DEVELOPMENT AND VALIDATION OF AN INSTRUMENT TO MEASURE MULTIDIMENSIONAL, INSTRUCTION-SPECIFIC STUDENT ENGAGEMENT IN UNDERGRADUATE MATHEMATICS

by

Daria Gerasimova
A Dissertation
Submitted to the
Graduate Faculty
of
George Mason University
in Partial Fulfillment of
The Requirements for the Degree
of
Doctor of Philosophy
Education

Committee:

________________________________ Chair

________________________________ Chair

____

____

____

________________________________ Program Director

________________________________ Dean, College of Education and Human Development

Date: ___________________________ Summer Semester 2020
George Mason University
Fairfax, VA

Development and Validation of an Instrument to Measure Multidimensional, Instruction-Specific Student Engagement in Undergraduate Mathematics

A Dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy at George Mason University

by

Daria Gerasimova
Master of Science
St. Petersburg State Polytechnical University, 2012
Bachelor of Science
St. Petersburg State Polytechnical University, 2010

Directors: Angela D. Miller, Associate Professor, Margret Hjalmarson, Professor
College of Education and Human Development

Summer Semester 2020
George Mason University
Fairfax, VA

Dedication

To my former students at School #136, St. Petersburg, Russia, who are the reason I know that the work we do in education matters.

Acknowledgements

The biggest thank you goes to my advisors: Dr. Angela Miller and Dr. Margret Hjalmarson. Neither this dissertation nor me becoming an educational researcher would have been possible without them.

It was because of Dr. Miller that I found how much I like quantitative educational psychology research and measurement. I am grateful for all the conversations I had with Dr. Miller about methods and measurement applied to my dissertation work and beyond. These conversations gave me the guidance I needed in conducting research, challenged my thinking, and helped me shape my research ideas. Dr. Miller's passion for and approach to quantitative research inspired me to dig deeper, learn more, and think critically to develop the best solutions to research problems.

It was because of Dr. Hjalmarson that I had a chance to pursue a Ph.D. I will be forever grateful to Dr. Hjalmarson for giving me this opportunity and supporting me all the way though the program. Under the guidance of Dr. Hjalmarson, I learned how to conduct qualitative research – the skill proved to be useful in my dissertation work. I am grateful to Dr. Hjalmarson for always solving all sorts of problems I had (and oh, there were many!) and for making sure that the financial aspect of the Ph.D. is not among my problems.

I have been fortunate to have advisors who truly cared for me, my research, my professional growth, and my success. Advisors who spent uncountable hours talking to me about research and life. Advisors who read so many drafts of so many documents (conference proposals, papers, job applications, and many more). Advisors who provided critical and constructive feedback that helped me improve my research and writing. Advisors whose support got me where I am today.

Next, I would like to thank all the other people who contributed to my development as an educational researcher. First, I want to thank the members of my committee. I am grateful to Dr. Marvin Powell for increasing my knowledge and skills in the area of measurement. I am grateful to Dr. Michelle Buehl for showing me the breadth and depth of the world of educational psychology. I am grateful to Dr. Jill Nelson for involving me in professional development research in STEM fields. I am grateful to all members of my committee for their feedback on my dissertation research – the feedback that helped me improve this dissertation.

Second, I would like to thank fellow Ph.D. students in mathematics education, educational psychology, and many other concentrations. It is priceless to have people to chat with about grad student matters and to support each other. And I am grateful I had such people. I want to particularly thank Kathy Matson (Ph.D.!) and her family for always inviting me to celebrate holidays in their house! Third, I would like to thank all of the STEM instructors at Mason who let me collect data from their students for this dissertation.

Next, I wanted to thank an online chat of Russian folks doing Ph.D.s in the U.S. I have never met them in person, but their support online helped me get through the toughest periods of the Ph.D. I particularly want to thank Yana Sosnovskaya who has always been there to talk about anything that mattered at any point in grad school. I am grateful for her cheering, advice, words of comfort – whatever I needed at any given moment. If there are online Ph.D. friendships for life, then this is it!

I am grateful for a furry companion I had during my Ph.D. studies – my cat Mira. Mira has been supporting me since my undergrad studies. Together, we moved from Russia to the U.S. when I started the Ph.D. program. It might be underestimated how much emotional support furry friends can provide. I have been fortunate to have such a friend.

I am grateful for my family in Russia: my mom, my farther, my brother, and my grandparents. Although they never quite understood why I decided to do a Ph.D. and why I needed to move to another continent for that, they have always supported me in my professional endeavors.

Finally, I would like to thank the Spencer Foundation and the College of Education and Human Development (CEHD) at George Mason University. A grant from the Spencer Foundation (Principal Investigator: Dr. Hjalmarson) and the CEHD Dissertation Research Fellowship funded this research.

Спасибо!

Table of Contents

Page

List of Tables

List of Figures

Abstract

DEVELOPMENT AND VALIDATION OF AN INSTRUMENT TO MEASURE MULTIDIMENSIONAL, INSTRUCTION-SPECIFIC STUDENT ENGAGEMENT IN UNDERGRADUATE MATHEMATICS

Daria Gerasimova, Ph.D.

George Mason University, 2020

Dissertation Directors: Dr. Angela D. Miller, Dr. Margret Hjalmarson

This study aimed to develop and provide initial validity evidence for an instrument to measure multidimensional, instruction-specific engagement in undergraduate mathematics-based classes. Multidimensionality included three engagement dimensions: behavioral, cognitive, and emotional (Fredricks et al., 2004). Instructional specificity included four instruction types: lecture, whole-class interaction, individual work, and group work. The study design had several phases. First, I reviewed the literature and conducted exploratory interviews to inform item writing. Then, an iterative cycle of pretesting via cognitive interviews, evaluating via expert reviews, and revising was implemented. Finally, the instrument was field-tested. Results provided initial validity evidence for combining multidimensionality and instructional specificity in engagement measurement. Yet, changes to the instrument's internal structure occurred. First, among behaviors in whole-class interaction, active behaviors separated out from passive behaviors. Second, excluding active behaviors in whole-class interaction, behavioral and cognitive engagement dimensions were not differentiated empirically and, therefore, were collapsed. Thus, the final version of the instrument measures nine multidimensional, instruction-specific engagement constructs. Additionally, I found that student engagement may be better conceptualized as a formative construct rather than a reflective construct. The instrument is designed to enable educators to identify how and where students are not engaged. This information would allow educators to develop more targeted and, therefore, more efficient interventions. The instrument is also designed to be used in research that aims to inform the development of such interventions. In particular, it will enable researchers to determine which instructional factors affect specific types of engagement and how these factors exert their influence.

Chapter One

Student engagement is essential for student success in mathematics-based university courses. Student engagement is a student's active involvement in learning (Skinner et al., 2009). President's Council of Advisors on Science and Technology (2012) noted that "students must be engaged to excel in STEM fields" (p. 3). Empirical studies at both K-12 and university levels showed that student engagement predicts achievement in mathematics and more broadly in STEM (Handelsman et al., 2005; Lau & Roeser, 2002; Skinner et al., 2017; M.-T. Wang et al., 2016; Whitney et al., 2019). Additionally, at the K-12 level, student engagement was found to predict STEM career aspirations (M.-T. Wang et al., 2016) and anticipated choices of science majors and careers (Lau & Roeser, 2002). At the university level, student engagement in science courses was positively associated with science identity and science career plans (Skinner et al., 2017). In turn, student university engagement was negatively associated with burnout (Assunção et al., 2020). Further, studies at the K-12 level showed that student engagement predicts dropping out of school (Archambault et al., 2009; Christenson et al., 2012; Fall & Roberts, 2012) and high school graduation (Finn & Zimmer, 2012). At the university level, studies found that more engaged students were less likely to have an intention to drop out (Assunção et al., 2020; Maroco et al., 2016) and more likely to have an intention to persist in college (Lerdpornkulrat et al., 2018).

To help students become more engaged, educators need to, first, improve their ability to identify the kind of engagement that students lack and, second, increase their knowledge about which instructional factors affect these kinds of student engagement and how these factors exert their influence. Critical to achieving these goals is improving engagement measurement. In general, engagement measurement incorporates two aspects: dimensionality and specificity. Designing an engagement measure, one needs to make decisions about the number of dimensions and the level(s) of specificity, at which engagement will be measured. Thus, I need to make my decisions about dimensionality and specificity in a way that will allow the instrument to maximize the ability of researchers and practitioners to meet the two needs stated above. Next, I describe my decision-making about the aspects of dimensionality and specificity.

Dimensionality

Engagement has been commonly seen as a complex construct with multiple features (Christenson et al., 2012). Thus, although unidimensional measures of engagement exist (Marks, 2000; Salmela-Aro et al., 2016), the multidimensional approach to engagement measurement has been more common. Fredricks et al. (2004) proposed a three-dimensional framework of student engagement that includes behavioral, cognitive, and emotional dimensions. Behavioral engagement refers to participation in academic and social or extracurricular activities (p. 60). Cognitive engagement refers to willingness to invest in learning (p. 60). Emotional engagement refers to students' reactions and feelings toward school, teachers, classmates, etc. (p. 60). Other engagement dimensions, such as social and agentic, have also been developed. Social engagement is concerned with student conduct (Finn & Zimmer, 2012) or with social relationships (Rimm-Kaufman et al., 2015; M.-T. Wang et al., 2016). Agentic engagement has been conceptualized by Reeve and Tseng (2011) as students' constructive contributions to the learning process. In another multidimensional framework, proposed by Schaufeli et al. (2002), engagement has been defined as a state of mind, characterized by vigor, dedication, and absorption. Here, vigor refers to energy, investment, and persistence in work. Dedication refers to feelings of significance, enthusiasm, pride, etc. Finally, absorption refers to deep concentration and full involvement in work.

From the definitions of different dimensions, one can see that there are qualitative differences between engagement dimensions. For example, behavioral engagement looks differently than emotional engagement. A behaviorally engaged student may be the one who actively participates in class discussions, pays attention, and works hard, whereas an emotionally engaged student may be the one who feels good in class, is interested in work, and enjoys learning (Skinner et al., 2009). Besides qualitative differences, there may also be quantitative differences between levels of engagement dimensions. For example, a student may be very active in class discussions, pay their full attention, and work very hard but simultaneously not feel good in class, not be interested in the work, and not enjoy learning. In the study of Skinner et al. (2008), students, on average, tended to be more engaged behaviorally than emotionally. In the study of Rimm-Kaufman et al. (2015), students' average cognitive engagement tended to be higher than emotional, which in turn tended to be higher than social. Further, correlations between some engagement dimensions were often found to be low or moderate. For instance, Reeve and

Tseng (2011) found low-to-moderate correlations between behavioral, cognitive, emotional, and agentic engagement. Similarly, Rimm-Kaufman et al. (2015) found low-to-moderate correlations between behavioral, cognitive, emotional, and social engagement. Mean differences between engagement dimensions and low-to-moderate correlations suggest that a student can be engaged in one way but not engaged in another. Measurement of engagement as a unidimensional construct does not provide this information. Yet, with this information, educators will be able to better help students become more engaged. First, they will be able to identify students who lack engagement of a particular type. The lack of engagement for such students can be masked by relatively high unidimensional engagement scores and, thus, overlooked. Second, knowing what type of engagement students lack will allow educators to develop more targeted and, therefore, more efficient instructional interventions.

In sum, qualitative and quantitative differences between engagement dimensions suggest that multidimensional measurement is more useful than unidimensional. Multidimensional measurement, in contrast to unidimensional, allows researchers and practitioners to capture qualitatively different dimensions of engagement. The information about these dimensions provides an opportunity to develop more strategic, dimension-specific efforts in helping students become more engaged.

Specificity

Student engagement has been measured at different levels of specificity. Broadly, student engagement has been measured at the school level (e.g., Salmela-Aro & Upadaya, 2012). Within school, the level of class has been used most frequently (e.g., Reeve, 2013;

Wang et al., 2014). The level of class also often incorporates domain specificity, such as math (e.g., Kong et al., 2003; M.-T. Wang et al., 2016) or science (e.g., Lau & Roeser, 2002; M.-T. Wang et al., 2016). Further, within a class, engagement at the level of instruction type was reported by the studies that employed the experience sampling method (ESM, Shernoff et al., 2003; Uekawa et al., 2007; Yair, 2000). In these studies, engagement was measured with respect to particular activities, which were later classified within broader types of instruction. The types of instruction, which partially overlap across the three studies, include lecture, class discussion, individual work, group work, work in laboratories, presentations, watching video, testing, and downtime.

A drawback of engagement measurement via ESM is its inability to capture qualitative differences between engagement in different types of instruction. For example, engagement in lecture is likely to look different than engagement in group work. In lecture, students would be expected to listen to the instructor and take notes, whereas in group work, students would be expected to work on tasks together and discuss ideas with each other. Yet, ESM measurement uses the same surveys at all time points, thus prohibiting the adaptation of survey questions to instruction types.

In addition to qualitative differences, there may also be quantitative differences between engagement in different types of instruction. For example, a student may listen to everything that their instructor is saying and take notes on the instructor’s explanations extensively but not participate in group work. Although ESM measurement is not able to capture qualitative differences, it provided evidence for quantitative differences. For example, Shernoff et al. (2003) found that students were more engaged during group and individual work than during the time listening to a lecture, watching TV/video, or taking exams. Yair (2000) reported that students were more likely to be engaged in discussion, work in labs, work in groups, individualized instruction, and TV/video but not in lecture or presentations than in the unknown type of instruction.

Similar to the quantitative differences between engagement dimensions, the quantitative differences in engagement levels between instruction types suggest that a student can be engaged in one type of instruction and not engaged in another type. Measurement of engagement at the broader levels, such as class or school, does not provide this information. Yet, with this information, educators will be able to better help students become more engaged. First, they will be able to identify students who are not engaged in a particular instruction type while being engaged in other instruction types. The lack of engagement for such students can be masked by relatively high class-level engagement scores and, thus, overlooked. Second, knowing where in class students lack engagement will allow educators to develop more targeted and, therefore, more efficient instructional interventions.

Further, measuring engagement at the level of instruction type does not prevent educators from having class-level engagement scores if such scores are needed for one’s purposes. In fact, creating class-level scores from instruction type scores helps avoid two threats to validity. These threats may be an issue for some class-level measures that use instruction-specific items. The first threat to validity is construct irrelevant variance, which may occur from using measures with items not applicable to the classes of measure administration. For instance, students may be asked about participating in class discussions (e.g., Ing & Victorino, 2016; Skinner et al., 2009) in a class where a teacher does not hold class discussions. As another example, students may be asked about working with their classmates (M.-T. Wang et al., 2016) or participating in small-group discussions (Handelsman et al., 2005) in a class where no such opportunities are provided. The second threat to validity is construct underrepresentation, which may occur from using a measure without items relevant to the class of measure administration. For example, students may be asked about their engagement in lecture and whole-class interaction but not in group work (Mazer, 2012) in a group work heavy class. It should be noted that the discussed threats to validity are not a concern for measures that use class-level items, such as paying attention in class (e.g., Skinner et al., 2009) or putting effort into learning (e.g., M.-T. Wang et al., 2016).

In sum, qualitative and quantitative differences between engagement in different instruction types suggest that instruction-specific measurement is more useful than more general, class-specific measurement. Instruction-specific measurement, in contrast to class-specific, allows for capturing qualitatively different ways of engagement in different instruction types. The information about these ways provides an opportunity to develop more strategic, instruction-specific efforts in helping students become more engaged. Additionally, employing instruction-specific measurement may produce more valid class-specific engagement scores.

Case for Combining Multidimensionality and Instructional Specificity

In the discussion of dimensionality and specificity, I demonstrated the benefits of multidimensional and instruction-specific measurement of engagement. Combining multidimensionality and instructional specificity will allow educators to identify both how (the dimension) or where (the instruction type) students are not engaged. Besides, instruction-specific measurement of engagement dimensions may lead to more valid dimensional scores at the class level. Thus, I aimed to develop an instrument that measures multidimensional, instruction-specific engagement in undergraduate mathematic-based STEM classes.

In terms of multidimensionality, I selected behavioral, cognitive, and emotional dimensions by Fredricks et al. (2004) because these dimensions comprehensively capture important aspects of engagement in learning and can be measured in different types of instruction. In contrast, social and agentic dimensions may already incorporate instructional specificity. The dimensions by Schaufeli et al. (2002; vigor, dedication, and absorption) are relatively narrow in scope and do not fully capture the complex construct of engagement.

In terms of instructional specificity, I first excluded testing and downtime from consideration because my instrument focuses on student engagement during instructional time. To ensure that categories for instructional time, used by other researchers (e.g., Shernoff et al., 2003; Uekawa et al., 2007; Yair, 2000), are collectively exhaustive, I specified instruction types based on two characteristics: a focus of instruction (instructor vs. students) and a type of interaction during the instruction. Thus, I distinguished between four instruction types: lecture (instructor-focused, no interaction), whole-class interaction (instructor-focused, interaction between the instructor and students), group work (student-focused, interaction between students), and individual work (student-focused, no interaction).

To my knowledge, the three-dimensional framework of Fredricks et al. (2004) and instruction-specific measurement of engagement have not been applied together. Existing multidimensional measures tend to be class-specific or even more general and, therefore, are not able to capture instruction-specific variations. For example, M.-T. Wang et al. (2016) included class- and subject-specific items in their measure, such as an item about feeling good in a math class or an item about the effort students put into learning math. At the broader levels of specificity, items can refer to class in general, to learning in general, or to school. For example, in the measure of Lam et al. (2014), students were asked about the effort they put in class, about being interested in learning, and about liking school. In contrast, studies that explored student engagement across instruction types did not measure all important dimensions of engagement. For example, Yair (2000) operationalized engagement as students' attention, and Shernoff et al. (2003) operationalized engagement as students' concentration, interest, and enjoyment.

Use and Significance of the Instrument

The measure is designed to be used for two purposes. First, educators will be able to use the measure to identify how (the dimension) and where (the instruction type) students lack engagement. This information will help educators with determining what kind of engagement should be a target of instructional interventions. For example, if students lack emotional engagement in group work, then intervention efforts should focus on that specific engagement dimension in that specific instruction type. Second, the measure is designed to be used in research that aims to inform the development of such interventions. In particular, this research will be able to determine how instructors can most effectively help their students become more engaged. To be more precise, it will be able to identify instructional factors that can facilitate specific kinds of student engagement. Further, this research will be able to determine the mechanisms through which instructional factors affect student engagement. For instance, instructional factors may affect student engagement through student personal factors, such as motivation (e.g., Lam et al., 2012), or through other contextual factors, such as classmates. Taking emotional engagement in group work as an example, one may hypothesize that this type of engagement may be affected by behaviors of other group members (e.g., willingness to work together, social loafing) and student relationships with them (e.g., friendships). Thus, instructional factors, such as developing group norms, forming groups in a particular way (e.g., self-selected or teacher-assigned), or implementing accountability practices, may affect students' behaviors, their attitudes toward each other, and ultimately their emotional engagement in group work. Further, emotional engagement in group work may be affected by students' motivation to work on tasks (e.g., students' self-efficacy or group efficacy to answer the task, as well as students' value of the task or of the instruction type). Thus, instructional factors, such as finding optimal levels of task complexity and number of tasks to work on, as well as using more relevant tasks and finding an optimal amount of instructional time to devote to group work, may lead to the increase in perceived efficacy and value, resulting in the increase in emotional engagement in group work. In sum, for each type of engagement in each instruction type there are multiple potential student-level factors that may affect this particular kind of engagement. Identifying these factors and determining how instructors can influence them is critical to increasing student engagement and subsequently achievement in undergraduate mathematics-based courses.

Summary

In this study, I aimed to develop and provide initial validity evidence for an instrument that measures student behavioral, cognitive, and emotional engagement in four types of instruction: lecture, whole-class interaction, group work, and individual work. The combination of engagement dimensions and instruction types is a novel approach in engagement measurement. This approach will allow educators to identify not engaged students more precisely and develop more targeted and, therefore, more efficient instructional interventions to help these students become more engaged. It will also allow researchers to determine instructional factors that affect particular types of engagement in particular instruction types and how these factors exert their influence. This knowledge will inform the development of instructional interventions that aim to increase student engagement. Thus, this study contributes to the advancement of engagement measurement and has the potential to be used in practice where it will help identify the specific kinds of engagement that students lack and, thus, help inform the foci of engagement interventions. It will also be instrumental in research that aims to develop effective instructional interventions to increase student engagement and, more broadly, to advance theories of effective teaching.

Chapter Two

In the first half of this chapter, I discuss how student engagement has been measured in the literature. Specifically, I focus on theoretical frameworks that guided the development of engagement measures, on the methods of engagement measurement, and on the levels of specificity that these measures employ. I also explore relationships that student engagement has been shown to have with its predictors and outcomes. Next, I present the design of the engagement measure I aimed to develop in this study. In the second half of this chapter, I review specific behaviors, cognitive processes, and emotions that students can experience in a mathematics-based classroom. As a result of the review, I develop precise conceptualizations of engagement dimensions used in this study.

Dimensionality

Although student engagement has been used extensively in educational research, there is little consensus among researchers on how it should be defined (Christenson et al., 2012; Sinatra et al., 2015). Student engagement has been conceptualized as a unidimensional or a multidimensional construct, although most researchers view student engagement as a multidimensional construct (Christenson et al., 2012). Yet, specific dimensions, their number, and conceptualizations varied among researchers. Some of these differences are rooted in the differences in theoretical frameworks that are used to guide the process of instrument development. Thus, in this section, I discuss the frameworks frequently utilized to conceptualize and operationalize student engagement. For each framework, I review dimensions and indicators that researchers have employed to operationalize student engagement. Broadly speaking, indicators are “markers or descriptive parts inside a target construct” (Skinner & Pitzer, 2012, p. 25). Specifically, I include both single-item and multi-item indicators (i.e., subscales), as some researchers chose to use the former whereas others chose to use the latter. I also review measures of student engagement that were not conceptualized within a particular theoretical framework. Studies included in this analysis were studies that developed an instrument of student engagement. These studies included instrument development studies and substantive studies that created instruments along the way. Studies that used or adapted existing items (but not full scales) from other studies or datasets were also included; however, studies that applied full existing scales were excluded.

The three-dimensional framework of Fredricks et al. (2004). The framework of Fredricks et al. (2004) includes three engagement dimensions: behavioral, cognitive, and emotional. In the engagement literature, behavioral engagement has also been referred to as physical (Burch et al., 2015), and emotional engagement has also been referred to as affective (Finn & Zimmer, 2012; Reschly & Christenson, 2012) or psychological (Appleton et al., 2006). In the discussion below, I included not only the studies that used all three dimensions but also the studies that used one or two of them.

Behavioral engagement. Behavioral engagement “draws on the idea of participation; it includes involvement in academic and social or extracurricular activities” (Fredricks et al., 2004, p. 60). This dimension is often indicated by behaviors within and outside of the classroom. Active on-task indicators within a classroom include on-task behaviors in general (Rimm-Kaufman et al., 2015), participation in class (M.-T. Wang et al., 2016) or, more specifically, participation in class activities (Kong et al., 2003; Z. Wang et al., 2014) and discussions (Kong et al., 2003; Li & Lerner, 2013; Miserandino, 1996; Rimm-Kaufman et al., 2015; Skinner et al., 2008; Z. Wang et al., 2014), working with other students (Maroco et al., 2016; Z. Wang et al., 2014), as well as not wanting to stop working (Z. Wang et al., 2014). Specific participation behaviors include asking questions, making suggestions, trying to answer questions (Hospel et al., 2016), and formulating questions in mind (Z. Wang et al., 2014). Self-reliance during in-class work has also been used as an indicator of behavioral engagement (Rimm-Kaufman et al., 2015). Further, more passive on-task behaviors are listening (Gunuc & Kuzu, 2015; Kong et al., 2003; Miserandino, 1996; Skinner et al., 2008; Z. Wang et al., 2014), paying attention (Miserandino, 1996; Rimm-Kaufman et al., 2015; Skinner et al., 2008; M.-T. Wang et al., 2011), concentrating (Kong et al., 2003; Rimm-Kaufman et al., 2015), and staying focused (M.-T. Wang et al., 2016). Writing-related behaviors include note-taking (Whitney et al., 2019).

Other behaviors commonly used to indicate behavioral engagement include effort (Burch et al., 2015; Kong et al., 2003; Lam et al., 2014; J.-S. Lee, 2014; Miserandino, 1996; Rimm-Kaufman et al., 2015; Skinner et al., 2008; M.-T. Wang et al., 2016) and persistence (Kong et al., 2003; Lam et al., 2014; J.-S. Lee, 2014; Miserandino, 1996; M.-T. Wang et al., 2016). Further, researchers have also used positive conduct-related indicators, such as attendance (Archambault et al., 2009), as well as following instructions (Hospel et al., 2016) and rules (Gunuc & Kuzu, 2015; Maroco et al., 2016). Behaviors, occurring prior to class, include completing assignments (Z. Wang et al., 2014), reviewing assignments before submission (Whitney et al., 2019), preparation (Li & Lerner, 2013; Reschly & Christenson, 2006), time spent on homework (Kong et al., 2003; Reschly & Christenson, 2006), and finishing homework on time (Gunuc & Kuzu, 2015; Hospel et al., 2016; Li & Lerner, 2013; M.-T. Wang et al., 2016). Other out-of-class behaviors consist of talking about the material outside of class (M.-T. Wang et al., 2016), studying on a regular basis (Whitney et al., 2019), time spent on out-of-class learning (Kong et al., 2003), and participation in extracurricular activities (Lam et al., 2014; Reschly & Christenson, 2006; Sciarra & Seirup, 2008).

Not desired in-class behaviors used to indicate behavioral engagement include daydreaming (M.-T. Wang & Holcombe, 2010), pretending to work (Lam et al., 2014; Miserandino, 1996), falling asleep (Miserandino, 1996), thinking about (Miserandino, 1996) or doing (M.-T. Wang et al., 2016) other things, experiencing helplessness (Miserandino, 1996), and asking off-topic questions (Rimm-Kaufman et al., 2015). Other negative behaviors were related to conduct, such as not following the rules (Fall & Roberts, 2012), disrupting class (Archambault et al., 2009; Hospel et al., 2016; Rimm-Kaufman et al., 2015), being rude to the teacher (Archambault et al., 2009), absenteeism (Fall & Roberts, 2012; Hospel et al., 2016; Li & Lerner, 2013; Reschly & Christenson, 2006; Sciarra & Seirup, 2008), tardiness (Awang Hashim & Murad Sani, 2008; Fall & Roberts, 2012; Reschly & Christenson, 2006; Sciarra & Seirup, 2008), coming to class unprepared (Awang Hashim & Murad Sani, 2008), fighting (Awang Hashim & Murad

Sani, 2008; Reschly & Christenson, 2006; M.-T. Wang et al., 2011), being sent to office (Reschly & Christenson, 2006; M.-T. Wang et al., 2011), and disciplinary actions (Sciarra & Seirup, 2008).

Cognitive engagement. Cognitive engagement “draws on the idea of investment; it incorporates thoughtfulness and willingness to exert the effort necessary to comprehend complex ideas and master difficult skills” (Fredricks et al., 2004, p. 60). Most commonly used indicators are those related to self-regulation (Miller et al., 1996; Thien & Razak, 2013; M.-T. Wang et al., 2011; M.-T. Wang & Holcombe, 2010), deep and shallow processing strategy use (Kong et al., 2003; Miller et al., 1996), and cognitive strategy use (M.-T. Wang et al., 2011). Specific indicators when not grouped into subscales include trying to figure out (Z. Wang et al., 2014) or understand (M.-T. Wang et al., 2016) mistakes, going back to not understood material, thinking deeply (Z. Wang et al., 2014), checking work (M.-T. Wang et al., 2016), asking oneself questions to check understanding (Awang Hashim & Murad Sani, 2008; Z. Wang et al., 2014), thinking about different ways to solve a problem (M.-T. Wang et al., 2016), connecting new knowledge with existing knowledge (Lam et al., 2014; M.-T. Wang et al., 2016) or with experiences (Lam et al., 2014), combining ideas from different courses, summarizing the material, identifying key information when reading (Whitney et al., 2019), paraphrasing, making up examples (Lam et al., 2014), and setting study goals (Awang Hashim & Murad Sani, 2008).

Other indicators of cognitive engagement include paying attention (Burch et al., 2015; Rimm-Kaufman et al., 2015), concentrating, and focusing (Burch et al., 2015).

Further, also used are effort (Rimm-Kaufman et al., 2015), persistence (Miller et al., 1996), willingness to learn (Archambault et al., 2009), importance of learning (Rimm-Kaufman et al., 2015), usefulness of the material in real world (Lam et al., 2014; Reschly & Christenson, 2006), and the perception of connection between school and students' lives (Li & Lerner, 2013). Besides, goal orientation (Li & Lerner, 2013), reliance on the teacher (Kong et al., 2003), identification with school (Li & Lerner, 2013), boredom (Reschly & Christenson, 2006), and talking about the subject with others outside of school (Maroco et al., 2016) were found among indicators of cognitive engagement, as well. In addition, negative indicators, such as thinking about other things in class (Rimm-Kaufman et al., 2015) or avoiding work (M.-T. Wang et al., 2016), have also been used to measure cognitive engagement.

Emotional engagement. Emotional engagement "encompasses positive and negative reactions to teachers, classmates, academics, and school and is presumed to create ties to an institution and influence willingness to do the work" (Fredricks et al., 2004, p. 60). A number of specific emotions have been used to indicate emotional engagement. In particular, researchers have included interest (Archambault et al., 2009; Burch et al., 2015; Kong et al., 2003; Lam et al., 2014; Miserandino, 1996; Rimm-Kaufman et al., 2015; Skinner et al., 2008; Z. Wang et al., 2014) and a related item of finding ways to make the course material interesting (Appleton et al., 2006). Other positive emotions included feelings of enjoyment (Lam et al., 2014; Li & Lerner, 2013; Skinner et al., 2008; M.-T. Wang et al., 2016), happiness (Lam et al., 2014; Li & Lerner, 2013; Miserandino, 1996; Z. Wang et al., 2014), excitement (Burch et al., 2015; Li &

Lerner, 2013; Maroco et al., 2016; Z. Wang et al., 2014), pride (Burch et al., 2015; Lam et al., 2014; Z. Wang et al., 2014), as well as pleasure and satisfaction (Kong et al., 2003). Other positive emotions include feeling relaxed and comfortable (Miserandino, 1996), amused (Z. Wang et al., 2014), energetic (Burch et al., 2015), and enthusiastic (Burch et al., 2015). Further, researchers have also used more general positive feelings, such as feeling good (Miserandino, 1996; Skinner et al., 2008; M.-T. Wang et al., 2016) or feeling positive (Burch et al., 2015).

Negative emotions have also been used to indicate emotional engagement. These emotions include boredom (Lam et al., 2014; Miserandino, 1996; Rimm-Kaufman et al., 2015; M.-T. Wang et al., 2016), feeling tired or sleepy (Miserandino, 1996), feeling frustrated (Kong et al., 2003; M.-T. Wang et al., 2016), anxious (Kong et al., 2003), worried (M.-T. Wang et al., 2016), as well as feeling scared, unhappy, sad, mad, and angry with respect to learning in class (Miserandino, 1996). More general negative feelings are concerned with feeling bad, terrible (Miserandino, 1996), and down (M.-T. Wang et al., 2016).

Additionally, measures of emotional engagement include attitudinal questions. For example, in some of such measures, students were asked about how much they like their school (Archambault et al., 2009; Lam et al., 2014), what they learn there (Lam et al., 2014), specific aspects of learning (e.g., the feeling of solving problems, Rimm-Kaufman et al., 2015), their teachers and communicating with the teachers, and seeing friends in class (Gunuc & Kuzu, 2015). Some researchers have also asked whether the class is fun (Rimm-Kaufman et al., 2015; Skinner et al., 2008) and whether a student looks forward to going to class (M.-T. Wang et al., 2016) or to school (Lam et al., 2014). Lastly, indicators also include the desire to understand the material (Appleton et al., 2006; M.-T. Wang et al., 2016). Negative attitudes, on the other hand, include not wanting to be in class and not caring about learning (M.-T. Wang et al., 2016).

Besides feelings and attitudes, emotional engagement questions have also asked about involvement in class work (Miserandino, 1996; Skinner et al., 2008), applying course material to one's life, and thinking about the course between class meetings (Appleton et al., 2006). Valuing of school education (Finn & Zimmer, 2012; M.-T. Wang et al., 2011) and finding ways to make the course material relevant to one's life (Appleton et al., 2006) were found among indicators of emotional engagement, as well. Further, the sense of belonging to a school (Awang Hashim & Murad Sani, 2008; Finn & Zimmer, 2012; J.-S. Lee, 2014; Thien & Razak, 2013; M.-T. Wang et al., 2011; M.-T. Wang & Holcombe, 2010) or a student group (Gunuc & Kuzu, 2015) also serves as an indicator of emotional engagement. The last set of indicators is concerned with peer and student-teacher relationships (Gunuc & Kuzu, 2015; Sciarra & Seirup, 2008), as well as interactions with teachers (Reschly & Christenson, 2006; see also Voelkl, 1995). Specifically, Gunuc and Kuzu (2015) included having close friends in class and the opportunities to share problems with the teachers. Besides sharing problems, interaction with teachers also included talking about jobs, courses, drug/alcohol abuse, etc. (Reschly & Christenson, 2006; Voelkl, 1995). Indicators that may also incorporate relationships are school safety and harmony among different racial groups (Sciarra & Seirup, 2008).

Modifications and additions to the framework of Fredricks et al. (2004). Some researchers have distinguished between other dimensions of student engagement as modifications of the dimensions of Fredricks et al. (2004) or as additions to them. These dimensions include academic, social, and agentic engagement.

Academic engagement. In the conceptualization of Finn and Zimmer (2012), academic engagement – a variation of behavioral engagement – refers to behaviors directly related to the learning process. These behaviors include attentiveness and completing assignments in class and at home or augmenting learning through academic extracurricular activities (p. 102). Examples of indicators that Finn and Zimmer (2012) used to measure academic engagement included paying attention, participating in class discussion, and completing assignments. Further, Fall and Roberts, (2012) indicated academic engagement with attention, withdrawal, homework completion, and effort. Reschly and Christenson (2006) defined academic engagement as “the amount of time that students spend on task” (p. 278). These authors included the following indicators of academic engagement: time on task, credit hours toward graduation (high school), homework completion rate and accuracy, and class grades (number of failing grades). The term of academic engagement, specifically behavioral academic engagement, was also used by Gasiewski et al. (2012). To measure academic engagement, these authors used indicators of asking questions in class, discussing course grades or assignments with the instructor, attending instructor’s office hours, participating in class discussions, tutoring other students, reviewing material before the class, attending review or help sessions, and studying with other students.

Social engagement. In the conceptualization of Finn and Zimmer (2012), social engagement – another variation of behavioral engagement – is concerned with student conduct, i.e., “the extent to which a student follows written and unwritten classroom rules of behavior, for example, coming to school and class on time, interacting appropriately with teachers and peers, and not exhibiting antisocial behaviors such as withdrawing from participation in learning activities or disrupting the work of other students” (p. 102). Example indicators that Finn and Zimmer (2012) used to measure social engagement include needing to be reprimanded and interfering with classmates’ work. In contrast, in the perspective of Wang et al. (2016), social engagement includes “the quality of social interactions with peers and adults, as well as the willingness to invest in the formation and maintenance of relationships while learning” (p. 17). Similarly, Rimm-Kaufman et al. (2015) refer to social engagement, or task-related interaction in the terminology of Patrick et al. (2007), as “students’ day-to-day social exchanges with peers that are tethered to the instructional content” (p. 172). In these conceptualizations, social engagement was operationalized by talking to others in class (Rimm-Kaufman et al., 2015), working with others (M.-T. Wang et al., 2016), sharing ideas and materials (Rimm-Kaufman et al., 2015), helping others (Rimm-Kaufman et al., 2015; M.-T. Wang et al., 2016), as well as trying to understand and build on other people’s ideas (M.-T. Wang et al., 2016).

Agentic engagement. A dimension of agentic engagement was proposed by Reeve and Tseng (2011) in addition to the three dimensions of Fredricks et al. (2004). According to Reeve and Tseng (2011), agentic engagement is “students’ constructive contribution into the flow of the instruction they receive" (p. 258). Indicators of agentic engagement include informing teachers about needs and interests, expressing preferences and opinions, and asking questions (Reeve, 2013).

The Self-System Process Model: Engagement vs. disaffection. The Self-System Process Model (Connell, 1990; Connell & Wellborn, 1991) was developed to describe the functioning of a self-system, specifically the relations between social context, self (psychological needs), actions, and outcomes of those actions. The model states that people have fundamental psychological needs – for competence, autonomy, and relatedness – that are affected by social context and that in turn affect people's actions, variability in which leads to variability in outcomes. The part of this model that is of interest to engagement researchers is the construct of action. For people's actions, the developers distinguished between engagement and disaffection. In this framework, engagement is defined as "patterns of action reflecting acceptance of and commitment to the goals of learning and successful school performance" (Connell, 1990, p. 87), and disaffection is defined as "patterns of action reflecting a lack of commitment to these goals" (Connell, 1990, p. 87). Those patterns of action also include cognitions, emotions, and behaviors, which connect this theory to the three-dimensional framework of Fredricks et al. (2004).

Following the Self-System Process Model, some researchers incorporated the engagement and disaffection distinction in their measures. For example, Skinner et al. (2008, 2009) developed separate engagement and disaffection scales for both behavioral and emotional engagement, resulting in the creation of four scales: behavioral engagement, behavioral disaffection, emotional engagement, and emotional disaffection. Skinner et al. (2008) emphasized that disaffection is more than the absence of engagement. Rather, behavioral and emotional disaffection refer to behaviors and emotions that reflect maladaptive motivational states. To indicate behavioral disaffection, Skinner et al. (2008, 2009) employed behaviors, such as acting like working, not trying very hard, doing just enough to get by, and thinking about other things. To indicate emotional disaffection, the authors used a number of negative emotions, including boredom, worry, nervousness, discouragement, frustration, as well as feeling bad, mad, and bothered. More recently, Z. Wang et al. (2014) also included a disengagement dimension, in addition to behavioral, cognitive, and affective dimensions. The disengagement dimension was indicated by two behavioral disaffection items of Skinner et al. (2008) and by the process of “zoning out” of Valentine and Painter (2007).

Flow theory. Flow theory was created by Csikszentmihalyi (1990, 1997) and originally was not associated with engagement measurement. However, in later years, several theoretical frameworks for student engagement were developed based on this theory. Flow, as Csikszentmihalyi (1990) defines it, is “the state in which people are so involved in an activity that nothing else seems to matter; the experience itself is so enjoyable that people will do it even at great cost, for the sheer sake of doing it” (p. 4). According to the original model of the flow state (Csikszentmihalyi, 1975/2000), people experience flow when perceived opportunities for action (challenges) match their capabilities (skills). If challenges are greater than skills, people experience anxiety; if, in contrast, skills are greater than challenges, people experience boredom. Later, conditions for experiencing flow were redefined. According to the current model of the flow state (Csikszentmihalyi, 1997), flow occurs when both skills and challenges are high. When both skills and challenges are low, instead of flow, people experience apathy. Other combinations of skill and challenge levels produce experiences such as worry (low skills, moderate challenges), anxiety (low skills, high challenges), arousal (moderate skills, high challenges), control (high skills, moderate challenges), relaxation (high skills, low challenges), and boredom (moderate skills, low challenges).

Cavanagh and Kennish (2009) applied flow theory to student engagement in classroom learning. They conceptualized the flow zone where skills match challenges as the zone of engagement in learning. In their model, Cavanagh and Kennish (2009) defined skills as learning capabilities, and challenges as expectations of learning. A balance of learning capabilities and expectations of learning produces engagement in learning, with high levels of each leading to high engagement and low levels of each leading to low engagement. Measuring engagement within this model includes measuring both dimensions: learning capabilities and expectations of learning (Cavanagh, 2015; Kennish & Cavanagh, 2011). Specifically, learning capabilities were indicated by self-esteem, self-concept, resilience, self-regulation, and self-efficacy; expectations of student learning were indicated by explanation, interpretation, application, perspective, empathy, and self-knowledge.

Shernoff et al. (2003) also conceptualized engagement within the framework of flow theory, yet they did it differently from Cavanagh and Kennish (2009). Specifically, Shernoff et al. (2003) used three components of flow theory, which need to be experienced simultaneously for the flow to occur: concentration, interest, and enjoyment. Thus, from the flow theory perspective, student engagement is defined as “the heightened, simultaneous experience of concentration, interest, and enjoyment in the task at hand” (Shernoff, 2013, p. 14). Shernoff et al. (2016) refer to this conceptualization of engagement as subjective engagement, noting that it is distinct from behavioral, cognitive, or affective engagement dimensions of Fredricks et al. (2004). Yet, subjective engagement intersects with the framework of Fredricks et al. (2004), as it includes both affective (enjoyment) and cognitive (concentration) elements (Shernoff et al., 2016).

Framework of vigor/energy, dedication, and absorption. Engagement within this framework originated from the burnout research in the work context. Specifically, in this research, engagement was hypothesized to be the opposite of burnout (Schaufeli et al., 2002). Engagement was defined as “a positive, fulfilling, work-related state of mind that is characterized by vigor, dedication, and absorption” (Schaufeli et al., 2002, p. 74). According to the authors, vigor was characterized by “high levels of energy and mental resilience while working, the willingness to invest effort in one’s work, and persistence even in the face of difficulties” (p. 74). Dedication was characterized “by a sense of significance, enthusiasm, inspiration, pride, and challenge” (p. 74). Finally, absorption was characterized by “being fully concentrated and deeply engrossed in one’s work, whereby time passes quickly and one has difficulties with detaching oneself from work” (p. 74). Schaufeli et al. (2002) note that the dimension of absorption is similar to the concept of flow of Csikszentmihalyi (1990), yet the two concepts are not identical. According to Schaufeli et al. (2002), flow is a more complex concept than absorption;

also, flow refers to short-term experiences, whereas absorption (as well as vigor and dedication) refer to a more pervasive and persistent state of mind. Schaufeli et al. (2002) developed two versions of the instrument: an employee version and a student version. Indicators of vigor in the student version included feeling like going to class, energy, perseverance, and resilience. Indicators of dedication in the student version included challenge, inspiration, enthusiasm, pride, and meaningfulness. Indicators of absorption in the student version included forgetting other things, time flying, immersion, and happiness. Schaufeli et al. (2006) later modified the employee version of the instrument to create its short version.

Salmela-Aro and Upadaya (2012) adopted the framework and adapted the instrument of Schaufeli et al. (2006) to develop the Schoolwork Engagement Inventory. In the context of school, energy (labeled as vigor in the work of Schaufeli et al., 2002, 2006) refers to “a positive approach to schoolwork,” dedication refers to “a positive cognitive attitude and perceiving schoolwork as meaningful,” and absorption refers to “full concentration on studying so that time seems to pass quickly” (p. 60). Salmela-Aro and Upadaya (2012) also connected this framework to the three-dimensional framework of Fredricks et al. (2004), describing energy as an emotional component, dedication as a cognitive component, and absorption as a behavioral component of engagement. Indicators used in the measure of Salmela-Aro and Upadaya (2012) were the same as in the measure of Schaufeli et al. (2006).

No framework. Some instruments were developed without an explicit reference to a particular framework. Some of these measures were unidimensional, whereas others were multidimensional. For example, Ing and Victorino (2016) created a classroom engagement scale using academic engagement items from a larger survey. This scale included indicators, such as contributing to class discussion, asking questions, interacting with faculty during lecture, helping classmates, etc. Marks (2000) created a measure of engagement in the instructional activity. This unidimensional scale consisted of effort, attentiveness, lack of boredom, and completing class assignments.

Further, Salmela-Aro et al. (2016) developed a unidimensional measure of situational engagement, indicators of which included feeling active and interested, enjoyment, and importance of the task. An instrument of Uekawa et al. (2007) was also situation-specific and unidimensional. In their study, engagement was indicated by paying attention, not feeling like listening, motivation, boredom, enjoyment, staying focused, wishing the class to end, and being completely into class. Another situation-specific engagement was investigated in the study of Yair (2000), who worked with the same dataset, as Shernoff et al. (2003), but conceptualized engagement differently. While Shernoff et al. (2003) used a flow-based conceptualization that combines concentration, interest, and enjoyment, Yair (2000) viewed student engagement in terms of attention. Specifically, Yair (2000) considered students to be engaged if their location (e.g., a classroom) matched their thoughts (e.g., class material). In contrast, Yair (2000) considered students to be disengaged if there was a mismatch between the place and thought (e.g., being in a classroom but thinking about non-class related things).

An example of a multidimensional measure, not conceptualized within any major engagement framework, was the instrument of Lau and Roeser (2002), which consisted of three dimensions: engagement in a classroom, engagement in extracurricular activities, and engagement in testing situations. The classroom engagement dimension was indicated by paying attention in class, completing homework, participating in classroom activities, and using self-regulatory strategies. The extracurricular engagement dimension was indicated by visiting science-related websites, reading science-related magazines or books in free time, watching science-related television programs, and talking to parents or other adults about science-related issues. Finally, the test engagement dimension was indicated by test mood, energy, effort, and use of test-taking strategies.

Mazer (2012) found a four-dimensional structure of engagement in their study. Specifically, the dimensions included silent in-class behaviors, oral in-class behaviors, thinking about course content, and out-of-class behaviors. Indicators of silent in-class behaviors included attentive listening, giving full attention, attending class, etc. Indicators of oral in-class behaviors were concerned with oral participation in class discussions. Indicators of thinking about course content consisted of various ways students may think about course material. Finally, indicators of out-of-class behaviors included reviewing notes, studying for a test, talking about the course material with others, and reading additional literature about the course topic.

Finally, a measure of Handelsman et al. (2005) had a four-dimensional structure that included dimensions of skills, emotional engagement, participation/interaction, and performance. Indicators of skills included putting effort, completing homework, doing readings, reviewing notes, being organized, taking notes, careful listening, etc. Indicators of emotional engagement were discussed earlier within the three-dimensional framework of Fredricks et al. (2004). Participation/interaction was indicated by raising a hand, asking questions, having fun, participating in small-group discussions, attending office hours, and helping other students. Finally, the performance dimension consisted of the following indicators: getting good grades, doing well on tests, and being confident in the ability to learn and do well in class.

Qualitative differences between dimensions. The description of theoretical frameworks, dimensions, and indicators showed their variability in engagement measurement. Engagement dimensions within and across frameworks have qualitative differences. In other words, how students are engaged in different dimensions looks differently. The three-dimensional framework is the most widely used framework, as can be seen from the reviewed studies. It should be noted that the three-dimensional framework has been applied in a variety of ways, and qualitative differences between dimensions in this framework are not necessarily consistent across measures. To illustrate qualitative differences within measures, I provide several examples. Awang Hashim and Murad Sani (2008) conceptualized behavioral engagement in terms of compliance with school and classroom rules, cognitive engagement as students' thinking, processing the information, and self-directed learning, and emotional engagement as a sense of identification with school and positive relationships with peers. M.-T. Wang et al. (2011) conceptualized behavioral engagement as attentiveness and school compliance, cognitive engagement as self-regulated learning and cognitive strategy use, and emotional engagement as school belonging and valuing of school education. Lam et al. (2014) conceptualized behavioral engagement as effort and persistence in schoolwork as well as participation in extracurricular activities, cognitive engagement as cognitive strategy use, and emotional engagement as feelings about learning and school. In these examples, the level of agreement on engagement conceptualizations and operationalizations differed within engagement dimensions across measures. In particular, behavioral engagement of Awang Hashim and Murad Sani (2008) consisted of compliance, behavioral engagement of M.-T. Wang et al. (2011) also included attentiveness, and behavioral engagement of Lam et al. (2014) did not include compliance at all and instead included effort, persistence, and participation in extracurricular activities. To indicate cognitive engagement, both Awang Hashim and Murad Sani (2008) and M.-T. Wang et al. (2011) included cognitive strategy use and self-regulated learning, whereas Lam et al. (2014) limited the construct only to cognitive strategy use. Finally, to indicate emotional engagement, Awang Hashim and Murad Sani (2008) included both belonging and peer relationships, M.-T. Wang et al. (2011) also included belonging but used value of school and not peer relationships, and Lam et al. (2014) focused on feelings more broadly. In sum, dimensions of engagement may have qualitative differences, although these differences are not necessarily consistent across measures.

Quantitative differences between dimensions. In addition to qualitative differences, there may also be quantitative differences between engagement dimensions. In other words, a student’s level of engagement in one dimension may be different from their level of engagement in another dimension. Low-to-moderate correlations may provide evidence for such quantitative differences. Specifically, low-to-moderate correlations were observed between behavioral, cognitive, and emotional engagement

(Lam et al., 2014; Lerdpornkulrat et al., 2018; Li & Lerner, 2013; Maroco et al., 2016; Reeve & Tseng, 2011; M.-T. Wang, 2010; Whitney et al., 2019), as well as for social (Rimm-Kaufman et al., 2015) and agentic (Reeve, 2013; Reeve & Tseng, 2011) engagement with behavioral, cognitive, and emotional engagement. Outside of the three-dimensional framework and its additions/modifications, low-to-moderate correlations were found between engagement and disaffection dimensions of behavioral and emotional dimensions (Skinner et al., 2008, 2017), test, classroom, and extracurricular engagement (Lau & Roeser, 2002), and the four dimensions of Mazer (2012; silent in-class behaviors, oral in-class behaviors, thinking about course content, and out-of-class behaviors). Yet, for some dimensions, in some studies, correlations were quite high. For example, quite high correlations were observed between behavioral, cognitive, and emotional dimensions in the studies of M.-T. Wang et al. (2011) and Z. Wang et al. (2014), as well as between the dimensions of energy, dedication, and absorption (Salmela-Aro & Upadaya, 2012). In sum, there may be quantitative differences between different dimensions of engagement, although some studies showed that this within-student variation may be rather small.

Rationale for measuring multidimensional engagement. The review of theoretical frameworks and dimensions within these frameworks showed that there are both unidimensional and multidimensional measures of student engagement. However, the multidimensional approach to engagement measurement was substantially more common. This observation is not necessarily surprising. Indeed, the multidimensional approach allows educators to capture qualitatively different dimensions that provide information about quantitative differences in engagement levels of a particular student. Thus, the multidimensional approach provides an opportunity to develop more strategic, dimension-specific engagement interventions.

Methods of Measurement

While a variety of methods have been used to measure student engagement, most widely used is a retrospective student report. It is a common method of measurement for behavioral, cognitive, emotional, academic, social, and agentic engagement dimensions, disaffection, as well as dimensions vigor/energy, dedication, and absorption. Examples of such measures include measures by Archambault et al. (2009), Burch et al. (2015), Hospel et al. (2016), Kong et al. (2003), Miller et al. (1996), Miserandino (1996), Reeve (2013), Reeve and Tseng (2011), Skinner et al. (2008), M.-T. Wang et al. (2011), and Z. Wang et al. (2014). Some researchers utilized student self-report measures in combination with other methods. For example, Rimm-Kaufman et al. (2015) used student self-reports to measure cognitive, emotional, and social engagement but teacher reports and classroom observations to measure behavioral engagement. Whereas Rimm-Kaufman et al. (2015) applied two methods only to one engagement dimension, M.-T. Wang et al. (2016) employed two methods – student and teacher reports – to measure all engagement dimensions under investigation: behavioral, cognitive, emotional, and social engagement. Differently from M.-T. Wang et al. (2016), Sciarra and Seirup (2008) combined student and teacher reports to develop a single scale. In particular, cognitive and emotional engagement dimensions of Sciarra and Seirup (2008) included both teacher- and student-reported items. Finally, behavioral, cognitive, and emotional engagement were also measured via neurophysiological measures. For example, Charland et al. (2015) measured behavioral engagement via eye tracking, cognitive engagement via electroencephalography (EEG), and emotional engagement via automatic facial emotion recognition software (to measure emotional valence) and electrodermal activity encoder/sensor (to measure emotional arousal).

Engagement within flow theory employed several methods of measurement. Shernoff et al. (2003) measured engagement via the experience sampling method (ESM). This method aims to measure people's experiences in random moments (Larson & Csikszentmihalyi, 2014). Specifically, participants receive signals (e.g., via a pager) that serve as a sign to complete a brief survey about their experience at the moment of the signal. Engagement within the framework of Cavanagh and Kennish (2009), which was also based on flow theory, was measured via two different methods. In the study of Kennish and Cavanagh (2011), engagement was measured via a researcher-completed rating scale instrument. Specifically, researchers interviewed students about their engagement and in the process assigned ratings for each engagement dimension based on the amount of evidence for the dimension. In contrast, Cavanagh (2015) measured engagement within this framework via student self-report. Finally, engagement that was not conceptualized within a particular framework was measured either via student self-report (Handelsman et al., 2005; Ing & Victorino, 2016; Lau & Roeser, 2002; Marks, 2000; Mazer, 2012) or ESM (situational engagement of Salmela-Aro et al., 2016; also Uekawa et al., 2007; Yair, 2000).

In sum, student engagement has been measured via a variety of methods, although a student self-report is the most common. This observation is not necessarily surprising considering the ease of collecting survey data from students. Although self-reports are subject to the social desirability bias, it may be argued that students may have the best knowledge about their engagement, especially for those types that cannot be observed (e.g., cognitive engagement). Further, teacher reports and observational measures may also limit the number of students whose engagement is measured, particularly in large classes. Student self-report measures can be differentiated between retrospective and in-the-moment (via ESM). ESM measurement is particularly useful when moment-to-moment fluctuations in the levels of engagement are of interest. However, when one is interested in engagement over a particular period of time, retrospective measurement can be more useful. It targets a specific, defined in advance time period rather than particular activities, i.e., "snapshots" of engagement during this time period, which might not represent the entire time period. Retrospective measurement is easier and cheaper than ESM; yet, it comes at a price of potentially lower precision compared to ESM with a sufficiently large number of administrations. Another advantage of retrospective measurement over ESM is potentially stronger content validity of measured constructs. As ESM measures are administered frequently, the number of items that can be included in an ESM survey has to be small. In contrast, one-time retrospective surveys can be longer. Thus, as in this study I was interested in student engagement over the course of the semester and aimed to measure multiple engagement dimensions, which may include unobservable dimensions, in classes of various size, the retrospective approach seemed to be the most reasonable option.

Specificity of Measurement

Skinner and Pitzer (2012) viewed student engagement as nested at four levels. The broadest level refers to engagement with prosocial institutions, such as church, youth groups, family, community, and, of course, school. Engagement with school is placed at the second level. In school, students are engaged with sports, clubs, government, and most importantly a classroom. Within the classroom, the third level, students engage with teachers, peers, and curriculum. Finally, at the lowest level, students engage directly with learning activities. Thus, engagement can occur at different levels of specificity, as well as across a variety of situations within levels and across different variations of levels (e.g., specific classes at the class level).

The broadest level of engagement measurement, used in engagement research, is the school level. For example, Li and Lerner (2013) used a sense of belonging and feelings toward school to measure emotional engagement. Items of Salmela-Aro and Upadaya (2012) measured energy, dedication, and absorption with respect to school or school work. In other measures, although item wording referred to engagement in class, this reference was not specific to a particular class but rather to a class in general terms. For example, in the study of Reeve and Tseng (2011), students rated their engagement with respect to all classes they were taking. In the Wilson et al. (2015) study, students also rated their engagement across classes but only within those in their major. Further, some researchers did not mention school or class in their items; instead, they referred to the process of studying or learning in general (e.g., in the cognitive dimension of Lam et al., 2014).

At the class/subject level, some measures were specific to a particular class or subject. For example, in the studies by Z. Wang et al. (2014) or Reeve (2013), students reported on their engagement in the class of survey administration. In the study of Mazer (2012), students were asked to use their first class in a week as a referent, as opposed to the class from which they were recruited. In other studies, students reported their engagement in a math class (Hospel et al., 2016; Kong et al., 2003; Marks, 2000; Miller et al., 1996; Rimm-Kaufman et al., 2015; M.-T. Wang et al., 2016), a French class (Hospel et al., 2016), a social studies class (Marks, 2000), or a science class (Lau & Roeser, 2002; M.-T. Wang et al., 2016). Some researchers also differentiated between in-class and out-of-class engagement. For example, Burch et al. (2015) separated in-class cognitive engagement (i.e., when a student is in a classroom for a particular class/course) and out-of-class cognitive engagement (i.e., when a student is reading or studying material related to a particular class/course). Mazer (2012) separated in-class and out-of-class behaviors. Ing and Victorino (2016) focused their measure on engagement in a classroom only. Lau and Roeser (2002) measured student engagement from what they called a situational perspective. These authors went beyond the in-class/out-of-class distinction, measuring classroom, test, and extracurricular engagement. The term situational engagement was also used in the studies that measured engagement via ESM (e.g., Salmela-Aro et al., 2016). In these studies, engagement was measured at the level of activity, the nature of which was also recorded (Shernoff et al., 2003; Uekawa et al., 2007; Yair, 2000). Measuring engagement via ESM, researchers were able to classify activities within broader types of instruction (e.g., lecture, group work, individual work, etc.), which in turn enabled them to describe engagement at the level of instruction type.

Finally, many measures mixed indicators at different levels and/or situations in a single dimension. For example, dimensions in some measures included items that referred to school or class in general (Miserandino, 1996; Skinner et al., 2008; M.-T. Wang et al., 2011). In addition to items about school or class in general, Lam et al. (2014) also incorporated items about learning in the affective dimension and items about homework and extracurricular activities in the cognitive dimension. In terms of subjects, Archambault et al. (2009) included questions about two subjects – separately for French and math – in their cognitive engagement subscale. Similarly, Fall and Roberts (2012) included questions about two subjects – separately for English and math – in their academic engagement subscale.

Qualitative differences between level-specific engagement. Some items can be applied to any level by simply changing a referent. For example, an item about student interest is applicable to a variety of levels, such as activity (Shernoff et al., 2003), class (Z. Wang et al., 2014), learning a subject (e.g., statistics, Whitney et al., 2019), specific aspect of a subject (e.g., knowing how to solve new mathematics problems, Kong et al., 2003), or school work (Maroco et al., 2016). As another example, an item about paying attention was used at the level of activity (Yair, 2000), class (Skinner et al., 2017), or classes in general (M.-T. Wang et al., 2011).

Some items can be applied across levels but are not applicable to all situations within levels. For example, an item about talking about a subject outside of class was included as an indicator of behavioral subject-specific engagement in the measure of M.-T. Wang et al. (2016). However, when subject-specific engagement was broken down into three situations – classroom, test, and extracurricular activities, – the item about talking about a subject outside of class was used to indicate extracurrucular subject-specific engagement (Lau & Roeser, 2002). This item is not applicable to classroom or test engagement.

Some levels may incorporate items that are specific or non-specific to level variations. For example, some items in the measure of Kong et al. (2003) included mathematics-specific language, such as problem solving. Yet, other items did not include such language. As another example, at the level of instruction type, interest, enjoyment, and concentration apply to any instruction type (Shernoff et al., 2003). Yet, some ways of engagement are specific to particular instruction types. For instance, working together on a task may be an indicator of engagement in group work, whereas volunteering to answer the instructor's questions may be an indicator of engagement in whole-class interaction. Other ways are specific to some but not all instruction types. For example, listening is applicable to engagement in lecture, whole-class interaction, and group work but not to engagement in individual work.

In sum, some measures may include items that can be applied at any levels or in any situations within levels. Yet, other measures may include items, some or all of which are specific to particular levels or to particular situations within levels. Thus, there may be qualitative differences in level-specific engagement measures.

Quantitative differences between level-specific engagement. Most instruments measured student engagement at a single level. However, when engagement was measured in multiple situations within levels, quantitative differences in engagement levels occurred. For example, Lau and Roeser (2002) found low-to-moderate correlations between test, classroom, and extracurricular engagement. Mazer (2012) found low-to-moderate correlations between in-class and out-of-class behaviors.

Considering level variations at the level of a subject, mean differences in actual or predicted engagement were observed across subjects. Marks (2000) found that elementary and high school students were more engaged in a math class compared to a social sciences class; yet, there were no differences for middle school students. When engagement was measured via ESM, Yair (2000) found that students were more likely to be engaged in math classes than in English, foreign language, or social science classes; however, no difference in the likelihood of engagement was observed for reading and science classes when compared with math. Shernoff et al. (2003) explored mean differences in engagement between the following subjects: math, English, science, foreign language, history, social studies, computer science, art, and vocational education. They found that students were more engaged in arts than in math, English, science, foreign language, history, social studies, and vocational education. Further, students were more engaged in computer science than in math, English, science, foreign language, and history. Finally, students were more engaged in vocational education than in math.

Considering level variations at the level of an instruction type, ESM measures provide mean differences in actual or predicted engagement across instruction types. Shernoff et al. (2003) found that students were more engaged during group and individual work than during the time listening to a lecture, watching TV/video, or taking exams. Yair (2000) reported that students were more likely to be engaged in discussion, work in labs, work in groups, individualized instruction, and TV/video but as likely in lecture or presentations than in the unknown type of instruction. Further, Uekawa et al. (2007) showed that students were more engaged in group work and less engaged in downtime, compared to lecture. However, these effects became statistically non-significant once content characteristics, students' perceptions of the material, and types of classroom conversations were controlled.

Low-to-moderate correlations and mean differences suggest that students may differ in their levels of engagement within a particular level of specificity (e.g., the level of instruction type). In sum, there may be quantitative differences in student engagement within levels of specificity, although evidence for these differences is currently limited.

Rationale for measuring instruction-specific engagement. The review of levels of specificity in engagement measurement showed that there are a variety of levels of specificity, from very broad (e.g., school) to very specific (e.g., activity). Broad levels, such as class or school, are most frequently used. In this study, I focused on undergraduate mathematics-based classes, as student success in such classes is critical for completion of university degrees, especially in STEM fields. Additionally, engagement was shown to be domain-specific in two ways. First, researchers found differences in student engagement levels across different classes (e.g., Marks, 2000; Yair, 2000). Second, domain-specific measures of engagement have the capability to capture domain-specific engagement more precisely by including domain-specific language in the items, such as items about problem solving in math-specific measures (e.g., Kong et al., 2003; Miller et al., 1996; Rimm-Kaufman et al., 2015; M.-T. Wang et al., 2016).

Engagement at the class level can be further broken down. Specifically, measuring engagement at the level of instruction type provides more information about students' levels of engagement in a particular undergraduate mathematics-based class. Instruction-specific measurement allows educators to capture qualitatively different engagement in different instruction types. Further, instruction-specific measurement provides information about quantitative differences in engagement levels of a particular student across instruction types. Yet, instruction-specific engagement is broad enough to be used as a target of instructional interventions. Thus, the instruction-specific approach provides an opportunity to develop more strategic, instruction-specific engagement interventions.

Threats to validity in class-level measurement. Instruction-specific measurement is not limited to producing instruction-specific scores. It can also produce class-level scores when engagement scores at the class level are of interest to researchers. In fact, instruction-specific measurement may produce more valid class-level scores than class-level measurement because instruction-specific measurement may help to avoid two threats to validity: construct-irrelevant variance and construct underrepresentation.

Construct irrelevant variance. Construct-irrelevant variance at the class level occurs when class-level measures ask about situations that are not applicable to the class, engagement in which is being measured. In general, when engagement is measured at the class level, while some situations might be safe to assume to exist (e.g., a situation of solving problems when measuring engagement in a math class, Rimm-Kaufman et al., 2015), other situations might be not. For example, Lam et al. (2014) and Miller et al. (1996) asked questions with respect to students' homework. Ing and Victorino (2016) asked about class discussions. Z. Wang et al. (2014) asked about working with other students in class. However, it is possible that the student's class did not have homework, class discussions, or opportunities to work with classmates.

Construct underrepresentation. Construct underrepresentation at the class level occurs when class-level measures do not ask about situations relevant to the class, engagement in which is being measured. For example, Mazer (2012) asked about engagement in lecture and whole-class interaction but not about engagement in group work. Thus, the use of this measure in a class with group work would produce potentially misleading scores, as the construct of class engagement does not include group work engagement and, therefore, is underrepresented. As another example, Handelsman et al. (2005) asked about engagement in whole-class and small-group settings but not about engagement in individual work. Thus, the use of this measure in a class with individual work would produce potentially misleading scores, as the construct of class engagement does not include individual work engagement and, therefore, is underrepresented.

Rationale for determining class-level engagement from instruction-specific measurement. I found that some class-level measures of student engagement, which employed instruction-specific indicators, may suffer from the two threats to validity – construct-irrelevant variance and construct underrepresentation. Determining class-level engagement scores from instruction-specific measures may help avoid these threats to validity. First, instruction-specific measurement does not assume that instruction-specific items are applicable to a particular class. Instead, it allows for explicitly asking students about the instruction type applicability prior to administering instruction-specific measures. Second, instruction-specific measurement helps to ensure that engagement in all major instruction types is captured. In other words, it helps to avoid situations when engagement in instruction types, prominent in a particular class, are not captured. In sum, instruction-specific measurement allows for determining class-level engagement scores that are potentially more valid than class-level measurement. It should be noted that these threats to validity are not a concern for class-level engagement measures that use class-level items and not instruction-specific items. For example, these threats to validity are not applicable to measures that used such items as paying attention in class (e.g., Skinner et al., 2009), putting effort into learning, trying to connect what is being learned with prior knowledge (M.-T. Wang et al., 2016), perceiving a math lesson to be fun (Rimm-Kaufman et al., 2015), or feeling bored in class (Marks, 2000).

Relationships Between Engagement and Other Constructs

In this section, I describe relationships that student engagement has been shown to have with other constructs of educational interest. Specifically, I discuss predictors of student engagement and its outcomes. Predictors may include facilitators of student engagement, i.e., “explanatory causal factors, outside the target construct, that have the potential to influence the target” (Skinner & Pitzer, 2012, pp. 25–26). In the review of predictors below, I do not differentiate between facilitators and predictors, which do not necessarily have the causal potential. Outcomes, according to Skinner and Pitzer (2012), are “the results that engagement itself can produce” (p. 26). However, similarly to the review of predictors, in the review of outcomes below, I do not differentiate between outcomes that have the potential to be causal effects of student engagement and outcomes that are predicted but not necessarily caused by student engagement. Additionally, I also describe relationships between student engagement and other constructs where no direction of prediction or causality is implied, hypothesized, or theorized. Studies examined in this analysis include not only the studies analyzed in the previous sections but also the studies that used existing measures of student engagement.

Predictors. Predictors that are positively related to emotional engagement include involvement in academic (associated with major) and non-academic co-curricular activities (Wilson et al., 2014). In contrast, co-curricular hours were not found to be a significant predictor of emotional engagement (Wilson et al., 2014). In the study of Reeve and Tseng (2011), grade level was a negative predictor of agentic engagement. However, grade level did not appear to predict ESM-measured engagement in the final model of Yair (2000). Retention in secondary school was negatively associated with behavioral, cognitive, and affective engagement (Archambault et al., 2009). Initial mathematics achievement was a significant predictor of behavioral engagement in mathematics classes but not of cognitive, emotional, or social engagement (Rimm-Kaufman et al., 2015). Prior achievement was also not a significant predictor of engagement in the study by Marks (2000). General ability did not predict test, classroom, or extracurricular engagement in the final models by Lau and Roeser (2002). Engagement comprised of energy, absorption, and dedication was positively predicted by GPA; students on the vocational track were also more engaged than students on the upper secondary track (Salmela-Aro & Upadaya, 2012). GPA was also found to be a positive predictor of engagement in the study by Marks (2000). Finally, freshmen appeared to be more academically engaged than other undergraduate students (Gasiewski et al., 2012).

In terms of demographics, female students were more behaviorally engaged than male students when behavioral engagement was measured via student self-report (Reeve & Tseng, 2011; M.-T. Wang et al., 2011) or classroom observations (Rimm-Kaufman et al., 2015). No gender differences were found when behavioral engagement was measured via teacher report (Rimm-Kaufman et al., 2015). Further, female students also had higher cognitive and social engagement than male students, but female and male students did not differ in their emotional engagement (Rimm-Kaufman et al., 2015). Different findings were obtained by M.-T. Wang et al. (2011), who showed no gender differences in cognitive engagement but higher levels of emotional engagement for female students. Gender differences were also explored in the levels of test, classroom, and extracurricular engagement, with the only significant difference found for classroom engagement (female students were more engaged than male; Lau and Roeser, 2002). Additionally, female students were also more engaged than male students when engagement was conceptualized as a combination of energy, absorption, and dedication (Salmela-Aro & Upadaya, 2012). Looking at engagement by school level, Marks (2000) found that elementary and high school female students were less engaged than their male peers, but there were no gender differences for middle school students. Finally, no gender differences were observed for academic engagement (Gasiewski et al., 2012) or when engagement was measured via ESM (Yair, 2000).

In terms of race, European American students were more behaviorally and less emotionally engaged than African-American students but did not differ in cognitive engagement (M.-T. Wang et al., 2011). Further, no differences between White and non-White students were found for test, classroom, or extracurricular engagement (Lau & Roeser, 2002) or for academic engagement (Gasiewski et al., 2012). In the study by Marks (2000), Hispanic students did not differ from White students in their engagement, but African-American students were more engaged than White students in elementary and middle school (no differences were found for high school). When engagement was measured via ESM, African-American students appeared to be less engaged than Asian students, although no differences in comparison to Asian students were observed for Hispanic or White students (Yair, 2000). Further, student age was negatively associated with behavioral, cognitive, and affective engagement in the study of Archambault et al. (2009) but was not a significant predictor in the study (Rimm-Kaufman et al., 2015). In the latter study, age also did not predict social engagement. Free or reduced-price lunch was a positive predictor of emotional engagement but not of behavioral, cognitive, or social engagement (Rimm-Kaufman et al., 2015). Parental education was not a significant predictor of test, classroom, or extracurricular engagement (Lau & Roeser, 2002). SES was also not a significant predictor of engagement in the study by Marks (2000) or of ESM-measured engagement in the study by Yair (2000). Parental involvement positively predicted engagement in elementary and high schools but was not a significant predictor in middle school (Marks, 2000). Parent support was found to positively predict behavioral and academic engagement (Fall & Roberts, 2012).

Psychological needs satisfaction was positively related to all engagement dimensions under investigation: agentic, behavioral, cognitive, and emotional (Reeve & Tseng, 2011). Identification with school was a positive predictor of behavioral and academic engagement; perceived control was a positive predictor of academic engagement but was not found to predict behavioral engagement (Fall & Roberts, 2012). Further, self-efficacy was positively related to emotional (Rimm-Kaufman et al., 2015; Wilson et al., 2014), cognitive, social, teacher-reported but not observed behavioral engagement (Rimm-Kaufman et al., 2015). In the study of Liu et al. (2018), self-efficacy positively predicted behavioral, cognitive, and emotional engagement. Competence beliefs positively predicted test and classroom engagement but not extracurricular engagement (Lau & Roeser, 2002). Task values, in contrast, were a positive predictor of test, classroom, and extracurricular engagement (Lau & Roeser, 2002). Additionally, students' motivational goal orientations were also examined to predict student engagement (Lerdpornkulrat et al., 2018). Specifically, mastery goal orientation positively predicted behavioral, cognitive, and emotional engagement. Performance approach goal orientation did not predict any engagement dimensions. Lastly, performance avoidance goal orientation negatively predicted behavioral engagement but did not predict cognitive or emotional engagement. Further, engagement comprised of energy, absorption, and dedication was negatively predicted by burnout and depression, and positively predicted by self-esteem (Salmela-Aro & Upadaya, 2012). It was also positively predicted by study and personal recourses and was not related to study demands (Salmela-Aro & Upadyaya, 2014). Further, engagement within this framework was positively predicted by autonomy, competence, and relatedness, and negatively by neuroticism; yet, extraversion, openness, conscientiousness, and agreeableness were not found to be significant predictors (Sulea et al., 2015). Douglas et al. (2016) found that industriousness positively predicted vigor, dedication, and absorption. Openness was a positive predictor of vigor and dedication but not absorption, whereas intellect was a positive predictor of absorption but not vigor and dedication. Other personality aspects – politeness, compassion, orderliness, enthusiasm, assertiveness, volatility, and withdrawal – did not predict vigor, dedication, or absorption. Further, Marks (2000) found alienation to be a negative predictor of engagement. Finally, academic coping and friendship quality were shown to be positively related to student engagement, comprised of behavioral, cognitive, and emotional dimensions (Thien & Razak, 2013).

Teacher characteristics also appeared to predict student engagement. Teachers' highest degree earned was found to matter only for cognitive engagement but not for behavioral, emotional, or social (Rimm-Kaufman et al., 2015). Specifically, students taught by a teacher with a Master's degree were more cognitively engaged than their peers taught by a teacher with a Bachelor's degree. Teachers' years of experience, in contrast, was negatively associated with social engagement but was not associated with behavioral, cognitive, or emotional (Rimm-Kaufman et al., 2015). For college instructors, Gasiewski et al. (2012) examined whether a tenure status mattered for student academic engagement. They found that students taught by tenured instructors were more engaged than students taught by non-tenure track instructors. However, students taught by tenure-track but not yet tenured instructors did not differ in their engagement from their peers taught by non-tenure track instructors. Classroom characteristics, such as class size (large, medium, or small) or seat type (individual, lab, or roundtable) did not appear to matter for ESM-measured engagement (Uekawa et al., 2007). However, students were more engaged when their seats were chosen by the teacher as opposed to the students themselves (Uekawa et al., 2007).

Teacher support positively predicted behavioral, cognitive, and emotional engagement (Liu et al., 2018). It also positively predicted behavioral and academic engagement in the study of Fall and Roberts (2012). Marks (2000) found that both school and classroom support positively predicted engagement. Rimm-Kaufman et al. (2015) found concurrent emotional support and concurrent classroom organization were positively related to cognitive, emotional, and social engagement. In terms of behavioral engagement, a significant (positive) relationship was found only for concurrent classroom organization and only with observed behavioral engagement. Concurrent instructional support did not predict either behavioral, cognitive, emotional, or social engagement (Rimm-Kaufman et al., 2015). Environmental support was also a positive predictor of ESM-measured flow-based engagement in the study of Shernoff et al. (2016). Further, some perceptions of school environment were found to predict student engagement (M.-T. Wang & Holcombe, 2010). Performance goal structure negatively predicted behavioral and emotional engagement but positively predicted cognitive engagement. Mastery goal structure, on the other hand, positively predicted all engagement dimensions. Autonomy support was found to positively predict emotional but was not found to predict behavioral or cognitive engagement. Promotion of discussion was a positive predictor of cognitive and emotional engagement but did not predict behavioral engagement. Finally, teacher social support was a positive predictor of behavioral and emotional engagement but did not predict cognitive engagement. Further, the amount of time spent on different instruction types was also examined in relation to engagement. Gasiewski et al. (2012) found that the proportion of time devoted to class discussion and the proportion of time devoted to group work positively predicted student academic engagement. In contrast, the proportion of time devoted to lecture was negatively related to academic engagement.

Content aspects may also make a difference in student engagement. Uekawa et al. (2007) tested whether content difficulty and content newness predicted ESM-measured engagement. In their later model, content difficulty was negatively related to ESM-measured engagement, and content newness did not appear to matter at all. They also found that perceiving content to be relevant to everyday life was positively related to engagement; yet, relevance to college, job, or test did not appear to predict engagement in the final model. In terms of students' perceptions of teaching, Yair (2000) also found that students are more likely to be engaged when instruction is more relevant to their lives, more challenging, and more academically demanding. Yet, the level of skills required from a student did not appear to predict engagement. Environmental challenge, however, did not appear to predict ESM-measured flow-based engagement in the study of Shernoff et al. (2016).

Lastly, students' feelings and behaviors during class were also examined in relation to ESM-measured engagement. Uekawa et al. (2007) found that students were more engaged when they felt more cooperative, competitive, and had fun in class. In contrast, students were less engaged when they felt sleepy or were confused. Further, having an academic conversation with a teacher positively predicted engagement, whereas having a social conversation with classmates negatively predicted engagement. Yet, having a social conversation with a teacher or an academic conversation with classmates did not appear to predict engagement.

Outcomes. Student engagement has been shown to predict a variety of outcomes. Overall engagement positively predicted students' self-perceived academic achievement (Assunção et al., 2020; Maroco et al., 2016). Cognitive, emotional, and agentic engagement were positively related to students' overall semester grade, whereas behavioral engagement was not found to be a significant predictor (Reeve & Tseng, 2011). In a different study by Reeve (2013), however, emotional and cognitive engagement did not significantly predict student course grades, but behavioral and agentic engagement did. Whitney et al. (2019) found that cognitive and affective engagement positively predicted statistics proficiency, whereas behavioral engagement did not appear to predict statistics proficiency. M.-T. Wang et al. (2016) found that student-reported behavioral engagement positively predicted math and science course grades, whereas teacher-reported behavioral engagement positively predicted science course grades but not math course grades. Student-reported social engagement negatively predicted math and science course grades, whereas teacher-reported social engagement negatively predicted math course grades but not science course grades. Notably, these results for social engagement from the tested model differ from the results of zero-order correlations, where social engagement was positively associated with course grades (with the exception of a non-significant correlation between student-reported social engagement and science course grade). Finally, emotional and cognitive engagement, either student- or teacher-reported, did not predict math or science grades. Fall and Roberts (2012) found that behavioral and academic engagement positively predicted achievement in mathematics and reading. Lee (2014) found that behavioral and emotional engagement positively predicted reading literacy. In the study by Lau and Roeser (2002), test engagement was a positive predictor of science test scores but did not appear to predict science second-semester grades. In contrast, classroom and extracurricular engagement did not predict science test scores. However, science second-semester grades were positively predicted by classroom engagement and negatively by extracurricular engagement. When no framework for measurement was used, engagement was shown to be a positive predictor of GPA (Ing & Victorino, 2016). M.-T. Wang and Holcombe (2010) found that behavioral, cognitive, and emotional dimensions of engagement also positively predicted GPA. Finally, overall engagement was a negative predictor of the number of failing courses (Maroco et al., 2016).

Besides achievement, student engagement was examined to predict school dropout. Archambault et al. (2009) found that behavioral engagement was a negative predictor of school dropout, whereas cognitive and emotional engagement did not predict school dropout. Fall and Roberts (2012) found that behavioral and academic engagement negatively predicted dropping out. In contrast, in the study of Reschly and Christenson (2006), dropout was positively predicted by cognitive, psychological, and behavioral engagement (except for the homework component of behavioral engagement, which was not a significant predictor, and the extracurricular activities component, which negatively predicted dropout). Dropout was also explored by Finn and Zimmer (2012) who tested whether engagement in Grade 4 and in Grade 8 predicted high school graduation. They found that Grade 4 academic engagement was a positive predictor of high school graduation, but Grade 4 social engagement was not a significant predictor. However, both academic and social engagement in Grade 8 positively predicted high school graduation, but affective engagement was not a significant predictor. Finally, when intention to drop out of a university was considered as an outcome, overall engagement was a negative predictor (Assunção et al., 2020; Maroco et al., 2016). Among behavioral, cognitive, and emotional engagement, only emotional engagement was found to be a positive predictor of the intention to persist in college, stable across tested models (Lerdpornkulrat et al., 2018).

Further, engagement was tested in relation to STEM career aspirations (M.-T. Wang et al., 2016). Specifically, STEM career aspirations were negatively predicted by student-reported behavioral engagement in math and science, positively predicted by teacher-reported behavioral engagement in math, and were not predicted by teacher-reported behavioral engagement in science. Cognitive engagement in math negatively predicted STEM career aspirations when measured by student report but was not a significant predictor when measured by teacher report; cognitive engagement in science was not a significant predictor of STEM career aspirations. Social engagement in math or science was also not a significant predictor of STEM career aspirations. The negative effects were interesting when predicting STEM career aspirations, considering that all zero-order correlations between engagement dimensions and STEM career aspirations were positive. Lau and Roeser (2002) studied a similar phenomenon – anticipated choices of science majors or careers – and found that extracurricular engagement positively predicted these choices but test and classroom engagement did not appear to predict them. Finally, engagement comprised of energy, absorption, and dedication was a positive predictor of life satisfaction (Salmela-Aro & Upadyaya, 2014). Similarly, student engagement, comprised of behavioral, cognitive, and emotional dimensions, was a positive predictor quality of school life (Thien & Razak, 2013). Handelsman et al. (2005) also found that emotional engagement was a positive predictor of the belief in incremental theory.

Other relationships. Self-efficacy, mastery goals, performance-approach-goals, performance-avoidance goals, school-prompted interest, and self-reported grades were positively related to behavioral, cognitive, and affective engagement, and negatively related to disengagement (Z. Wang et al., 2014). Self-efficacy and self-reported grades, as well as task value, were also positively related to behavioral engagement in the study of Hospel et al. (2016). Relationships between student engagement and personality traits were also examined (Qureshi et al., 2016). Behavioral, cognitive, and emotional engagement were positively related to agreeableness, conscientiousness, and imagination. Extraversion was found to be positively correlated only to behavioral engagement. Finally, emotional stability was not found to be related to any engagement dimension. Next, overall engagement was negatively related to burnout (Assunção et al., 2020). Further, student- and teacher- reported teacher support, as well as perceived control, autonomy, and relatedness had positive relationships with behavioral and emotional engagement, and negative relationships with behavioral and emotional disaffection (Skinner et al., 2008). Similar relationships with perceived control, autonomy, and relatedness were also found in the study of Skinner et al. (2017). Further, behavioral and emotional engagement were positively related to science identity, science career plans, purpose in science, and science course grades; in contrast, behavioral and emotional disaffection were negatively related to these constructs (Skinner et al., 2017). Behavioral, cognitive, and emotional engagement were also positively related to understanding of and attitudes to employability (Qureshi et al., 2016). In terms of teaching characteristics, student-reported whole-class activities and teachers' questioning practice had a positive relationship with behavioral, cognitive, and affective engagement, and a negative relationship with disengagement (Z. Wang et al., 2014). Additionally, learning climate, efficient classroom management, clarity of instruction, activating teaching, differentiation, and teaching learning strategies were positively related to behavioral and emotional engagement (Inda-Caro et al., 2018).

Conclusion. The review of relationships of student engagement with its predictors and outcomes showed that student engagement is related to a variety of constructs. In particular, student engagement has the potential to affect a number of important educational outcomes, such as achievement and dropout. Yet, student engagement is also a malleable construct (Fredricks et al., 2004). Thus, efforts to improve student learning outcomes may focus on improving student engagement. Further, student engagement has also been shown to be related to a number of potential facilitators that can be manipulated. Of a particular interest to educators are instructional factors that teachers can manipulate in order to affect student engagement. Thus, student engagement may be an important construct through which educators can increase student learning outcomes.

Proposed Measure

In the literature review, I demonstrated the importance of student engagement for learning outcomes, including the outcomes in STEM disciplines. To help students become more engaged, we need to improve our measurement of engagement. In the discussion of dimensionality and specificity, I showed potential benefits of multidimensional and instruction-specific measurement of engagement. Combining multidimensionality and instructional specificity will allow educators to identify both how (the dimension) or where (the instruction type) students are not engaged. This information will enable educators to develop more targeted and, therefore, more efficient instructional interventions than those that target only different engagement dimensions or only engagement in different instruction types. Further, combining multidimensionality and instructional specificity in engagement measurement will allow researchers to inform the development of such interventions. Specifically, researchers will be able to identify instructional factors that affect different dimensions of student engagement in different instruction types and to understand how these factors exert their influence (e.g., directly or indirectly through, for example, personal factors). Researchers are likely to be able to identify such factors, as the literature review showed that student engagement has a number of potential facilitators. An additional benefit of instruction-specific measurement of engagement dimensions is its potential ability to produce more valid dimensional scores at the class level. Thus, I aimed to develop an instrument that measures multidimensional, instruction-specific engagement in undergraduate mathematic-based classes.

Selecting dimensions. The discussion of indicators above suggests that the three-dimensional framework of Fredricks et al. (2004) is most commonly used to measure student engagement. This framework is also comprehensive and flexible, as it has the capability to capture different aspects of engagement via a number of different ways (i.e., by applying a broad range of indicators). However, such flexibility also has its drawbacks for measurement. First, as there are no specific, agreed upon, conceptualizations of engagement dimensions, there are also no clear-cut lines between them, resulting in the use of the same or very similar indicators for different dimensions in different measures. For example, listening and paying attention, as well as effort and persistence, were used to indicate both behavioral and cognitive engagement. Boredom, as well as similar constructs of identification with school and school belonging, were used to indicate both cognitive and emotional engagement. Involvement in class work, in turn, was used to indicate both behavioral and emotional engagement. Second, the range of indicators became broad to the extent that facilitators and outcomes appeared among indicators (see also Sinclair et al., 2003; Skinner & Pitzer, 2012). For example, constructs, such as effort, self-regulation, or interest may be facilitators of engagement, whereas performance is more typically used as an outcome of engagement. Including constructs that could be facilitators or outcomes of engagement as indicators makes an examination of their relationships with engagement impossible.

The three dimensions of Fredricks et al. (2004) – behavioral, cognitive, and emotional – can also be applied to a variety of instruction types. In contrast, social and agentic dimensions may already incorporate instructional specificity. In particular, social engagement may incorporate instructional specificity where students interact with each other in whole-class or small group settings (i.e., whole-class interaction or group work). Agentic engagement may incorporate instructional specificity where students interact with their instructor (i.e., whole-class interaction). Thus, these dimensions cannot be applied to all instruction types. Further, dimensions of Schaufeli et al. (2002; vigor, dedication, and absorption) are relatively narrow in scope and do not fully capture the complex construct of engagement. Finally, dimensions of disaffection were also not used, as I aimed to measure engagement rather than disaffection. Future research may want to take a step further and conceptualize instruction-specific dimensions of disaffection. Thus, due to the comprehensiveness and flexibility of the three-dimensional framework of Fredricks et al. (2004), as well as due to its applicability to instruction types, I adopted this framework. However, I conceptualized engagement dimensions in a way that (1) makes the distinction between the dimensions clear, and (2) excludes constructs that could and should be separate from engagement. Thus, broadly speaking, behavioral engagement in this study was conceptualized as students' behaviors, cognitive engagement as students' cognitive processes, and emotional engagement as students' emotions.

Selecting instruction types. Prior research classified class time in several categories. For example, Shernoff et al. (2003), using ESM to measure engagement, classified the most frequent activities in five categories: individual work, listening to a lecture, taking exams, watching TV, films, or videos, and group work, which also included lab activities. Similarly, classifications of ESM-measured engagement in the study of Yair (2000) included lecture, group work, individualized work, watching television and video presentations; however, Yair (2000) separated lab work and added classroom presentations and class discussions. Uekawa et al. (2007) also used ESM to measure engagement but classified activities into four categories, three of which overlap with the categories of Shernoff et al. (2003) and Yair (2000): whole-class instruction (i.e., lecturing time), individual work (i.e., independent seat work), and group work (e.g., solving problems or doing experiments together). The fourth category was downtime (i.e., doing nothing). Other studies employed categories of instruction types when measured the composition of class instruction. Walter et al. (2016) measured the percentage of instruction spent on lecture, small group work, individualized instruction, and other forms of instruction. Gasiewski et al. (2012) also measured percentages of instruction but distinguished between three instruction types: lecture, class discussion, and group work.

Selecting instruction types for my study, I used the categories of instruction types, specified in prior research, as a guidance. First, I excluded testing and downtime from consideration because my instrument focuses on student engagement during instructional time. Second, to ensure that categories of instructional time, used by other researchers, are collectively exhaustive, I specified instruction types based on two characteristics: a focus of instruction (instructor vs. students) and a type of interaction during the instruction. Thus, I distinguished between four instruction types: lecture (instructor-focused, no interaction), whole-class interaction (instructor-focused, interaction between the instructor and students), group work (student-focused, interaction between students), and individual work (student-focused, no interaction).

Gap in the literature. To my knowledge, the three-dimensional framework of Fredricks et al. (2004) and instruction-specific measurement of engagement have not been applied together. Existing measures of behavioral, cognitive, and emotional engagement tend to be class-specific or even more general and, therefore, are not able to capture instruction-specific variations. In contrast, studies that explored student engagement across instruction types did not measure all important dimensions of engagement. Thus, in this study, I aimed to develop and validate a self-report instrument that measures student behavioral, cognitive, and emotional engagement in lecture, whole-class interaction, group work, and individual work in an undergraduate mathematics-based classes. Through the literature review in the next section, I developed more precise conceptualizations of behavioral, cognitive, and emotional engagement dimensions. This literature review also served as a source of potential indicators of each dimension in the proposed instrument. Conceptualizations of instruction types were produced during the instrument development process.

Conceptualizing Engagement Dimensions

In the previous section, I broadly conceptualized behavioral engagement as students' behaviors, cognitive engagement as students' cognitive processes, and emotional engagement as students' emotions. In this section, I review specific behaviors, cognitive processes, and emotions that students can experience in a mathematics-based classroom. The goals of this review are (1) to develop more precise conceptualizations of engagement dimensions and (2) to identify behaviors, cognitive processes, and emotions that can serve as a basis for item development. The first goal is addressed in this literature review, and the second goal is addressed in Chapter Three. The examined literature includes the work on behavioral, cognitive, and emotional engagement, as well as, more broadly, on behaviors, cognition, and emotions as applied to a classroom setting.

Behavioral engagement. Educational researchers frequently include a general category of participation in measures of behavioral engagement. Participation items typically ask about participation in class (M.-T. Wang et al., 2016), class discussions (Kong et al., 2003; Mazer, 2012; Miserandino, 1996; Skinner et al., 2008), class activities (Lam et al., 2014), or small group discussions (Handelsman et al., 2005). A variation of participation items is an item asking about being an active student in class (Gunuc & Kuzu, 2015). In other cases, researchers include more specific verbal (or oral) and nonverbal (or silent) behaviors as indicators of behavioral engagement (e.g., Mazer, 2012). In this section, I discuss these behaviors as well as their classifications. I conclude with the conceptualization of behavioral engagement for my instrument.

Verbal behaviors. In-class verbal behaviors can take place either within the whole-class or small group context. Further, students can verbally interact either with a teacher or with other students. In measures of behavioral engagement, some researchers use general interaction behaviors, such as talking to other students (Rimm-Kaufman et al., 2015), trying to work with others (M.-T. Wang et al., 2016), and interacting with faculty (Ing & Victorino, 2016). Bennett and Dunne (1991) proposed a classification of student talk in small groups. According to this classification, all talk can be separated into non-task related and task related talk. The latter can be further split into not directly relevant talk and directly relevant talk. Among the directly relevant talk, the authors also differentiated between cognitively-oriented talk (i.e., talk related to the cognitive demand of the task) and socially-oriented talk (i.e., talk that is concerned with the management of the group). These categories can also be applied to the whole-class context. Below, I discuss behaviors that may fit into the categories of this framework.

Cognitively-oriented talk. In general, regardless of the context and focus of interaction, cognitively-oriented talk may include sharing ideas, thoughts, or information (Hospel et al., 2016; Ing & Victorino, 2016; Mazer, 2012; McCrone, 2005; Rimm-Kaufman et al., 2015; Shachar & Sharan, 1994; M.-T. Wang et al., 2016; Webb, 1980), offering comments, expressing opinions (Fassinger, 1995), and making suggestions (Hospel et al., 2016). Specific to a task or activity, students may hypothesize (Wong et al., 2002), reason, justify (McCrone, 2005; Wong et al., 2002), clarify (Wong et al., 2002), explain (Kosko, 2014; Shachar & Sharan, 1994; Webb, 1980, 1984), or elaborate (Gillies & Khan, 2009). They may also criticize, positively evaluate, and correct (Webb, 1980). The classification of Shachar and Sharan (1994), developed based on cognitive-intellectual features in the students' speech, adds more specific categories of talk to this list. For example, one category of Shachar and Sharan (1994) is an unstructured idea, which is similar to thinking out loud. The authors described it as "an association expressed verbally that is not organized syntactically" (p. 327). It can take a form of a short or disjoint sentence or a series of sentences. Another category is repetitions that are "almost verbatim restatement of one's own or of someone else's comment" (p. 327). Similar behaviors to repetitions are restating (Wong et al., 2002) and revoicing (Otten et al., 2011). When a student restates not verbatim but in their own words, the behavior is referred to as paraphrasing (Rosenzweig et al., 2011). A separate but related category of Shachar and Sharan (1994) is repetition with expansion. Here, a speaker repeats someone else's ideas but adds a new idea or a connecting link between the ideas. The next category of Shachar and Sharan (1994) is a generalization, i.e., an abstract principle that is formulated on the basis of statements and is also applicable beyond the task at hand. Other categories include providing concrete examples and organizing ideas (e.g., relating new comments with those made earlier).

Further, students, participating in a cognitively-oriented talk, also have different and changing roles during the interaction. Specifically, they may be asking or responding to others. When asking, students may be posing questions or making requests. In particular, they may be requesting information, explanation, or clarification (Gillies & Khan, 2009; Shachar & Sharan, 1994; Webb, 1980, 1982), or asking for help more generally (Webb, 1984). When responding, students may be answering questions (Fassinger, 1995; Hospel et al., 2016) or addressing requests by giving the requested information, explanation, or clarification (Shachar & Sharan, 1994). They can also be providing the requested help (Handelsman et al., 2005; Ing & Victorino, 2016; Rimm-Kaufman et al., 2015; M.-T. Wang et al., 2016; Webb, 1982, 1984). Further, some researchers distinguish between the types of responses. For example, a response can be classified as a short answer, an extended response, an answer with full justification (McCrone, 2005), or as a correct or incorrect answer (Webb et al., 2008). Explanations, in turn, can be classified as correct and complete, ambiguous and incomplete, or incorrect (Webb et al., 2008). Another way of classifying responses includes agreements and disagreements (Shachar & Sharan, 1994).

Socially-oriented talk. As mentioned above, socially-oriented talk is concerned with management. In the framework of Bennett and Dunne (1991), the object of management is the group. However, extending the framework, the object of management may also be the class more broadly. In the small group context, a commonly used management category is giving directives, which may include verbal instructions (Gillies, 2004), directions (Gillies & Khan, 2009), or comments about procedures for the conduct of group discussion (Shachar & Sharan, 1994). Gillies (2004) also noted a separate category of directives with a physical prompt, i.e., verbal instructions with hand gestures. Behaviors concerning efforts to refocusing may also be considered as socially-oriented.

For example, a student may refocus the discussion when it goes off-track (Shachar & Sharan, 1994) or may discipline another student to refocus his/her attention (Gillies & Khan, 2009).

Further, some behaviors, described earlier as cognitively-oriented talk, may also be socially-oriented if the behavior is concerned with management rather than cognition. One example is the behavior of asking questions. This behavior can be classified as cognitively-oriented talk if a student asks a question about the task. However, it can be classified as socially-oriented talk if a student asks a question about group conduct or classroom norms. Other examples of such double-natured behaviors are expressing opinions, offering comments, and making suggestions. In the engagement or participation research, such behaviors, when cognitively-oriented, may be used to indicate behavioral engagement (e.g., Handelsman et al., 2005; Hospel, Galand, & Janosz, 2016). When socially-oriented, such behaviors may be used to indicate agentic engagement (Reeve, 2013; Reeve & Tseng, 2011). Yet, sometimes, the orientation of these behaviors is not specified. For instance, the behavior of asking questions (e.g., Fassinger, 1995) can refer to any kind of questions, cognitively-oriented, socially-oriented, or both.

Non-task related talk and not directly relevant talk. In engagement research, behaviors that may be classified as non-task related talk are typically disruptive to instruction. In engagement measures, verbal disruptive behaviors include chatting in a loud voice (Hospel et al., 2016) and being rude to the teacher (Archambault et al., 2009). In a mathematics class, for example, any nonmathematical contributions to a class discussion (McCrone, 2005) may be considered as non-task related talk. Not directly relevant talk, according to Bennett and Dunne (1991), is concerned with the acquisition or manipulation of materials needed to complete the task (e.g., finding paper or sharpening pencils). However, this kind of talk is not typical for engagement research.

Nonverbal behaviors. Nonverbal behaviors, similarly to verbal, can be classified into task related and non-task related behaviors. Task related nonverbal behaviors that are typically used in measures of behavioral engagement, include listening (Gunuc & Kuzu, 2015; Kong et al., 2003; Mazer, 2012; Miserandino, 1996; Skinner et al., 2008) and related behaviors of paying attention (Lam et al., 2014; Mazer, 2012; Miserandino, 1996; Skinner et al., 2008; M.-T. Wang et al., 2016) and concentrating (Kong et al., 2003). Another category of task related non-verbal behaviors is reading, which can be applied to different instruction types. For example, in lecture, students may read the instructor’s notes or presentation slides. During problem solving, students read (Rosenzweig et al., 2011) and re-read (Wong et al., 2002) a problem in question. Reading is also involved in checking and reviewing the solution (Rosenzweig et al., 2011; Wong et al., 2002).

The next category of nonverbal task related behaviors is writing. During listening, students may take notes on what is being said (Canpolat et al., 2015; Imhof, 1998). As part of problem solving, students may draw diagrams and do calculations by hand (Rosenzweig et al., 2011; Wong et al., 2002). Finally, nonverbal task related behaviors can be psychomotor. In the engagement research, a behavior of raising a hand is sometimes used (Handelsman et al., 2005). Other behaviors in this category are concerned with maintaining attentive posture and giving a nonverbal response (Ford et al., 2000). These behaviors, emphasized primarily in the research on listening, may include making eye contact (Canpolat et al., 2015; Cooper & Buchanan, 2010; Fontana et al., 2015; Ford et al., 2000; Imhof, 1998), utilizing methods of non-verbal communication (e.g., nodding; Canpolat et al., 2015; Cooper & Buchanan, 2010; Fontana et al., 2015), sitting up straight, following along with both head and eyes (Canpolat et al., 2015), and watching speaker's body language (Imhof, 1998), such as paying attention to gestures, facial expressions, tone of voice, and stresses in speech (Canpolat et al., 2015).

Non-task related nonverbal behaviors, similarly to non-task related verbal behaviors, can include disruptive behaviors. An example of nonverbal disruptive behaviors is throwing things in the air (Hospel et al., 2016). Nonverbal non-disruptive behaviors, in turn, can include pretending to work (Hospel et al., 2016; Lam et al., 2014; Skinner et al., 2008) and lying over the chair or desk (Hospel et al., 2016). Finally, not working (Hospel et al., 2016) and doing something else (Canpolat et al., 2015; M.-T. Wang et al., 2016) can also be classified as non-task related nonverbal behaviors.

My conceptualization of behavioral engagement. For my instrument, I partially adopt the framework of Bennett and Dunne (1991). First, I extend their classification of student talk to also include nonverbal behaviors. Second, due to the focus of my instrument on engagement, I include only task related behaviors in the items; non-task related behaviors are not included as they would indicate disaffection rather than engagement. Third, to distinguish behaviors from cognition, behaviors need to be observable. Lastly, while there are numerous behaviors in which a student can be engaged during class, I focus specifically on those that are expected of a student within each type of instruction. Thus, I conceptualize behavioral engagement as students' expected observable on-task behaviors, including both verbal and non-verbal.

Cognitive engagement. In this section, I identify cognitive processes that occur in students' minds during listening and problem solving, arguably the two major student behaviors in undergraduate mathematics-based classrooms. I start the section by describing two major theories that constitute the conceptual framework for cognitive processes under investigation: the information processing theory and the schema-based theories. Then, I describe the cognitive processes involved in listening and problem solving within this framework. I conclude with the conceptualization of cognitive engagement for my instrument.

Conceptual framework. The literature on cognition in listening and problem solving is situated largely within the two theories: the information processing theory and the schema-based theories. I describe these theories below.

Information processing theory. As summarized by Mayer (2012), the information processing view consists of two main elements: (1) mental representations that are formed by the human mind and (2) cognitive processes that a learner applies to mental representations. Two versions of the information processing view exist: the classic (or information acquisition) view and the constructivist (or knowledge construction) view. The former represents learning as simply adding information to the long-term memory; the latter represents learning as developing a cognitive structure in the working memory by combining the incoming information and existing knowledge. The constructivist view tends to be supported by the modern education research community as it is aligned with the current understanding of the nature of learning. Specifically, learning nowadays is commonly understood as a personal and constructive activity occurring through integrating incoming information with the existing knowledge. The constructivist view of information processing has its roots in, among others, the work of Piaget. Thus, when referring to the information processing view throughout this paper, I will be referring to the constructivist view.

The information processing theory can be presented in the following model (see Mayer, 2012). Information from the outside is received by the learner's ears or eyes and stored in the sensory memory in the form of an exact sensory copy of the received information for a very short period of time (less than a second). Then, the learner selects particular information and transfers it to the working memory for further processing. There, the information is held for a short period of time (less than 30 seconds) and is organized into coherent mental representations. Next, the learner activates relevant knowledge in the long-term memory, i.e., a memory that permanently stores an unlimited amount of information. Then, this knowledge is transferred from the long-term memory to the working memory and integrated with the current mental representations, leading to their modifications. After that, the constructed knowledge is encoded and stored in the long-term memory. The three cognitive processes – selecting, organizing, and integrating – can serve as overarching categories for more specific cognitive processes.

Schema-based theories. Mental representations mentioned in the previous section are essentially contextualized schemas (Derry, 1996), which are central to investigations in the schema-based research. This research examines how schemas are constructed and revised but does so within several theoretical perspectives: the schema theory (or the cognitive schema theory) and the constructivist view, specifically the radical or schema-driven constructivism. I describe each view in more detail below. I also discuss the nature of schemas, as well as their classification and functions.

The schema theory (or the cognitive schema theory) is a version of the information processing theory (Derry, 1996). The purpose of this theory is to "identify cognitive mechanisms that underlie schema construction and revision" (p. 167), where schemas refer to any memory structure. Research within this view was largely influenced by Bartlett (1932), who studied memory, particularly how and what people remember from stories (Marshall, 1995). For Bartlett, schemas were a means to explain how and why people distort and ignore aspects of a new experience (Marshall, 1995). The second view is the constructivist view, specifically the radical or schema-driven constructivism, which suggests that "all new logical-mathematical and conceptual understanding is constructed on the basis of previously constructed schemes" (Derry, 1996, p. 165). Within this view, schemas tend to be understood from a Piagetian perspective (Derry, 1996). Piaget (1952), who studied the development of reasoning in children, viewed schemas as a means of making sense of the environment (Marshall, 1995). Piagetian schemas represent the "big ideas" that underlie students' constructed understanding of mathematics and science (Derry, 1996). To explain knowledge construction within his theory, Piaget introduced two principles: assimilation and accommodation. Assimilation refers to the incorporation of new experiences into the schema, and accommodation refers to the modification of the schema if a new experience does not fit the current schema.

Synthesizing Barrett's and Piaget's research, Marshall (1995) concluded that although the emphases of their investigations were different, both theorists thought of a schema in a similar way. Indeed, Bartlett was concerned with the influence of schemas on retrieval or recall, and Piaget was concerned with exploring how schemas develop (Marshall, 1995). However, both theorists viewed a schema as "a memory structure that develops from an individual's experiences and guides the individual's response to the environment" (Marshall, 1995, p. 15). Moreover, the theoretical views on schemas – the schema theory and the constructivist view – may not be as different as they may seem. Derry (1996) compared the two views and concluded that the differences between them may be in the terminology rather than in the actual meaning.

Current conceptualizations of a schema are essentially variations of memory structures of Bartlett and Piaget, such as a conceptual structure (Skemp, 1987), a conceptual frame (Weinberg et al., 2014), or a knowledge cluster (Thorndyke, 1984). Essentially, a schema is an abstraction of a particular phenomenon developed and adjusted through encountering this phenomenon multiple times (Thorndyke, 1984). Being abstract, it permits a person to recognize and organize new information or experience by "filling in" general parts of the schema with specific information from the encountered phenomenon. Further, the abstract nature of a schema enables an individual to infer from and reason based on the incomplete data. Finally, schemas are hierarchically organized.

One example of a schema hierarchy is described by Derry (1996). This author differentiates between the three classes of schemas: memory objects, mental models, and cognitive fields. According to Derry (1996), memory objects are basic components of knowledge permanently stored in memory. They are formed by combining different types of representations (such as pictorial, declarative, procedural, auditory, etc.) and can be used to interpret new events. Within memory objects, Derry (1996) distinguished between three schema types of different order. The simplest schemas that represent minimal abstractions of events are phenomenological primitives, or p-prisms (diSessa, 1993). Above them are more integrated kinds of memory objects that, as stated by Derry (1996), reflect the meaning of schemas described by Sweller and Cooper (1985). These schemas are complex and structured; they allow people to recognize encountered patterns and classify them within the existing knowledge so that appropriate responses can be generated. Next memory-object schemas are object families, which Derry (1996) describes as loosely organized collections of ideas that activate each other. Object families behave as a single memory object.

Another class of schemas, presented by Derry (1996), are mental models. A mental model is an organization of memory objects that represent a mental representation, or interpretation, of a phenomenon. This mental representation is constructed via mapping activated memory objects onto the components of the phenomenon, followed by reorganizing and connecting these objects. Importantly, mental models as understandings of particular situations do not exist outside of these situations. Finally, the last class of schemas is cognitive fields – preconceptions that are activated during the mental modeling process in response to a particular phenomenon prior to their organization into a mental representation.

Schemas have multiple functions but of particular relevance to mathematics learning are the two uses of schema identified by Thorndyke (1984). First, schemas serve as a structure for acquiring new knowledge. Specifically, when new information is encountered, it activates schemas in the memory that help a student organize the information into a mental representation and integrate it back in memory. Second, schemas are useful in problem solving. Here, a problem activates particular schemas that may help to solve it. A student "fills in" the schema with details from the problem to determine a solution. Thus, it is not surprising that the notion of schemas are commonly used in the research on listening and problem solving that I discuss next.

Cognitive processes during listening. Listening is "the process of receiving, constructing meaning from, and responding to spoken and non-verbal messages" ("An ILA Definition of Listening," 1995, p. 4). In this section, I focus particularly on the cognitive side of listening and identify cognitive processes that occur during listening. From the view of cognitive psychology, listening is often conceptualized within the information processing framework (Imhof, 2010), with some researchers using the terminology of schemas (e.g., Wolvin, 2010). This conceptualization allows for listening to be defined as the process of selecting, organizing, and integrating information (Imhof, 2010). Thus, I will use these categories to classify the identified cognitive processes. A note needs to be made that in practice the separation of the organizing and integrating processes may not be as definite as it is in theory. A primary reason for the boundaries blurring between these processes is the reciprocal relationship between them. Indeed, mental representations are constructed and adjusted through the use of abstract schemas retrieved from the long-term memory, and the abstract schemas are modified or expanded when mental representations are integrated.

Selecting. In the framework of the information processing theory, selecting in its broad sense refers to paying attention to the incoming information held in the sensory memory so that this information can be further processed in the working memory (Mayer, 2012). Differences within the selection processes are primarily attributed to what is selected. Many researchers specify selection of main, substantial, relevant, or most important points, ideas, or other pieces of information over minor, trivial, irrelevant, or less important ones (e.g., Aryadoust et al., 2012; Dermitzaki et al., 2009; Halone et al., 1998; Vermunt, 1996; Weinstein et al., 2016).

Researchers also consider the cognitive process of distinguishing major points or ideas from details, examples, or supporting information (Powers, 1986; Richards, 1983). Additionally, Schmeck et al. (1977) noted the process of differentiating between similar ideas. Other subjects of attention may be sequences, cause and effect relationships, or parts that have practical utility (Al-Musalli, 2015; Vermunt, 1998). Further, Richards (1983) identified the selection processes that are not directly related to the lecture content but nevertheless important for listening comprehension. These processes include recognizing functions of non-verbal cues as markers of emphasis and attitude, instructional/learner tasks (e.g., warnings, suggestions, recommendations, advice, instructions), a function of intonation to signal information structure (e.g., pace, pitch, volume, key), markers of cohesion, and key lexical items related to subject/topic.

Notably, distinguishing between particular parts of content or aspects of the class does not necessarily mean that some parts would be selected and others would not. In other words, a difference exists between the differentiation process and the selection process with the former enabling the latter. For instance, a student may distinguish between the main idea and its supporting information but may decide to select both for the further processing. Thus, selection is affected by a student's differentiating abilities and ultimately determined by personal choices of what should be processed further. These personal choices are also influenced by a person's perceptual filter that screens the incoming information through one's predispositions, background, experience, mental and physical states, etc. (Wolvin, 2010). An example of filtering that negatively impacts selection during academic listening is filtering new information by a personal bias (e.g., Imhof, 1998).

Organizing. In general terms, within the information processing framework, organizing refers to a mental arrangement of the selected information into mental structures, i.e., mental representations (Mayer, 2012). Earlier, I discussed that from the schema-based theoretical perspective, mental representations are developed through the manipulations with the incoming information as well as via the involvement of schemas retrieved from the long-term memory. For the purposes of this study, I refer to the former as organizing processes and to the latter as integrating processes. However, this separation is rather artificial, as these processes occur simultaneously and depend on each other. Thus, in this subsection, I discuss the cognitive processes of developing mental representations without regard to retrieved schemas.

An overarching cognitive process necessary for organization is mentally following someone when they are speaking (e.g., Richards, 1983). Specifically, a student may need to follow a hypothesis, persuasion, or argument in lectures (Aryadoust et al., 2012). For example, following an argument may involve analyzing the successive steps in an argumentation (Vermunt, 1998). Relevant to the process of following are also the processes of relating pieces of incoming information to each other and making connections between them (Vermunt, 1996). Specifically, a student may need to identify relationships among units within discourse, such as major ideas, generalizations, hypotheses, supporting ideas, and examples (Richards, 1983). Inferring the nature of such relationships (e.g., cause, effect, conclusion) may also be important (Richards, 1983). Lastly, logically, the processes of relating or connecting may also take a form of comparing, contrasting, or associating. A variation of this process is the process of relating information obtained from different sources. For example, while listening to a lecture, a student may connect the information they hear with the information from a textbook or handouts (Aryadoust et al., 2012). Kardash and Amlund (1991) considered the process of mentally combining different pieces of new information from course materials into some new order that makes sense to a student.

A different but also frequently encountered in the literature organizing strategy is paraphrasing (e.g., Aryadoust et al., 2012; Thompson et al., 2004). Some researchers describe this process as putting or expressing new information, ideas, or concepts into students' own words (Greene et al., 2004; Pintrich & de Groot, 1990; Schmeck et al., 1977). Processes that often subsume both relating and paraphrasing are summarizing (e.g., Aryadoust et al., 2012; Halone et al., 1998; Thompson et al., 2004) and synthesizing (e.g., Thompson et al., 2004). For example, one of the elaborative processing strategies for learning new material of Schmeck et al. (1977) refers to summarizing material in a student's mind in their own words. Another variation of the relating and summarizing processes is the process of self-explanation, i.e., the process of "generating explanations for oneself in an attempt to make sense of relatively new information" (Rittle-Johnson et al., 2017, p. 600). In the context of organizing, self-explaining helps integrate pieces of new information together. Overall, paraphrasing, summarizing, and synthesizing strategies are especially useful for students as they take notes during lectures (Aryadoust et al., 2012; Canpolat et al., 2015; Imhof, 1998). A study of Van Meter et al. (1994) showed that while some content is recorded verbatim (e.g., definitions and examples), other material tends to be noted in a paraphrased form (e.g., concepts and ideas).

Organizing as a general term is also sometimes referred to as structuring (Imhof, 1998; Vermunt, 1998). It needs to be noted that not all organizing or structuring strategies reported in the literature are cognitive in nature. For example, making charts or diagrams (Pintrich et al., 1991) certainly helps with the organization but essentially is a physical action (i.e., a behavior) rather than a cognitive process. Nevertheless, the process of mental development of such charts or diagrams can be considered cognitive.

Integrating. Integrating, within the information processing framework, refers to making connections between the selected information and existing knowledge from the long-term memory (Mayer, 2012). This process is similar to the process of relating discussed above. However, there is one substantial difference: relating in the previous subsection is made within the incoming information, while in this subsection relating goes beyond the new information. Researchers, studying listening, reading comprehension, or learning in general, conceptualize this process as relating, connecting, comparing, contrasting, or associating new information with prior knowledge or experience (e.g., Canpolat et al., 2015; Imhof, 1998; Kardash & Amlund, 1991; Pintrich et al., 1991; Pokay & Blumenfeld, 1990; Schmeck et al., 1977; Weinstein et al., 2016).

A variation of the relating process is relating new information to practice or to a context different from a learning situation (i.e., a classroom). This process may take a form of thinking about practical applications of concepts or visualizing situations in which they may occur (Schmeck et al., 1977). Another way of relating to practice is through creating examples or making analogies (Canpolat et al., 2015; Vermunt, 1996). Self-explaining, mentioned in the previous subsection, may be viewed as a variation of relating and summarizing in this subsection, as well. Here, self-explanations may serve as a means for integrating new information with existing knowledge (Rittle-Johnson et al., 2017).

Next, applying previous knowledge to new situations may take a form of critical thinking (Pintrich et al., 1993). In other words, students critically evaluate new information using the knowledge they already have (Pintrich et al., 1993) and make judgments about this information (Al-Musalli, 2015; Fontana et al., 2015). This process may also include questioning (Pintrich et al., 1991). Further, the process of integrating new information with the existing knowledge, especially when critical thinking is involved, may reveal conflicts between what is being learned and what is already known. Kardash and Amlund (1991) considered the process of resolving such conflicts as another cognitive process. Self-explaining may also facilitate this process (Chi, 2000). Thus, through conflict resolution, new information may lead to the processes of re-evaluating (Thompson et al., 2004) and revising (Chi, 2000) prior knowledge.

As previously discussed, the processes of organizing and integrating are not independent in practice, though I presented them separately. These processes taken together seem to correspond to the processes of understanding, comprehending, decoding, or interpreting within the listening theories (e.g., Thompson et al., 2004; Wolvin, 2010) or to the processes of elaborating in educational research (e.g., Pintrich et al., 1991).

Cognitive processes during problem solving. From the cognitive psychology perspective, problem solving is viewed as a series of mental operations that transform knowledge representations (Mayer, 1982). In this section, I explore these mental operations, or cognitive processes, that occur during problem solving. I start by describing existing models of problem solving and proceed to discussing specific cognitive processes.

Models of problem solving. In this section, I describe two models of problem solving – the schema-driven model of Gick (1986) and the transfer model of Nokes-

Malach and Mestre (2013). I close this section by connecting the terminology used in these models with the terminology of the information processing theory.

To start, I describe a holistic picture of the schema-driven problem solving process developed within the information processing framework and presented by Gick (1986). According to Gick (1986), the global problem solving process consists of the two sub-processes: (1) generation of a problem representation or problem space (i.e., the problem-solver's view of the problem) and (2) a solution process that involves a search through the problem space (p. 101). Constructing a problem representation involves understanding the problem, i.e., developing a representation of what is given in the problem and a representation of the problem goal (Greeno, 1977). The process of understanding is conducted through analyzing the relationships between problem elements and identifying patterns of these relationships. Further, as part of constructing a problem representation, particular problem features activate relevant knowledge in memory, i.e., schemas. Citing Gick and Holyoak (1983), Gick (1986) conceptualized a schema as "a cluster of knowledge related to a problem type" that "contains information about the typical problem goal, constraints, and solution procedures useful for that type of problem" (p. 102). If an appropriate schema is found and activated, a problem-solver can proceed to solving a problem (the third stage of the problem-solving process), i.e., implement solution strategies and procedures the schema provides. Otherwise, a problem-solver needs to search for a solution first (the second stage of the problem-solving process). If a solution is not implemented successfully, the problem-solver may go back to the stages of representation construction or solution search and try different strategies.

I discuss these problem solving strategies in more detail in the subsection about specific cognitive processes below.

Another model of problem solving as a process of transfer has been developed more recently by Nokes-Malach and Mestre (2013). In the view of Nokes-Malach and Mestre (2013), transfer is “a dynamic process in which the learner engages in the highly selective activation and application of knowledge to create a representation that allows her or him to make sense of the situation in order to accomplish some goal or perform some task” (p. 185). Their model was built on the schema-driven model of problem solving, developed by Gick (1986). Thus, similarly to the model of Gick (1986), the model of Nokes-Malach and Mestre (2013) reflects the information processing framework and includes two major processes: representation construction and solution generation. One difference from the model of Gick (1986), pointed out by Nokes-Malach and Mestre (2013), is the inclusion of situational aspects, such as social, motivational, and ecological factors. Thus, the representation construction process is extended beyond understanding the problem (framing) and activating relevant knowledge to also take into account the influence of the environment (e.g., physical or social).

Another difference between the two models is in the process of solution generation. In particular, as noted by Nokes-Malach and Mestre (2013), their model uses transfer mechanisms for solution generation, whereas Gick's model (1986) uses problem solving strategies. I discuss both – transfer mechanisms and problem solving strategies – in the next subsection. A final major difference between the models is the specification of the evaluation process. While both Gick (1986) and Nokes-Malach and Mestre (2013) stated the presence of such a process, Nokes-Malach and Mestre (2013) explained the evaluation criterion. According to their model, a result of each process – a mental representation of a problem from the representation construction process and a solution during the solution generation process – is evaluated during the sense-making process. Specifically, a problem-solver decides if they are satisfied with the result (a process called satisficing by Nokes-Malach and Mestre, 2013).

In conclusion, I discuss how the models of problem solving are connected with the information processing theory. The two major parts in the considered models of problem solving are the processes of construction a problem representation and generation of a solution. The information processing theory, on the other hand, operates on the processes of selecting, organizing, and integrating. However, the processes within the two theories may not be as different as they may seem, as both representation construction and solution generation involve the processes of selecting, organizing, and integrating. Indeed, during the process of representation construction, a student selects information from the problem statement, organizes it into a mental representation (i.e., interprets), and integrates this representation with the prior knowledge (i.e., uses this knowledge to revise the mental representation). Similarly, during the process of solution generation, a student may employ the process of selecting while searching for information needed to solve a problem. Further, the student organizes the information found outside and/or the prior knowledge retrieved from memory during the integration process. Another form that integrating may take is modifying prior knowledge as a result of problem solving.

Specific cognitive processes. In this section, I present specific cognitive processes, suggested by the models of problem solving that I discussed above. These cognitive processes include problem solving strategies and transfer mechanisms, involved in the models of Gick (1986) and Nokes-Malach and Mestre (2013), respectively. I also describe other cognitive processes, found in the problem solving literature, and relate them to the cognitive processes that I identified for listening.

I start the discussion of specific cognitive strategies with problem solving strategies. Gick (1986), referencing Mayer (1983), defined problem solving strategies as "techniques that may not guarantee solution but serve as a guide in the problem solving process" (p. 100). Gick (1986) emphasized that these strategies can be content specific (i.e., specific to a particular topic in a particular domain) or general (i.e., applicable across topics and domains). In terms of the latter, Gick (1986) described several commonly used strategies. One strategy is problem decomposition, i.e., breaking the problem into sub-problems. Another strategy is a means-ends analysis, which involves reducing the difference between the current state of the problem and its goal. She also mentioned, citing Polya (1957), the strategy of using analogies, which entails searching for an analogous problem with a known solution. These problem solving strategies may be used during the process of either representation construction or solution search.

Prior to discussing transfer mechanisms, a note about the notion of transfer needs to be made. In its classic form, transfer refers to transportation of knowledge elements that are learned in one situation or task and applied to another (Nokes-Malach & Mestre, 2013). The former situation is typically referred to as a learning situation, and the latter –

as a test or transfer situation. This approach produced inconsistent results and subsequently received a substantial critique from the research community that aimed to determine the reasons for this inconsistency (Nokes-Malach & Richey, 2015). Concerns included limiting the context to learning and test tasks, ignoring other factors that may influence the process of transfer (e.g., environmental aspects or individual differences), and measurement issues (the restriction to measures of problem solving accuracy, solution strategy, or reaction time). Thus, a number of alternatives perspectives on transfer occurred.

One alternative model is the model of Nokes-Malach and Mestre (2013) described above. Another alternative view, called Actor-Oriented Transfer, was proposed by Lobato (2012). Both theories deviated from the classical view in a sense that the transfer process no longer occurs from a single learning situation but from a collection of prior knowledge and experiences. The view of Lobato (2012) also makes an important point that what a student transfers to a particular test situation may or may not be what they may be expected or assumed to transfer. Similarly, incorrect performance during the transfer situation may or may not indicate the lack of transfer, as a student may have transferred an incorrect interpretation of a learning situation or other experiences. Thus, a modern understanding of transfer is more complex but also more flexible than its classic view.

Nevertheless, some aspects of the classic view remain to be useful and are incorporated in the modern theories, e.g., in the model of Nokes-Malach and Mestre (2013). These aspects are classic mechanisms of transfer: identical rules, analogy, knowledge compilation, and constraint violation (Nokes-Malach & Mestre, 2013). Identical rules refer to the process of rule application if the rule conditions match the current context. Analogy, also mentioned by Gick (1986) as a problem solving strategy, consists of three sub-processes: (1) retrieving an example, (2) aligning and mapping it to the current context, and (3) drawing an inference for the current context. Knowledge compilation involves interpreting knowledge components into specific problem solving actions. Lastly, constraint violation can be viewed as a cycle of generating, evaluating, and revising. Further, these mechanisms differ not only in the cognitive processes used but also in the knowledge structures they operate on (ranging from procedural to declarative knowledge), efficiency levels (from highly efficient to least efficient), and scope (from narrow to wide).

Notably, not all mathematics educators, working in the constructivism paradigm, recognize transfer as a valid approach to explaining the contributions of prior knowledge to learning (e.g., Carraher & Schliemann, 2002). In the perspective of Carraher and Schliemann (2002), transfer is “a relatively passive “carrying over” and deployment of learning from one situation to another once learners recognize the “similarity” between those situations” (p. 19). They argue that learning, as a constructivist process, goes beyond such “carrying over.” Specifically, as the authors show through clinical interviews, students use a variety of prior knowledge and experiences, adjusting them throughout the learning process. For this reason, Carraher and Schliemann (2002) called for abandoning the concept of transfer in its entirety.

Yet, the definition of Carraher and Schliemann (2002) and, subsequently, their critique seem to be very similar to the definition and critique of the classic view on transfer. This critique has been largely addressed by the modern transfer theories. Responding to Carraher and Schliemann (2002), Nokes-Malach and Mestre (2013) agreed that the classic view cannot capture the dynamic constructive process of transfer as this view permits the use of only one transfer mechanism per transfer process. However, explaining this process is possible within the model of Nokes-Malach and Mestre (2013), which permits triggering of multiple mechanisms throughout the learning process. The authors supported their point by re-examining the interview data of Carraher and Schliemann (2002), showing how the knowledge construction process of the study participants may be explained through multiple transfer mechanisms. Notably, Carraher and Schliemann (2002) recognized the existence of modern transfer theories but did not find re-conceptualization of transfer appropriate. Instead, they argued for the adoption of Piaget’s theory (specifically, the notions of assimilation and accommodation) to explain learning.

Finally, beyond problem solving strategies and transfer mechanisms, problem solving may involve other cognitive processes. These processes, similarly to the cognitive processes of listening, can be classified by the information processing categories of selecting, organizing, and integrating. Moreover, some cognitive processes occur in both problem solving and listening, as discussed below.

Processes of selecting in problem solving can be represented as identifying the nature of a problem (Bonner, 2013) or focusing on particular information in a problem statement (Hegarty et al., 1995). Questioning and self-explaining are used in problem solving, as well (Rittle-Johnson et al., 2017; Rosenzweig et al., 2011). Particularly, students may ask themselves about what is given when examining a problem statement (Pokay & Blumenfeld, 1990). In term of organizing, problem solving processes may include paraphrasing and mentally visualizing a task (Rosenzweig et al., 2011). Paraphrasing is also used in listening, and visualizing may roughly correspond to the process of synthesizing information into a schematic representation. Further, comparing is another process used in both problem solving and listening. However, the nature of this process differs. In problem solving, it may involve comparing solutions among students (Kosko, 2014), which can occur in collaborative environments. Additionally, students may also engage in comparing when an instructor presents the solution students had worked on before.

In terms of integrating, the relating process, largely used in listening, is of similar importance in problem solving. However, problem solving researchers do not typically refer to this process as relating. Indeed, the work on problem solving discusses the processes of retrieving relevant knowledge and using it to generate a solution (Wong et al., 2002), which is essentially a process of integrating prior knowledge with a current mental representation of a problem. Also, similarly to listening, problem solving includes critical thinking processes, specifically the processes of evaluating the solution through, for example, reviewing and checking (Rosenzweig et al., 2011; Wong et al., 2002).

My conceptualization of cognitive engagement. To conceptualize cognitive engagement, I adopt the information processing framework as it is applicable to both listening and problem solving. Specifically, I use cognitive processes, classified within the categories of selecting, organizing, and integrating, to indicate cognitive engagement. With problem solving items in particular, I aim to measure cognitive engagement in a broad range of tasks, including mathematical and non-mathematical tasks (e.g., conceptual questions, codes, etc.). Therefore, I do not include problem solving strategies and transfer mechanisms in the operationalization of cognitive engagement, as they are specific to mathematical problems and are not appropriate for non-mathematical tasks. Lastly, while there are numerous cognitive processes that can occur in a student's mind during class, I focus specifically on those that are expected of a student within each type of instruction. Thus, for my instrument, I conceptualize cognitive engagement as students' expected cognitive processes of selecting, organizing, and integrating.

Emotional engagement. In this section, I first describe the differences between emotions, moods, and affective traits. Next, I discuss two major theoretical views of emotions: discrete and dimensional. Finally, I focus specifically on educational research and examine the types of academic emotions as well as specific emotions used in educational research studies. I conclude with the conceptualization of emotional engagement for my instrument.

Emotions, moods, and affective traits. Rosenberg (1998) proposed a hierarchical model of affective organization that was comprised of three levels of affect, specifically affective traits and two classes of affective states – moods and emotions. According to Rosenberg (1998), affective traits are "stable predispositions toward certain types of emotional responding" (p. 249). Moods are affective states that are shorter in duration, less pervasive, and narrower in distributive breadth (the range of different psychological and physiological processes that can be influenced by the level of affect), compared to affective traits. Emotions are “acute, intense, and typically brief psychophysiological changes that result from a response to a meaningful situation in one's environment” (p. 250). Thus, they are the shortest in duration, least pervasive, and narrowest in distributive breadth. According to the hierarchical model of Rosenberg (1998), affective traits influence emotions both directly and indirectly through moods. While this direction is predominant, another direction is possible as well, i.e., emotions also influence moods, which in turn influence affective traits.

Pekrun (2006) views emotions more broadly than only psychophysiological changes, as Rosenberg (1998) does. According to Pekrun (2006), emotions are “multi-component, coordinated processes of psychological subsystems including affective, cognitive, motivational, expressive, and peripheral physiological processes” (p. 316). Yet, as Rosenberg (1998), Pekrun (2006) also views emotions to be intense. Moods are comprised of the same components but are less intense (Pekrun, 2006). Therefore, in Pekrun’s perspective, moods can be defined as low-intensity emotions. The difference between moods and emotions can be also described in terms of a referent. Moods typically do not have a specific referent, whereas emotions do (Pekrun & Linnenbrink-Garcia, 2012). For example, emotions can be caused by a particular learning activity. Nevertheless, both moods and emotions can influence student learning (Pekrun & Linnenbrink-Garcia, 2012). In the next section, I discuss how affect is conceptualized in the literature.

Discrete vs. dimensional view. Two views on how emotions can be distinguished from one another have been proposed. One view – discrete or modular – suggests that emotions are separate and fundamentally different from each other (Ekman, 2016; Ekman & Cordaro, 2011). One of the biggest theories within this view is the theory of basic emotions. For example, Tomkins and McCarter (1964) proposed eight primary affects, which they named at both moderate and high intensity: interest/excitement, enjoyment/joy, surprise/startle, distress/anguish, fear/terror, shame/humiliation, contempt/disgust, and anger/rage. Izard, in their Differential Emotions Theory, also identified interest, joy, surprise, distress, fear, contempt, disgust, and anger, and added guilt and shyness (e.g., Izard et al., 1971). In 2011, Ekman concluded that evidence exists for seven basic emotions: anger, fear, surprise, disgust, contempt, and also happiness and sadness (Ekman & Cordaro, 2011). In short, the list of basic emotions is not well-established and differs among researchers.

Another view proposes that emotions are differentiated via dimensions. Most common dimensions are valence and arousal (or activation), which can be applied either separately or simultaneously. In terms of valence, emotions can be positive or negative (Watson & Tellegen, 1985), also referred to as pleasant or unpleasant (e.g., Barrett & Russell, 1998). The former may include such emotions as happy and content; the latter may include such emotions as unhappy and miserable (Barrett & Russell, 1998). In terms of arousal, researchers typically distinguish between activating and deactivating emotions (e.g., Barrett & Russell, 1998). Activating emotions may include such emotions as aroused and alert, and deactivating emotions may include such emotions as sleepy and quiet (Barrett & Russell, 1998). However, Thayer (1986) further distinguished between energetic arousal (energy and tiredness dimensions) and tense arousal (tension and calmness dimensions). Examples of emotions in the energy dimension are energetic and active, in the tiredness dimension – tired and sleepy, in the tension dimension – tense and jittery, and in the calmness dimension – still and placid.

When valence and arousal are applied simultaneously, they create a two-dimensional space, proposed by Russell (1980). Specifically, Russell (1980) developed the circumplex model of affect where pleasure (0°) and misery (180°) define the horizontal dimension, and arousal (90°) and sleepiness (270°) define the vertical dimension. Further, Russell (1980) also identified two supplemental dimensions that are not independent but helpful in defining the quadrants: excitement (45°) and depression (225°) as one supplemental dimension, and distress (135°) and contentment (315°) as another. Examples of pleasant activating emotions are interested and excited, examples of pleasant deactivating emotions are relaxed and calm, examples of unpleasant activating emotions are irritable and nervous, and examples of unpleasant deactivating emotions are bored and tired (Barrett & Russell, 1998).

Since each of the dimensions has two ends, a question arises of whether there is one bipolar dimension or two independent dimensions. For example, there may be a single bipolar valence dimension (with positive emotions on the one end and negative on the other), or there may be two independent valence dimensions, one for positive emotions and one for negative. The same logic also applies to activation. Green et al. (1993) suggested that true dimensionality of affect may be masked by measurement error.

They investigated the question of valence dimensions, using multiple measures of affect. Measurement models were tested via Confirmatory Factor Analysis (CFA) to account for both random and systematic errors. (The systematic error was hypothesized to occur for items from the same measures.) If positive and negative emotions were truly bipolar, their latent factors were expected to have a large negative correlation. In contrast, if they were independent, the correlation between the latent factors was expected to be close to zero. The authors found that once random and systematic errors were taken into account, affect was bipolar in terms of the happy-sad dimension as well as the positive-negative dimension.

Later, Barrett and Russell (1998) used the approach of Green et al. (1993) to examine the dimensionality of affect in terms of not only valence but also activation. They found that valence and activation are independent of each other; however, positive and negative affect are bipolar, and activation and deactivation are also bipolar. Further, Barrett and Russell (1998) also described two possible sets of independent dimensions: (1) pleasant-unpleasant and activation-deactivation, and (2) pleasant activation-unpleasant deactivation and unpleasant activation-pleasant deactivation. The authors noted that mathematically there is no difference between the sets.

Finally, in addition to valence and activation, other dimensions were proposed. For example, the third dimension in the three-dimensional framework of Mehrabian (1996) is dominance-submissiveness. In this framework, dominance refers to “a person's characteristic feelings of control and influence over his life circumstances,” and submissiveness refers to “feelings of being controlled and influenced by others or events”

(Mehrabian, 1996, p. 266). A different third dimension was proposed in the context of educational research by Pekrun (2006). Pekrun (2006) refers to this dimension as object focus, which is described in the next section. To conclude the discussion of discrete vs. dimensional view of emotions, I will note that both views are currently held by emotion researchers. Furthermore, the two views are not necessarily mutually exclusive. Ekman (2016) surveyed emotion researchers and reported that 18% of them hold a dimensional view, 16% hold a discrete view, and 55% hold both views. Pekrun (2016) suggested that the two approaches can be integrated by viewing discrete emotions as lower-level factors and dimensions as higher-order factors that describe common properties of the discrete emotions.

Academic emotions. For student learning, most important are academic emotions, i.e., “emotions that are experienced in an academic context” (Goetz et al., 2003, p. 11). Measuring academic emotions, some researchers use the two-dimensional framework of valence and activation (e.g., Linnenbrink-Garcia et al., 2011). One type of academic emotions is achievement emotions, i.e., “emotions tied directly to achievement activities or achievement outcomes” (Pekrun, 2006, p. 317). This definition introduces the third dimension – object focus – that differentiates between activity emotions and outcome emotions. The former refers to emotions toward ongoing achievement-related activities, and the latter refers to emotions toward the outcomes of these activities (Pekrun, 2006). Outcome emotions can be further differentiated between prospective and retrospective emotions. Prospective emotions are “related to future success and failure, such as hope and anxiety”; retrospective emotions are “related to success and failure that already occurred, such as pride, shame, relief, and disappointment" (Pekrun & Linnenbrink-Garcia, 2012, p. 262).

Other academic emotions include epistemic, topic, and social emotions (Pekrun & Stephens, 2012). Epistemic emotions are triggered by the cognitive characteristics of a task and the processing of task information (Pekrun et al., 2017; Pekrun & Stephens, 2012). Pekrun et al. (2017) identified seven categories of epistemic emotions: surprise, curiosity, enjoyment, confusion, anxiety, frustration, and boredom. To be epistemic, these emotions need to occur as a result of experiencing cognitive incongruity. For example, in the study by Pekrun et al. (2017), cognitive incongruity was achieved by having students read conflicting texts about climate change.

Topic emotions are emotions that are "triggered by the contents of learning material" (Pekrun & Stephens, 2012, p. 5). Differently from achievement and epistemic emotions, topic emotions are not directly related to learning. An example of topical emotions is students' negative emotions about the re-classification of Pluto (a finding of Broughton et al., 2013). Finally, social emotions are emotions caused by the interaction with other participants in the educational process (e.g., teachers or classmates) or by socially constructed aspects of learning, such as goals, contents, and learning outcomes (Pekrun & Stephens, 2012, p. 5). An example of social emotions is anger or gratitude felt about teachers. Pekrun and Stephens (2012) also suggested that achievement emotions and social emotions can overlap, resulting in social achievement emotions. An example of social achievement emotions is feeling envious of other students' success.

Finally, measurement of academic emotions also differs in the level of specificity. For example, the Achievement Emotions Questionnaire (AEQ) of Pekrun, Goetz, Frenzel, Barchfeld, and Perry (2011) has been designed to measure three types of emotions with respect to their corresponding situational contexts: class-related, learning-related, and test-related emotions. The AEQ can be further adjusted for different temporal specificity. By adapting instructions, three types of emotions – trait, course-specific, and state types of emotions – can be measured. Moreover, academic emotions can also be reported with respect to a particular activity (e.g., Linnenbrink-Garcia et al., 2011; Pekrun et al., 2017).

Emotions in educational research. In this section, I explore which emotions were used in conceptualizations and measures of emotional engagement, as well as outside of emotional engagement in educational research more broadly. While, as discussed above, some educational researchers use particular frameworks in their work, many others do not. Thus, below, I analyze emotions regardless of the framework, within which they may have been conceptualized. Additionally, since many measures of emotional engagement (or engagement more broadly if it is conceptualized as unidimensional) include not only emotions but other indicators as well, only emotions are discussed. Finally, since my interest in academic emotions concerns any emotions that can occur in a classroom regardless of its nature, I do not differentiate between different types of academic emotions (i.e., achievement, epistemic, topic, or social).

In sum, I analyzed 15 papers that used emotions. Results of this analysis are presented in Table 1. Overall, the results suggest that emotions, most frequently encountered in educational research, are interest, enjoyment, and boredom. Also repeatedly used are such emotions as excitement, happiness or unhappiness, pride, and curiosity, as well as anxiety, worry, anger, nervousness, and sadness. Occasionally, educational researchers are also interested in whether students feel enthusiastic, frustrated, ashamed, tired, comfortable, discouraged, energetic, hopeful or hopeless, mad, relaxed, satisfied or dissatisfied, sleepy, afraid, amazed, astonished, calm, confused, dull, inquisitive, irritated, muddled, puzzled, relieved, scared, surprised, tense, thrilled, uneasy, worn out, and more. In conclusion, I will note that sometimes, to indicate emotional engagement, educational researchers use feelings that are not strictly emotions, such as feeling good (Miserandino, 1996; Skinner et al., 2008; M.-T. Wang et al., 2016), positive (Burch et al., 2015), fine (Miserandino, 1996), bad (Miserandino, 1996; Skinner et al., 2008), or down (M.-T. Wang et al., 2016).

My conceptualization of emotional engagement. For my instrument, I adopt the dimensional view of emotions. Thus, I conceptualize emotional engagement as students' positive activating, positive deactivating, negative activating, and negative deactivating emotions. In the engagement literature, some researchers used negative emotions to indicate disaffection (e.g., Skinner et al., 2009); yet, others reverse-scored them to indicate engagement (e.g., Wang et al., 2016). In this study, I did not aim to separate engagement and disaffection. Thus, I aimed to reverse-score negative emotions to indicate engagement. By emotions, I mean any type of academic emotions that a student can experience during classroom learning, regardless of its nature or source (e.g., activity-related achievement emotions, epistemic emotions, topic emotions, or social emotions). Since outcome emotions are not the focus of my instrument, the third – object focus – dimension is not applicable to this study. Further, I do not employ the dominance-submissiveness dimension either, as it is not applicable to educational research.

Table 1. Frequencies of emotion occurrence in the educational research literature

Emotion	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	Total
Bored		1	1	1	1			1	1	1			1	1	1	10
Enjoy / joyful		1	1	1	1		1	1	1		1			1	1	10
Interested		1	1		1	1		1		1	1	1		1	1	10
Anxious		1	1	1	1			1			1					6
Worried	1	1	1		1				1		1					6
Angry		1		1	1			1		1						5
Excited	1		1				1			1		1				5
Happy		1	1				1			1				1		5
Nervous	1		1		1					1	1					5
Curious		1	1							1	1					4
Proud				1	1							1		1		4
Sad		1			1			1		1						4
Enthusiastic	1				1							1				3
Frustrated			1		1				1							3
Ashamed				1	1			1								3
Tired	1	1									1					3
Comfortable	1	1														2
Discouraged					1					1						2
Energetic	1											1				2
Hopeful				1				1								2
Mad		1			1											2
Relaxed	1	1														2
Satisfied					1						1					2
Sleepy	1	1														2
Afraid											1					1
Amazed			1													1
Astonished			1													1
Calm	1															1

Emotion	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	Total
Confused			1													1
Disinterested					1											1
Dissatisfied			1													1
Dull			1													1
Hopeless				1												1
Inquisitive			1													1
Irritated			1													1
Monotonous			1													1
Muddled			1													1
Pleasurable											1					1
Puzzled			1													1
Relieved				1												1
Scared		1														1
Surprised			1													1
Tense	1															1
Thrilled	1															1
Uncomfortable											1					1
Uneasy	1															1
Unhappy		1														1
Vitality					1											1
Worn out	1															1
Zest					1											1

Note: A: Linnenbrink-Garcia et al. (2011), B: Miserandino (1996), C: Pekrun et al. (2017), D: Pekrun et al. (2011), E: Skinner et al. (2008), F: Archambault et al. (2009), G: Li and Lerner (2013), H: Hospel et al. (2016), I: Wang et al. (2016), J: Connell (1990), K: Kong et al. (2003), L: Burch et al. (2015), M: Marks (2000), N: Lam et al. (2014), O: Rimm-Kaufman et al. (2015); "1" represents the presence of the emotion in the paper; the total column presents the number of papers in which an emotion was used.

Chapter Three

This chapter describes procedures for instrument development and initial validation. The instrument under development is a self-report instrument designed to measure student behavioral, cognitive, and emotional engagement within lecture, whole-class interaction, group work, and individual work over the course of the semester in undergraduate mathematics-based classes. I start by describing the argument-based approach to validity I adopted, and the frameworks for instrument development and validation I based my argument and sources of evidence on. Next, I present the interpretation/use and validity arguments for my instrument. Finally, I describe the procedures and analyses conducted to develop the instrument and validate instrument scores.

Argument-Based Approach to Validity

To be valid and, therefore, useful, scores on an educational or psychological measure need to have clearly stated interpretations and uses, specified by the developers and supported by the appropriate validity evidence (AERA, APA, & NCME, 2014). The argument-based approach to validity (Kane, 2013; 2015) provides a framework for validating proposed interpretations and uses. Specifically, two types of arguments are included in the argument-based approach: the interpretation/use argument and the validity argument. In the interpretation/use argument, the developer states the intended interpretations and uses of the instrument and specifies inferences and assumptions that warrant these interpretations and uses. The validity argument evaluates the interpretation/use argument by providing evidence for the coherence of the interpretation/use argument, the reasonableness of its inferences, and the plausibility of its assumptions. In other words, the argument-based approach to validity suggests that there is no "one size fits all" way of validation; rather, means of validation are selected and employed based on the necessity of a particular type of validity evidence for the proposed interpretations and uses. Prior to specifying my interpretation/use and validity arguments, I describe two general methodological frameworks for instrument development and validation. The information that these frameworks provide – the aspects of validity and means to enhance it – can be used in one's interpretation/use and validity arguments. My use of these frameworks entails the critical evaluation of the necessity of their elements for my arguments, the selection of elements deemed necessary, and the inclusion of such elements in the process of instrument development and validation.

Methodological Frameworks for Instrument Development and Validation

Model of Gehlbach and Brinkworth (2011). The model of Gehlbach and Brinkworth (2011) consists of a multi-step process of survey development and validation that aims to enhance validity. The model particularly emphasizes early stages of the instrument development that typically do not receive sufficient attention. I adopted this model to describe sources of evidence and provide a chronology for data collection.

The first step in the process of scale construction involves reviewing the literature. The goal of this step is (1) to precisely define the construct as it relates to the literature, and (2) to determine how existing instruments can be useful. The second step includes interviews and focus groups, which aim to examine whether the researcher-developed conceptualization of the construct matches the respondents' thinking about it. I refer to these type of interviews as exploratory interviews throughout this document. The third step involves synthesizing findings from the literature review (Step 1) and interviews/focus groups (Step 2). The synthesis results in (1) the conceptualization of the construct acceptable by both researchers and respondents, and (2) a list of indicators that can be used to measure the construct. The fourth step is item development. During this step, researchers write preliminary items in accordance with the construct conceptualization and using potential indicators that were identified in Step 3. The fifth step involves expert validation, through which the quality of the scale and individual items are examined. Specific characteristics that can be investigated in expert reviews include item relevance and clarity, as well as scale representativeness. The sixth step is cognitive pretesting, also referred to as cognitive interviewing. The goal of this step is to explore how respondents understand and answer the items. Gehlbach and Brinkworth (2011) conclude the process with the pilot testing of the instrument, which involves administering it to a larger sample of respondents and conducting quantitative analysis to provide further validity evidence. In this document, I refer to pilot testing as field-testing. This testing may be repeated until the set of high-quality items is established and is ready for testing on a large sample, representative of the specified population.

The unified construct-based model of validity (Messick, 1995). The unified construct-based model of validity (Messick, 1995; interpreted by, for example, Dimitrov,

2012) is one of the most comprehensive contemporary models of instrument validation, established in the field of educational measurement. It includes six aspects – content, substantive, structural, external, generalizability, and consequential – which "function as general validity criteria or standards for all educational and psychological measurement" (Messick, 1995, p. 744). I adopted this model as a general guide for validating my instrument. In particular, I used the aspects of Messick's model as building blocks for my interpretation/use and validity arguments.

The content aspect refers to content relevance and representativeness of items as well as their technical quality. The former can be established through expert judgments or developing the logical design of test items (classifying items by content area and test objective); the latter – through item analysis (e.g., item-total correlations) or IRT modeling (e.g., item-measure correlations). The substantive aspect refers to the consistencies between theoretically expected and observed response processes. Evidence for this aspect can be obtained through cognitive interviews by comparing the observed response processes with the expected ones as well as with respondents' behaviors. Another piece of evidence for the substantive aspect includes evidence for scale functioning, which refers to the consistency between the observed and expected response characteristics of the items. This evidence can be provided by examining shape characteristics, item difficulties, patterns of responses to items measuring the same constructs, etc. The structural aspect refers to establishing the internal structure of the instrument, which can be done through, for example, factor analyses. The generalizability aspect includes generalization of score properties and interpretations to and across groups, settings, etc., including the generalization of test criterion relationships. Evidence for this aspect can be provided through testing for factorial invariance, contextual stability, differential prediction, and reliability. The external aspect seeks convergent and discriminant evidence (e.g., via multitrait-multimethod models). Convergent evidence shows the measure under validation is similar to the existing measures of the same construct. Discriminant evidence, in turn, shows that the measure under validation is distinct from measures of other constructs. The external aspect also seeks evidence of criterion relevance (e.g., via correlational analysis, between-group differences, within-person changes), and evidence of applied utility (e.g., via logistic regression). Lastly, the consequential aspect concerns the implications of score interpretations and uses, especially unintended negative consequences. However, the inclusion of this aspect as part of validity is controversial. Some researchers argue that examination of consequences, though undoubtedly important, should be separate from validation (Cizek et al., 2010).

Interpretation/Use Argument for the Instrument Under Development

In this section, I describe the interpretation/use argument for the instrument under development. I start by describing the proposed interpretations and uses of the instrument. Then, I proceed to the discussion of the inferences and assumptions that warrant the proposed interpretations and uses.

Proposed score interpretations and uses. There are several types of composite scores that the instrument is designed to produce. First, twelve subscale composite scores are designed to be interpreted as the levels of each engagement dimension (behavioral, cognitive, and emotional) in each instruction type (lecture, whole-class interaction, individual work, and group work) in a particular undergraduate mathematics-based class. Second, three engagement dimension composite scores are designed to be interpreted as the levels of student behavioral, cognitive, and emotional engagement in a particular undergraduate mathematics-based class. Third, four instruction type composite scores are designed to be interpreted as the levels of student engagement in four types of instruction (lecture, whole-class interaction, individual work, and group work) in a particular undergraduate mathematics-based class. Fourth, global engagement composite scores (i.e., composite scores on the overall instrument) are designed to be interpreted as the level of overall student engagement in a particular undergraduate mathematics-based class. For each of these composite scores, the term “level” refers to the frequency of student engagement over the course of the semester.

Scores, produced by the instrument, are designed to be used in two ways. First, educators can use the scores to identify the kinds of engagement that students lack. Different kinds of engagement composite scores will enable educators to identify whether students lack overall engagement in a particular undergraduate mathematics-based class or whether they lack a particular dimension of engagement, engagement in a particular instruction type, or both. Second, the scores can be used in research, the goal of which is to inform the development of instructional interventions that aim to improve student engagement. In particular, researchers can use the scores to identify instructional facilitators of engagement (specifically, facilitators of overall engagement, a particular dimension of engagement, engagement in a particular instruction type, or a particular dimension of engagement in a particular instruction type). These instructional characteristics can be further used as a focus of engagement interventions.

Inferences and assumptions. Kane (2013; 2015) suggested four general inferences for achievement tests and beyond. The first inference is the scoring inference that provides an observed score based on the observed performances. The second inference is the generalization inference that provides a universe score, i.e., an expected score over a universe of possible observed performances. This inference may include generalization over such conditions, as items, occasions, context, etc. The third inference is the extrapolation inference that provides an expected score in a broader target domain. Finally, the fourth inference is the decision inference (or inferences) that enables making decisions based on the scores. For theory-based interpretations, Kane suggested that at least three inferences are relevant. The two inferences have been already mentioned: the scoring inference and the generalization inference. The third inference is a theory-based inference that links indicators and constructs. As the interpretation of the scores, produced by my instrument, is theory-based, I proceed with these three inferences. However, considering the proposed uses of the scores, I also include the decision inference in the argument. Additionally, I split the theory-based inference into two: one that makes an inference from items (before scoring) and another one that makes an inference from composite scores (produced as a result of scoring). Below, I describe the inferences in more detail and discuss assumptions that need to be met for inferences to be reasonable. I select assumptions from aspects of Messick’s model where applicable. The aspects are noted in parentheses.

Theory-based inference for item scores. The theory-based inference for item scores links construct measurement (here, items) and constructs, where the constructs are engagement dimensions in instruction types, engagement dimensions, engagement in instruction types, or global engagement. For this inference to be made, the following assumptions need to be satisfied. First, items need to be relevant to and representative of the construct being measured (the content aspect). Second, observed response processes need to match the intended ones and be aligned with respondents' behaviors (the cognitive modeling component of the substantive aspect). Third, item response characteristics need to be consistent with expected characteristics (the scale functioning component of the substantive aspect). Fourth, the internal structure of the instrument needs to be determined (the structural aspect).

Scoring inference. The scoring inference provides composite scores, i.e., scores that indicate levels of the constructs, based on item scores. To compute composite scores for each construct, scoring rules are needed. Subscale composite scores are designed to be computed via averaging the corresponding items. Engagement dimension composite scores are designed to be computed via summing the subscale composite scores across instruction types, weighted by the amount of instruction, and dividing the sums by the total amount of time spent on the four types of instruction. For example, in order to create behavioral engagement composite scores, scores on behavioral engagement in lecture, whole-class discussion, individual work, and group work are weighted by the corresponding amount of time. A sum of these weighted scores, divided by the total amount of time spent on all four types of instruction, constitutes composite scores on behavioral engagement. Instruction type composite scores are designed to be computed via averaging the subscale composite scores across engagement dimensions. For example, composite scores of engagement in lecture will be created by averaging scores on behavioral, cognitive, and emotional engagement in lecture. Global engagement composite scores are designed to be computed via averaging dimension engagement composite scores. Assumptions for the scoring inference are as follows. First, the scoring rules need to be plausible. Second, all information required for computing composite scores needs to be available.

Theory-based inference for composite scores. The theory-based inference for composite scores provides a further link between construct measurement (here, composite scores) and constructs, where the constructs are engagement dimensions in instruction types, engagement dimensions, engagement in instruction types, or global engagement. Satisfying assumptions for the theory-based inference for item scores are required but not sufficient conditions for making the theory-based inference for composite scores. Two additional assumptions apply to the theory-based inference for composite scores. First, I extended the assumption within the scale functioning component within the substantive aspect to composite characteristics. In particular, characteristics of composite scores need to be as expected. Second, expected relationships between composite scores and relevant constructs need to be demonstrated (the external aspect). These constructs were selected based on the three criteria: (1) relevance to student engagement as indicated by the previous research, (2) lack of overlap with my conceptualization and operationalization of student engagement, and (3) availability of a validated measure. In terms of predictive validity, as student engagement has been typically shown to have positive or statistically non-significant effects on achievement (e.g., Reeve, 2013; Reeve & Tseng, 2011; Whitney et al., 2019), I expect to find such relationships in this study. Further, to support discriminant validity, moderate positive relationships of student engagement with effort, persistence, feeling and value components of interest, and metacognitive strategies need to be demonstrated. These constructs are often included in measures of engagement; however, the present instrument is designed to be distinct from them. Prior research typically used effort and persistence (e.g., Kong et al., 2003; Lam et al., 2014; Wang et al., 2016) to indicate behavioral engagement, feelings (e.g., Rimm-Kaufman et al., 2015; Skinner et al., 2009; Z. Wang et al., 2014) and values (e.g., Finn & Zimmer, 2012; M.-T. Wang et al., 2011) to indicate emotional engagement, and metacognitive strategies (e.g., Awang Hashim & Murad Sani, 2008; see also a related concept of self-regulation, e.g., Miller et al., 1996) to indicate cognitive engagement. Thus, it is particularly important to show moderate correlations for effort and persistence with behavioral engagement, feeling and value components of interest with emotional engagement, and metacognitive strategies with cognitive engagement. Further evidence of discriminant validity will include low-to-moderate relationships between engagement and intellect, as suggested by prior research (Douglas et al., 2016). Finally, I also selected three constructs – social efficacy with peers, preference for group work, and public speaking anxiety – to discriminate between engagement in different instruction types. Social efficacy with peers is expected to correlate positively with engagement in group work but is not expected to correlate with engagement in other instruction types. Preference for group work is expected to correlate positively with engagement in group work, negatively with engagement in individual work, and non-significantly with engagement in lecture and whole-class interaction. Public speaking anxiety is expected to correlate negatively with engagement in whole-class interaction and not to be correlated with engagement in other types of instruction. Convergent validity will not be addressed in this study as the subscale constructs are unique and cannot be measured via a different method or instrument.

Generalization inference. The generalization inference provides expected composite scores over a universe of possible composite scores. For this inference to be made, assumptions of generalizability and reliability of scores across settings, groups forms, and formats, needs to be met (the generalizability aspect). These assumptions ensure that scores are generalizable and reliable regardless of the setting, group membership, form, or format. Another assumption within the generalizability aspect posits that relationships between engagement and external variables (discussed in the theory-based inference for composite scores) also need to be generalizable across settings, groups, forms, and formats. This assumption ensures that score relationships with other variables are also the same regardless of the setting, group membership, form, or format. Settings include different course disciplines, course levels (lower vs. upper), and course types (required, elective, general education, or pre-requisite). These settings are selected because they are common in undergraduate education. Groups are based on student classification (freshman, sophomore, junior, and senior), status (full-time or part-time), enrollment in the Honors Program, major, domicile (domestic in-state, domestic out-of-state, or international), native language (English vs. non-English), gender, and race/ethnicity. Forms include eight forms developed to mitigate the potential effect of item block order. Formats include paper-and-pencil and online. Within the online format, there are two versions: PC and mobile. The generalizability aspect also includes an assumption about the internal consistency of subscales. Thus, subscales need to exhibit adequate internal consistency.

Decision inference. The decision inference allows for the use of scores produced by the instrument to identify the kind of engagement that students lack and implement instructional interventions that aim to increase this kind of engagement. The first assumption for this inference states that students need to benefit from instructional interventions that aim to increase particular kinds of engagement. The second assumption states that the benefits need to substantially outweigh the negative consequences of the interventions.

Validity Argument for the Instrument under Development

In the validity argument, I describe the procedures for evaluating the assumptions of the interpretation/use argument in order to warrant the inferences and support the proposed score interpretations and uses. The present study aims to develop the instrument and provide initial validity evidence for the instrument scores. Thus, some inferences and assumptions are outside of the scope of the present study and will be addressed in future validation work. The main effort of the initial validation aims at providing evidence for the theory-based inference, i.e., the inference that the instrument measures the intended construct.

Theory-based inference for item scores. The model of Gehlbach and Brinkworth (2011) was adopted (and slightly adapted) to describe sources of evidence and provide a chronology for data collection (see Figure 1). First, the development of items was conducted via the two methods described by Gehlbach and Brinkworth (2011): literature review and exploratory interviews. The exploratory interviews provided me with students' first-hand engagement experiences, and the literature review provided me with theory-based indicators supported by empirical evidence. As a result, behaviors, cognitive processes, and emotions, which may serve as indicators of each construct, were identified, and a list of potential indicators was created. Based on this list, the items were drafted. Next, assumptions within the content and substantive aspects were examined through expert reviews and cognitive interviews, respectively. Experts evaluated item relevance and representativeness of the construct. During cognitive interviews, I explored student response processes (the cognitive modeling component of the substantive aspect). I also conducted a preliminary evaluation of the assumption within the scale functioning component of the substantive aspect. Specifically, I examined item descriptive statistics using responses from cognitive interviews. The information from expert reviews and cognitive interviews was used to revise the items, which were then evaluated via expert reviews and tested via cognitive interviews again, creating an evaluation-revision loop until the instrument was ready for field-testing. Field-testing data were statistically analyzed to further investigate the assumptions within the scale functioning component of the substantive aspect, as well as to investigate the assumtion within the structural aspect. Specifically, for scale functioning, I examined item response characteristics using descriptive statistics and item correlations. Next, I employed Exploratory Structural Equation Modeling (ESEM) to explore the assumption within the structural aspect. The confirmatory analysis of the internal structure is outside of the initial validation and should be conducted as part of the further validation efforts.

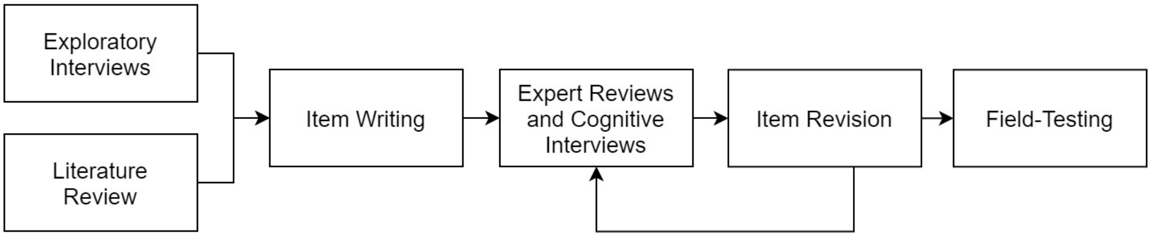


Figure 1. The process of instrument development and validation (based on the model of Gehlbach and Brinkworth, 2011)

Scoring inference. To provide all necessary information needed for scoring, the instrument included (1) engagement items, grouped in blocks by instruction type (the Engagement Survey) and (2) a question that provides information about the percentage of time spent on the four types of instruction (the Instructional Time Form). The latter was needed in order to compute engagement dimension composite scores, which were weighted by the amount of instruction type. Further, the assumption of the plausibility of the scoring rule for subscales – averaging across items – is justified when subscale items are parallel, i.e., have equal loadings and error variances (McNeish & Wolf, 2020). Subscale composites intended to represent particular engagement dimensions in particular instruction types. To examine whether items within each subscale were parallel, I aimed to extend the ESEM model by constraining item loadings to be the same within subscales and item error variances to be the same within subscales. The scoring rule for engagement dimensions (i.e., weighted averages of subscale composite scores across instruction types, where weighting is done by the amount of instruction type) was assumed to be reasonable because weighting accounts for the duration of instruction-specific engagement. Finally, the scoring rule for engagement in instruction types (i.e., averages of subscale composite scores across engagement dimensions) and for global engagement (i.e., averages of engagement dimension composite scores) was assumed to be reasonable because, to my knowledge, there is no evidence in the engagement literature that one engagement dimension is more important than another one to justify the use of weights. I am also not aware of any weighting recommendations in the literature with respect to engagement dimensions.

Theory-based inference for composite scores. To evaluate the assumption within the scale functioning component of the substantive aspect, applied to composite scores, I explored descriptive statistics and correlations of engagement composite scores. Next, to investigate the assumption within the external aspect, I conducted correlational analyses with effort, persistence, feeling and value components of interest, metacognitive strategies, intellect, social efficacy with peers, preference for group work, and public speaking anxiety, and regression analyses with achievement.

Generalization inference. The first step in evaluating the assumption of generalizability of scores (the generalizability aspect) is to examine whether the field-testing sample includes a variety of classes from different settings, all forms, all formats, and students from a variety of backgrounds. Initial reasonableness for the sample representativeness rested upon the overall demographics of George Mason University as well as my efforts to recruit classes from different settings and to use all forms to a similar extent. (See the sections about the target population and data collection site sections). Next, I investigated frequencies of students across settings, forms, formats, and groups. To examine whether format-specific aspects of instrument administration could prevent generalizability across formats (paper-and-pencil and online), I pretested both formats via cognitive interviews. Additionally, I aimed to examine the assumption of the internal consistency of subscales via Cronbach's alpha. A more rigorous investigation of generalizability of scores (e.g., via testing for measurement invariance) as well as an examination of reliability of scores (e.g., via test-retest reliability) across settings, forms, formats, and groups is outside of the scope of initial validation. Investigations of the generalizability of external relationships (e.g., via differential prediction) are also outside of the scope of the initial validation.

Decision inference. Evaluating assumptions of the decision inference is outside of the scope of the initial validation. These assumptions should be investigated once the theory-based, scoring, and generalization inferences are fully supported.

IRB Approval

The study has been approved by the George Mason University Institutional Review Board. Exploratory interviews, expert reviews, and cognitive interviews have been approved under IRB #1015453 (see Appendix A), and field-testing has been approved under IRB #1321959 (see Appendix B).

Target Population

The population for which the instrument is being developed consists of students enrolled in undergraduate mathematics-based STEM classes. I define mathematics-based classes as classes that use mathematical (including logical and statistical) concepts and procedures and involve solving problems or other mathematical tasks. Further, I particularly targeted classes that include group and/or individual work. Put differently, students needed to have an opportunity to interact with each other (e.g., solve problems in formal or informal groups/pairs, participate in small group discussions) and/or work on tasks individually (excluding exams and formal quizzes). The decision to search for classes that incorporate either group or individual work was made because in classes with at least one of these instruction types, students may also be engaged in another one. For example, students who are asked to work in a group may start solving a problem by themselves before turning to others. Alternatively, students who are asked to work individually may still decide to talk to other students. Lecture and whole-class interaction were not explicitly included in the eligibility criteria because I assumed that these instruction types are fairly common and, therefore, do not need to be emphasized.

Data Collection Site

Data collection was conducted at George Mason University. This site is well suited for the purposes of this study for two reasons. First, George Mason is a large and diverse institution. As of Fall 2017, it served 24,987 undergraduate students from a variety of backgrounds (*Office of Institutional Research and Effectiveness*, n.d.). Specifically, over 50% were female students, over 50% were students of color (*Office of*

Institutional Research and Effectiveness, n.d.), 40% were first-generation students (*F1rst Gen Mason*, n.d.), and nearly 30% were Pell Grant recipients (*FA09: Pell Grant Report*, n.d.). One-third of undergraduate students were enrolled in the College of Science and School of Engineering, with over 30% of these students being female and almost 60% being students of color (*Office of Institutional Research and Effectiveness*, n.d.). Such a diverse student body allowed me to field-test the instrument across multiple student background characteristics.

Second, George Mason University is dedicated to providing students with innovative learning experiences. (The university has Innovative Learning as Strategic Goal #1 in its 10-year strategic plan, *George Mason University Strategic Plan*, n.d.). To achieve this goal, the university has built multiple active learning classrooms ("Signature Learning Spaces," 2015) and offered a number of teaching development opportunities for its faculty (*Stearns Center for Teaching and Learning*, n.d.). Such a culture of teaching innovation was conducive to my data collection efforts, specifically to the identification of classes that satisfied the recruitment criterion.

Exploratory Interviews

The goal of exploratory interviews was to identify potential indicators from students' first-hand learning experiences. In this section, I describe exploratory interview participants, the procedure used to conduct interviews, the approach to analysis, and major findings.

Participants and recruitment. To recruit participants for exploratory interviews, I first compiled a list of potentially eligible classes (see more details about eligibility in the Target Population section) based on my knowledge about instructors and their teaching styles. The list included five course sections (within four courses) across two semesters. Course disciplines included mathematics and electrical engineering. Further, the courses represented three course levels: 100, 200, and 300. Student contact information, needed to invite students to participate in the interviews, was requested from their instructors.

To select participants, I first randomized students in each class. Then, a small number of students were selected and invited to participate in the study via email. I did not invite all students at once for two reasons: (1) to make the scheduling process manageable and (2) to ensure that participants were relatively evenly distributed across courses, i.e., there were no courses with a substantially larger number of participants than in other courses. Thus, from some, typically larger, courses, not all students received an invitation, whereas from other, typically smaller, courses, all students were invited. Students in the courses with low response rate received an invitation twice. In total, 15 interviews were conducted (4 student from electrical engineering classes and 11 students from mathematics classes). All interviews took place during one semester.

The sample was diverse in terms of student demographic and background information (see Table 2). The student average age was 21 (*SD* = 3.42), ranging from 18 to 32. Students also varied in majors, which included Electrical Engineering, Computer Science, Biology, and more.

Table 2. Demographic and background information of exploratory interview participants

Characteristic	Frequency
Expected grade in the course	
- A	6
- B	7
- C	2
Student classification	
- Freshman	4
- Sophomore	6
- Junior	2
- Senior	2
Status	
- Full-time	14
- Part-time	1
Domicile	
- Domestic, in-state	14
- Domestic, out-of-state	1
GPA	
- 3.51 or better	4
- 3.01 up to 3.50	6
- 2.51 up to 3.00	4
- 2.01 up to 2.50	1
Gender	
- Male	8
- Female	7
Race/ethnicity	
- White	7
- African-American	1
- Hispanic	2
- Asian	5

Note. N = 15.

Materials and procedure. All exploratory interviews were conducted on a university campus and lasted between 20 minutes and one hour. I was the only interviewer. All interviews were audio-recorded. During the interview, students first learned about the study and signed a consent form. Next, they completed a Demographic

Form, followed by the Instructional Time Form. After that, students were interviewed in accordance with the interview protocol. Each student received a $25 Amazon gift card for their participation.

Demographic Form. The Demographic Form included questions about student learning in the course, their background, and demographic information.

Instructional Time Form. The Instructional Time Form was created to collect information about the amount of time spent on each type of instruction. In the form, students were asked to report the percentage of time they spent on each instruction type and make sure that the total amount of time equals 100%. Brief descriptions and/or examples of instruction types were also provided in the form. Students were asked to report specifically on the lecture section of their class (i.e., not on the recitation or lab sections). The role of the form in exploratory interviews was primarily to set the structure of the interview and to refer to particular types of instruction during the interviews.

Interview protocol. I started interviews by asking students to describe their typical class session. Then, I asked questions about students' engagement in their class. Specifically, I asked about students' behavioral, cognitive, and emotional engagement within each type of instruction. Example questions for behavioral engagement included "What do you typically do in class?" and "Do you always participate? When you don't, why?" Example questions for cognitive engagement included "What do you do to understand the material in class?" and "Do you always understand? When you don't, why?" Example questions for emotional engagement included "How do you feel?" and "How do you like it?" To ask about engagement in a particular instruction type, the questions had appropriate referents. For example, to ask about student behavioral engagement in group work, the question was modified to “What do you typically do when you work on a problem with other students?” Questions were also customized based on the specifics of instruction types and based on what the students told me about instruction in their class. Lastly, probes were used to obtain more information when needed (e.g., to inquire about students’ reasons for engagement or disaffection).

Data analysis and major results. During the interviews, I focused on identifying potential indicators that I could later use in item writing. Besides this pre-determined goal, the interviews also provided implications (i.e., aspects to consider) for item writing. Beyond listening to students, no specific analytic strategy was used, as exploratory interviews turned out to be not very useful in generating items, compared to the literature review. Below, I will describe my major take-aways from exploratory interviews and discuss why exploratory interviews lacked usefulness.

During the interviews, students named a number of indicators (see examples in Table 3). While some of these indicators were later used in item writing, some (see crossed examples in the table) were discarded. For instance, indicators “do something else” and “chat with others” were not used because they indicate disaffection rather than engagement, which the instrument is designed to measure. Further, students often brought up feelings of confusion. Yet, this feeling was excluded from emotional engagement subscales because of its cognitive nature. Overall, exploratory interviews were useful in producing indicators for behavioral engagement; yet, they were less successful in producing indicators for cognitive and emotional engagement. I found that students had difficulty formulating their cognitive processes and articulating their feelings. For example, when asked about how they make sense of the material, students either did not know how to answer a question or referred to behaviors (e.g., “I listen to the professor” or “I draw a diagram”). A question about feelings or emotions, in turn, often resulted in general statements, such as “I feel fine.” Notably, these problems seemed to be specific to open-ended questions, as during cognitive interviews later, students did not tend to have problems answering specific cognitive and (most of) emotional items.

Table 3. Example indicators identified during exploratory interviews

	Lecturing	Whole-class interaction	Group work	Individual work
Behavioral Engagement	Listen Take notes ~~Do something else~~	Listen Take notes Ask questions Answer questions	Listen Discuss Ask questions Explain Check answers ~~Chat with others~~	Write Identify key words Draw a picture Look back at notes
Cognitive Engagement	Pay attention Follow	Pay attention Follow		Recall what I know
Emotional Engagement	Bored, Enjoy, ~~Confused~~			

Further, during exploratory interviews, students also shared information that had implications for the instrument development process. One of such findings was the occurance of situations when students cannot be engaged even when they want to. For example, a student does not answer a question in class if someone else answers it. Or, a student may be disengaged if they finish the work before the time is up. Or, a student does not help others because they did not know the material themselves. An implication for item writing was to avoid asking questions about what I call unavoidable disengagement. Going back to the examples, the item about answering questions in class was worded as “volunteering to answer.” To address the second example, no items were created that tap into this extra time. For the third example, indicators of providing or receiving help were excluded due to the possibility of making an inappropriate assumption (i.e., an assumption that any student can provide help or that help is requested).

With respect to answering questions, students also revealed that they would often answer questions in their heads, even if they do not respond aloud. The implication for item writing was to separate this item into two: one to indicate behavioral engagement (volunteering to answer instructor’s questions) and one to indicate cognitive engagement (answering instructor’s questions in the head). Further, some students mentioned behaviors that were not typical for the types of instruction, in the context of which these behaviors were mentioned. For example, students shared that while their instructor was lecturing, they would ask another student a question. The implication for item writing was to exclude the use of such atypical indicators and narrow down the conceptualization of behavioral and cognitive engagement to expected behaviors and cognitive processes, respectively.

Literature Review

From the literature review, I identified potential behaviors, cognitive processes, and emotions that I could use to indicate engagement dimensions of my instrument.

Specifically, I compiled lists of potential indicators for each engagement dimension. A behavior, cognitive process, or emotion was selected if (1) it was aligned with my conceptualizations of the corresponding engagement dimension, and (2) I expected that students would be able to estimate and self-report its frequency for each type of instruction over the course of the semester. Thus, behaviors, cognitive processes, and emotions, chosen as a starting point for the instrument development process, are presented in Table 4, Table 5, and Table 6, respectively. In Table 6, emotions in bold were emotions most frequently used in educational research. These emotions were considered first for the inclusion in the instrument for pretesting.

Table 4. Potential indicators of behavioral engagement

Verbal behaviors	Non-verbal behaviors
• Give feedback • Contribute to class discussions • Shared ideas / thoughts / information with others • Express opinions • Talk to other students • Try to work with others • Offering new ideas / comments • Interact with faculty • Try to answer when the teacher asks the class a question / Volunteer to answer • Ask a question	• Pay attention • Listen • Take notes • Drawing a diagram / Visualize • Do calculations / Compute • Re-read • Raise hand in class • Check • Review • Maintain attentive posture • Following along with both head and eyes • Sitting up straight • Make / Keep eye contact • Watch speaker's body language / paying attention to gestures, facial expressions, tone of voice, and stresses in speech

Table 5. Potential indicators of cognitive engagement

Selecting	Organizing	Integrating
• Selecting / Identifying / Noticing / Distinguishing between major and minor points / relevant and irrelevant points / substantial and trivial points • Differentiating between similar ideas	• Following • Organizing / Structuring / Devising a system • Paraphrasing / Translating into or Expressing in own words • Relating / Comparing / Contrasting / Connecting / Synthesizing / Summarizing / Putting together / Combining / Grouping or bracketing ideas or parts of text/lecture • Self-explaining • Analyzing / Analyzing logic	• Relating / Comparing / Contrasting / Connecting / Combining / Summarizing / Synthesizing between different sources; Resolving conflicts between information from different sources • Integrating / Relating / Connecting / Associating / Making analogies / Comparing / Contrasting / Synthesizing new things with prior knowledge or experience • Relating to practice or other contexts / Applying / Thinking of examples (concretizing) • Critical thinking / Questioning / Evaluating / Judging • Reconstruction / Re-evaluation of prior knowledge in light of new information

Table 6. Potential indicators of emotional engagement

Positive Activating	Negative Activating	Positive Deactivating	Negative Deactivating
• Determined	• **Angry**	• At ease	• Apathetic
• Elated	• Annoyed	• Calm	• **Bored**
• **Enjoying**	• **Anxious**	• **Comfortable**	• Droopy
• Enthusiastic	• Ashamed	• Content	• Dull
• **Excited**	• Bothered	• Placid	• Exhausted
• Hopeful	• Disturbed	• **Relaxed**	• Fatigued
• **Interested**	• Embarrassed	• Relieved	• Hopeless
• Proud	• Frustrated	• Serene	• Sleepy
	• Irritated		• Sluggish

Positive Activating	Negative Activating	Positive Deactivating	Negative Deactivating
	• Jittery • Mad • **Nervous** • Scared • Stressed • Tense • **Worried**		• **Tired**

Item Writing

In this section, I describe how I approached the development of the Instructional Time Form and the Engagement Survey.

Instructional Time Form. The Instructional Time Form for pretesting asked students to report the percentages of time in the lecture section of the class, spent on the instruction types (which were the same as in exploratory interviews). Also, as in exploratory interviews, students were asked to make sure that the total amount of time equals 100%. Different from exploratory interviews, for pretesting, I first excluded examples of instruction types but provided more precise definitions. The goal was to first gather information from students during cognitive interviews by asking them what kind of activities they did within each type of instruction. Based on this information, I modified the form by adding example activities, which are likely to be well understood by students and common in undergraduate mathematics-based classes.

Engagement Survey. Items for the engagement survey were developed (1) in accordance with the conceptualizations of each engagement dimension and instruction type, (2) using the identified behaviors, cognitive processes, and emotions, and (3) taking into account implications from exploratory interviews. In the student version of the instrument, items were grouped in blocks by instruction type. Specifically, a set of items measuring behavioral, cognitive, and emotional engagement in a particular instruction type were mixed together under the same item stem. An example stem for lecture was "During the time in class when your instructor explains the material without interacting with the class, how often have you…" All stems asked about the frequency of engagement.

During the process of item development, guidelines for item writing were consulted. Applicable recommendations were selected and used. Specific guidelines included (1) writing simple, short, clear, and direct statements, (2) avoiding double-barreled items and compound statements, (3) avoiding slang, (4) avoiding non-distinguishing statements, and (5) avoiding negatives (Wolfe & Smith, 2007). I also used the Question Appraisal System (QAS-99; Willis, 2015) as a source of potential problems I should avoid while writing items. Potential pitfalls to avoid, based on QAS-99, included making inappropriate assumptions, having vague language, including sensitive items, using technical terms, or asking about things that are difficult to recall. After the initial version of the instrument was developed, it was revised based on the results of cognitive interviews and expert reviews.

Cognitive Interviews

Evidence from cognitive interviews was used to examine the assumption within the cognitive modeling component of the substantive aspect of Messick's model. The assumption stated that observed response processes need to match the intended ones and be aligned with respondents' behaviors. Additionally, quantitative responses on the Engagement Survey were used to evaluate the assumption within the scale functioning component of the substantive aspect. This assumption stated that item response characteristics need to be consistent with expected characteristics.

Participants and recruitment. To recruit participants for cognitive interviews, I first compiled a list of potentially eligible classes (see more details about eligibility in the Target Population section) based on my knowledge about instructors and their teaching styles. All classes were in session during one semester. The original pool consisted of 28 sections (within 21 courses) taught by 19 instructors. I obtained the email information of students enrolled in these courses through the institutional Office of the Registrar (requested by my co-adviser, Dr. Margret Hjalmarson). Next, the instructors were contacted and asked to confirm the eligibility of their classes. Nine sections (within eight courses) taught by six instructors were reported as ineligible, according to the eligibility criteria (see the Target Population section). For three more sections (within three courses) taught by three instructors, I did not receive information; these sections were retained in the pool. The instructors were also asked to indicate if they would like their students to not be invited to participate in this study. No instructors made such an indication.

The final pool consisted of 19 sections (within 15 courses) taught by 15 instructors. Yet, the ratio of courses per instructor was not 1:1 as one instructor taught two courses, and one course was taught by two instructors. The courses ranged in the discipline, including mathematics, physics, astronomy, electrical engineering, civil engineering, and computer science. Class sizes also varied from 22 to 194 students, with an average of 75 (*SD* = 45). The classes were taught in different settings, including lecture halls and alternative active learning environments. Further, the courses represented three course levels: 100, 200, and 300. Finally, the courses also included those designed for non-STEM majors as part of their general education requirement. The process of participant selection was identical to the process used in exploratory interviews. In total, 66 students were interviewed. All students were interviewed for one particular class and only once, i.e., if a student was enrolled in several eligible classes, he/she was interviewed only for one of them. One student participated in both exploratory and cognitive interviews.

Cognitive interviews were conducted during two semesters (35 interviews in the first semester and 31 interviews in the second semester). In both semesters, the students were asked to respond with respect to the class they were enrolled in the first semester. The decision to keep the reference to the class in the first semester for the interviews conducted in the second semester was made for two reasons: (1) cognitive interviews in the second semester started at the beginning of the semester when students could not yet meaningfully report on their engagement in their current classes, and (2) the first wave of field-testing was scheduled for the end of the second semester, for which I needed a new set of classes, different from those used for cognitive interviews.

The sample was diverse in terms of student demographic and background information (see Table 7). The student average age was 20.83 (*SD* = 2.88), ranging from 18 to 33. Students also varied in majors, which included Computer Science, Electrical Engineering, Computer Engineering, Biology, Bioengineering, Chemistry, Civil

Engineering, Economics, Forensic Science, Mathematics, Mechanical Engineering, and more.

Table 7. Demographic and background information of cognitive interview participants

Characteristic	Frequency
Expected/actual grade in the course	
- A	27
- B	29
- C	6
- D	2
- F	2
Student classification	
- Freshman	10
- Sophomore	20
- Junior	20
- Senior	16
Status	
- Full-time	65
- Part-time	1
GPA	
- 3.51 or better	29
- 3.01 up to 3.50	23
- 2.51 up to 3.00	11
- 2.01 up to 2.50	1
- Less than 2.00	1
- Not applicable	1
Domicile	
- Domestic, in-state	48
- Domestic, out-of-state	12
- International	6
Native language	
- Yes	54
- No	12
Gender	
- Male	33
- Female	33
Race/ethnicity	
- White	38
- Black or African-American	8
- Hispanic or Latinx	6

Characteristic	Frequency
- Asian	12
- Other	2

Note. N = 66. Due to the timing of data collection across two semesters, data for student classification, status, and GPA indicates the information during the course or next semester.

Materials and procedure. All cognitive interviews were conducted on a university campus and lasted approximately one hour (with some exceptions). I was the only interviewer. All interviews were audio-recorded. Each student received a $25 Amazon gift card for their participation. Overall, three types of cognitive interviews were implemented.

First type of cognitive interviews. In the first type of cognitive interviews, I first described the study to students, and they signed a consent form. Next, students completed the Demographic Form, followed by the Instructional Time Form and the sorting task (Brewer & Lui, 1996). After that, a cognitive interview about the Engagement Survey was conducted. In total, 42 cognitive interviews of the first type were conducted.

Demographic Form. The Demographic Form included questions about students' learning in the course, their background, and demographic information.

Instructional Time Form. The Instructional Time Form that I used in the first type of cognitive interviews did not include examples. Students were asked to report the percentage of time they spent on each type of instruction and make sure that the total equals 100%. After students completed the form, I asked them to explain their responses. This information was used to (1) examine whether students understand the instruction type in the expected way and (2) identify example activities to include in the form.

Sorting task. For the sorting task, I created example activities within each instruction type. The development of these examples was informed by the information from cognitive interviews. I printed example activities on small cards and instruction types on large cards. Students were asked to sort the small cards (example activities) by large cards (instruction types). Two rounds of sorting were conducted with each participant. For the first round, students were instructed to select only those small cards that included activities applicable to their classes; students were told to put non-applicable small cards on a large "Not applicable" card. During the second round, students were asked to sort all of the non-applicable small cards, as if activities on them were applicable to their classes. After each round, I examined how students sorted the cards and asked them to explain why they classified a particular small card within a particular type of instruction if their classification differed from the one I hypothesized. Additionally, after each round of sorting, I recorded how students sorted the cards by putting sorted small cards into envelopes with the round number and the type of instruction (these data were not collected for one student). Initially, there were four large cards, with one per each instruction type. However, as interviews progressed, I added another large card about the time when students did not work on the task (see results for more details). Definitions of instruction types on the large cards did not always match the definitions in the Instructional Time Form, as the Instructional Time Form was updated more frequently to incorporate changes based on the cognitive interview data. This is a limitation of the procedure; yet, I consider this limitation to be minor as students also provided verbal reports during the sorting task. The overall goal of the Sorting Task was

(1) to examine whether students classify activities by instruction type in the expected way, thus exploring the similarity between their understanding of instruction types with hypothesized, and (2) to determine which of the example activities used are most often classified correctly and, thus, are best for including in the Instructional Time Form.

Engagement Survey. After the sorting task, students participated in the cognitive interview about the engagement items. Specifically, they were asked to read the items aloud one by one, answer each question, and explain why they answered in the way they did. Additional probes were used when necessary, e.g., to solicit a more elaborated explanation, to clarify students' responses, or to ask about the meaning of a particular term. Engagement items (and their variations) that were tested during cognitive interviews are presented in Appendix C.

The second type of cognitive interviews. The second type of cognitive interviews was conducted after I gathered sufficient information on the activities within each instruction type from the cognitive interviews of the first type. Thus, the second type of cognitive interviews differed from the first type in two ways. First, example activities were included in the Instructional Time Form. Second, the Sorting Task was excluded, as its major role was to help with developing example activities. In total, 12 cognitive interviews of the second type were conducted. Out of these 12 interviews, one interview was conducted earlier then others (after the first 15 interviews of the first type were conducted). Yet, after conducting it, I realized that the move from the first type of cognitive interviews to the second was premature, and I went back to the first type of cognitive interviews.

The third type of cognitive interviews. The third type of cognitive interviews took place concurrently with the second type. In the third type, students were first asked to complete a booklet that included a consent form, the Instructional Time Form (with examples), the Engagement Survey, measures of additional constructs needed for validation, and the Demographic Form, in this order. Students worked on the booklet on their own without interruption. While students were completing the booklet, I was measuring (via stopwatch) the amount of time the students spent on each page of the booklet (these data are available for all but one student due to technical problems). This process was conducted to determine an estimate for the time a student needs to complete the booklet. Students were also informed about this process before starting working on the booklet.

After students completed the booklet, a retrospective cognitive interview was conducted. Specifically, students were first asked about their general perceptions of the survey. Then, I probed students on some of the survey items of my selection. I typically selected items that were new, recently revised, or potentially problematic. Different from the first and second types of cognitive interviews, here I read the questions myself (as opposed to asking students to read them). However, the probes were the same. In total, 12 cognitive interviews of the third type were conducted.

Data analysis and results. Analysis of cognitive interviews was conducted on (1) verbal reports on the Instructional Time Form, (2) Sorting Task data, (3) quantitative responses on the Engagement Survey, and (4) verbal reports on the Engagement Survey. In this section, I describe my approaches to the analysis of each of these data and major results. For verbal reports, no specific analytic strategy was used beyond listening to students. Thus, no approaches to analysis of verbal reports are described.

Verbal reports on the Instructional Time Form. Students' verbal reports, i.e., explanations of their responses, on the Instructional Time Form revealed problems with the definitions of instruction types. The problems were identified as cognitive interviews progressed. One example of these problems was the inclusion of out-of-class activities, such as homework. To address this problem, the word "in class" was emphasized in the definitions. Another problem was concerned with classifying interactions with the instructor during individual or group work, with some students classifying these interactions in the whole-class interaction category, and others into group or individual work. To address this problem, I emphasized that the whole-class interaction category includes interaction with the class as a whole. Another problem was concerned with calculating percentages of instruction time. Some students reported percentages that did not add up to 100%. Formatting changes were made to make the task easier. Additionally, students' verbal reports on the Instructional Time Form suggested potential activities to be included as example activities, descriptive of instruction types.
For example, participants often mentioned an activity "Instructor lectures in a traditional sense;" thus, it was later included in the Instructional Time Form to illustrate Lecture.

Sorting Task data. Sorting Task data included quantitative data and verbal reports that students produced while doing the task. Data from the Sorting Task were used for two purposes: descriptive and reparative. The descriptive purpose refers to determining the best ways to describe instruction types to students, in terms of both definitions and example activities. The reparative purpose refers to identifying how my understanding of instruction types differs from students' understanding and to determining how to modify descriptions of instruction types to minimize the difference between students' and my understanding.

To analyze the quantitative sorting task data, I recorded the categories of instruction types in which each student sorted each card. Table 8 presents my approach to data organization. In the columns, there are cards, pre-classified by instruction types. Ommitted in the table are cards pre-classified by a combination of Individual and Group Work. This combined category was created for activities that students may misclassify by Whole-Class Interaction. Each row represents a classification of cards by instruction types for each participant. A numeric indicator of an instruction type in each cell indicates the sorting exercise: "1" for the first sorting exercise where students sorted only those cards that described activities applicable to their classes, and "2" for the second sorting exercise where students sorted all non-applicable cards. The table can be read as follows: Student #1 correctly classified all instruction types, whereas Student #2 misclassified Card #4 as Lecture (correct classification is Whole-Class Interaction), and Card #7 as Whole-Class Interaction (correct classification is Group Work). When referring to correct classification, I refer to the classification that I hypothesized. Finally, some cells may be blank, as in the present example for Student #1. These missing data occur when some cards were not used during the sorting task, as some cards were developed later in the process in reaction to discoveries from already conducted interviews.

Table 8. My approach to sorting task data organization (an example)

	Lecture (L)		Whole-class interaction (W)		Individual work (I)		Group work (G)		Not working on a task (N)
	Card 1	Card 2	Card 3	Card 4	Card 5	Card 6	Card 7	Card 8	Card 9
Student #1	L1	L1	W2	W1		I2	G1		N1
Student #2	L2	L1	W2	L1	I1	I2	W1	G1	N2
Student #3	L1	W1	L1	G1	I1	L1	G1	G1	I1

Next, I developed an aggregate matrix that produced classification frequencies across participants and sorting exercises (see Table 9). Specifically, I recorded the count of each card being classified by each instruction type across all participants. (Again, ommitted in the table are cards pre-classified by a combination of Individual and Group Work.) For example, Card #2 was classified as Lecture by one participant and as Whole-Class Interaction by two participants; no participants classified Card #2 as Individual or Group Work. Numbers that are at the intersection of the same instruction type (e.g., Lecture by Lecture) indicate the number of participants who correctly classified a card (presented in bold in Table 9). Other non-zero numbers, in turn, indicate incorrect classification.

Table 9. Aggregate matrix of the card by instruction type classification (based on the Table 8 example)

	Lecture		Whole-class interaction		Individual work		Group work		Not working on a task
	Card 1	Card 2	Card 3	Card 4	Card 5	Card 6	Card 7	Card 8	Card 9
Lecture	**3**	**2**	1	1	0	1	0	0	0
Whole-class interaction	0	1	**2**	**1**	0	0	1	0	0
Individual work	0	0	0	0	**2**	**2**	0	0	1
Group work	0	0	0	1	0	0	**2**	**2**	0
Not working on a task	0	0	0	0	0	0	0	0	**2**

To determine which activities tended to be classified correctly by the hypothesized instruction type and can be used to describe the instruction types, I developed the correct classification rate for each card. The correct classification rate was calculated by dividing the number of times a card with the activity was correctly classified by the total number of times this card was included in the sorting task. I used any classification rate above 80% as indicative of activities perceived to be descriptive of the instruction types. In turn, any rate below 80% indicated activities perceived not to be descriptive of the instruction types. The rate close to 100% was considered ideal as it indicated that the activity was consistently associated with a particular instruction type.

Additionally, to identify activities that were or were not applicable to participants' classes, I developed the selection rate with which the activities were selected during the first sorting exercise, as opposed to the second one. This rate was calculated by dividing the number of times the card with the activity was selected in the first sorting exercise by the total number of times it was included in the sorting task. I used any rate below 60% as indicative of activities not applicable to participants' classes; in turn, any rate above 60% indicated applicable activities. The rate close to 100% was considered ideal, as it indicated that the activity occurred in the classes of most participants, in participants' perception.

Both rates were used in the analysis for the descriptive purpose. Specifically, as the goal of conducting the Sorting Task was to identify example activities that can be included in the Instructional Time Form to illustrate the instruction types, activities to be selected needed to have both rates high. In other words, activities needed to be (1) perceived to be descriptive of the instruction types and (2) actually occur in students' classes so that respondents can relate to the activities. In the analysis for the reparative purpose, only the correct classification rate was used. Specifically, low correct classification rates were further analyzed to explore misclassification of activities in order to understand how activities can be modified to become descriptive of instruction types, from the students' perspective.

Descriptive purpose. Correct classification and selection rates were analyzed to identify activities that can be used to illustrate instruction types in the Instructional Time Form. The correct classification rate (see Table 10) was above 80% for all activities within Whole-class Interaction and Group Work, as well as for some activities within Individual Work. This finding suggests that these activities were perceived to be descriptive of the corresponding instruction types. For example, the correct classification of the Individual Work activity "I work on a worksheet individually" was 95.12%, meaning that this activity was said to illustrate Individual Work by 95.12% of participants. Other activities, including the activities hypothesized to illustrate Lecture or a joint category of Individual or Group Work, did not tend to be correctly classified by participants (the correct classification rate lower than 80%), suggesting that these activities were not perceived to be descriptive of the corresponding instruction types. For example, the correct classification rate of the Lecture activity “Instructor goes through the problems students were asked to work on” was only 26.83%, meaning that only 26.82% of students thought that this activity illustrates Lecture.

The selection rate at the first sorting exercise (see Table 10) was high (above 60%) for most of the activities, suggesting that these activities are applicable to students’ classes. For example, the activity “The instructor talks through the material” had the selection rate of 100%, indicating that this activity occurred in the classes of all participants. The low selection rate (below 60%) was rare but yet appeared for several activities, such as “I work on a worksheet individually” (48.78%) and “Students talk about projects with each other” (56.10%). Thus, these activities did not frequently occur in students’ classes.

Interestingly, for some activities, one rate was low, whereas another rate was high. For example, the correct classification rate of the Lecture activity “Instructor goes through the problems students were asked to work on” was low (26.83%), yet its selection rate was high (90.24%), suggesting that the activity was not descriptive of the hypothesized instruction type but was applicable to participants’ classes. An opposite example is the Individual Work activity “I work on a worksheet individually,” which had a high correct classification rate (95.12%) but a low selection rate (48.78%), suggesting that the activity was descriptive of the hypothesized instruction type but was not very applicable to participants' classes. Although these activities were descriptive of the instruction types or were applicable to participants' classes, they were not considered for inclusion in the Instructional Time Form (the ultimate goal of the sorting task), as activities needed to have both rates high in order to be included in the form. An example of the activities eventually included in the Instructional Time Form is the Whole-Class Interaction activity "Instructor asks questions to the whole class." This activity had the correct classification rate of 87.80% and the selection rate of 100%, suggesting that it was both descriptive of the hypothesized instruction type and applicable to participants' classes.

Reparative purpose. Low correct classification rates were further analyzed to explore the misclassification of activities. For example, for Lecture, I found that in most cases misclassified activities were classified within Whole-Class Interaction. Verbal reports revealed the reason why this misclassification occurred. In the perception of participants, their instructors did not tend to do these activities (e.g., talk through the material) without interacting with the class. Thus, this information was used to modify the description of the Lecture instruction type in the Instructional Time Form.

Similarly, activities designed to be descriptive of Individual or Group Work tended to be frequently misclassified as Whole-Class Interaction. Verbal reports showed that this misclassification occurred due to the differences in understanding of Whole-Class Interaction between the participant perception and my definition. Specifically, participants thought that if an instructor walks around the classroom to talk with students while they are working on a task, then the instructor interacts with the class. However, in my view, such interactions were examples of Individual or Group Work, i.e., the time when students worked on the task. The identification of this problem led to the modification of the description of the Whole-Class Interaction instruction type in the Instructional Time Form.

Table 10. Results of the quantitative sorting task data

Activities and hypothesized instruction types	N	Correct classification rate (%)	Selection rate at the first sorting exercise (%)
Lecture			
Instructor talks through the material	41	65.85	100.00
Instructor shows or discusses examples	41	46.34	97.56
Instructor goes through the problems students were asked to work on	41	26.83	90.24
Whole-class interaction			
Instructor asks questions to the whole class	41	87.80	100.00
Instructor asks, “Who can remind me what this term means?”	41	87.80	92.68
Other students ask questions to the instructor in front of the whole class	41	90.24	95.12
Instructor does interactive lecture	41	92.68	87.80
Individual work			
I work on a worksheet individually	41	95.12	48.78
I solve a problem by myself	41	87.80	78.05
Instructor asks students to think about a question	41	9.76	95.12
I try to solve a problem by myself before turning to other students	41	75.61	90.24
I continue to solve a problem by myself after talking to other students	27	59.26	77.78
Group work			
Students work on a worksheet together	41	100.00	56.10
Students discuss homework with each other	41	82.93	68.29
I ask other students for help with a problem	41	92.68	87.80
Students check answers with their classmates	41	92.68	90.24
Students talk about projects with each other	41	87.80	56.10

Activities and hypothesized instruction types	N	Correct classification rate (%)	Selection rate at the first sorting exercise (%)
Students work in a group	41	97.56	75.61
Instructor asks students to discuss a question with their classmates	41	82.93	87.80
Individual or Group work			
I ask the instructor a question while solving a problem	27	40.74	85.19
The instructor comes to check on how students are solving a problem	27	40.74	85.19
I interact with the instructor while solving a problem	27	40.74	85.19
Not working on a task			
I wait for the professor to help me with the problem	19	26.32	84.21
I wait for the instructor to explain how to solve a problem	19	26.32	100.00
I talk to other students about things not related to the task	13	84.62	46.15
I do other things when I am supposed to work on the task	14	78.57	50.00
I decide not to work on a problem	19	73.68	52.63

Quantitative responses on Engagement Survey. All engagement items (and their variations) that were tested during cognitive interviews are presented in Appendix D. The initial analysis included an examination of descriptive statistics and frequencies of engagement items (see Appendix E). The results indicated that some items showed low variability and negative skewness. These items were behavioral and cognitive engagement items that were commonly endorsed by students, i.e., behaviors and cognitive processes that are easy for students to do. To address the issues of low variability and negative skewness, three methods were employed. First, I tried to change some items to extreme meaning (e.g., “paying attention” was changed to “giving full attention”). However, for some items (e.g., “Read what the instructor is writing or showing (e.g., instructor’s notes, PowerPoint slides, etc.)”), transformations to extreme meaning did not sound natural (e.g., “Read all of what the instructor is writing or showing (e.g., instructor’s notes, PowerPoint slides, etc.)”). Thus, for such items, transformations were not used. Second, I expanded the 5-point response scale to a 7-point response scale. Third, I removed items that showed ceiling effect. For example, I removed items about trying to understand instructor’s explanations during lecture, trying to follow what is being said during whole-class interaction, trying to relate the task to existing knowledge during individual work, and listening to other students during group work.

Some items that showed low variability and negative skewness were retained for field-testing for two reasons. First, these items were needed to preserve content validity. Second, given the small sample of cognitive interview participants, I was not sure if the items would have the same characteristics when administered to a larger sample. It was possible that these characteristics were an artifact of student self-selection into cognitive interviews. The problem with the quantitative analysis of responses from cognitive interviews was not only in the small number of participants but also in the small number of responses per item. Not every student completed all sections of the survey because some sections were not applicable (i.e., the amount of time spent on the corresponding instruction type was indicated as zero). Sometimes, not all survey blocks were completed during the interview due to the lack of time. Further, for some students on some items, data were not available because items were added or removed from the survey based on the results on cognitive interviews. Finally, a limitation of the quantitative analysis of responses from cognitive interviews was combining different variations of an item into a single variable. Multiple versions of the same items occurred due to making changes to the items based on the participants' feedback in an attempt to improve the items. Thus, it was unclear whether the distribution of item responses was affected by responses on the versions of items that were later improved.

Verbal reports on Engagement Survey. From students' verbal reports on Engagement Survey, I aimed to identify problems with the instrument. The problems may be explicit (i.e., problems raised by participants) or implicit (i.e., problems that were evident from the participants' responses, although not recognized by the participants themselves). An example of the former is a student's question about the meaning of the word; an example of the latter is a mismatch between the respondent's interpretation of the question and the intended meaning. Additionally, problems may be differentiated between two levels. At the higher level, problems are concerned with the section stem (e.g., a stem for engagement within Lecture). At the lower level, problems are concerned with items (i.e., items on the Engagement Survey). In identifying problems, I used the problem classification coding scheme (Forsyth et al., 2004) for guidance. This scheme is a hybrid coding scheme that combines cognitive coding (i.e., categories based on the cognitive model of the survey response process, Tourangeau, 2000) and question feature coding (Willis, 2015). The cognitive categories of Tourangeau (2000) include comprehension, retrieval, judgment, and response. The problem classification coding scheme classifies problems within sub-categories where appropriate, and further groups them by the cognitive categories (see Table 11). I used the information from the coding scheme as a guide for the kind of problems that may occur and, therefore, I should look for. However, I did not constrain myself to this scheme and examined what students say for any kind of problems. As problems were identified, I revised the items in an attempt to address or at least minimize the problems. After revisions, the item was further pretested during next cognitive interviews. Below, I describe major problems that I identified from students' verbal reports on the Engagement Survey.

Table 11. The problem classification coding scheme (Forsyth et al., 2004, p. 530)

Cognitive categories	Sub-categories	Problems
1. Comprehension and communication	Interviewer difficulties	1. Inaccurate instructions 2. Complicated instruction 3. Difficult to administer 4. Vague topic/term
	Question content	5. Complex topic 6. Topic carried over from earlier question 7. Undefined term(s) 8. Transition needed
	Question structure	9. Unclear respondent instruction 10. Question too long 11. Complex, awkward syntax 12. Erroneous assumption 13. Several questions 14. Carried over from earlier question 15. Undefined
	Reference period	16. Unanchored or rolling
2. Memory retrieval		17. Shortage of cues 18. High detail required or information unavailable 19. Long recall period
3. Judgment and evaluation		20. Complex estimation 21. Potentially sensitive or desirability bias
4. Response selection	Response terminology	22. Undefined term(s) 23. Vague term(s)
	Response units	24. Responses use wrong units 25. Unclear what response options are
	Response structure	26. Overlapping categories 27. Missing categories

The biggest problem at the section stem level was related to the attribution of the engagement items. Specifically, students tended not to pay attention to the instruction type, to which items refer. A number of changes to the mode of administration and format were made. At the beginning, items were administered one by one on small cards to help students focus on one item at a time. Each card included an item stem, an item itself, and response options. Due to the stem being the same for a set of items (one set for each instruction type), cards looked partially repetitive, and students tended to stop reading the stem and forgetting about the attribution. Hypothesizing that the problem with attribution may have been related to the mode of administration (cards), I changed the mode. Specifically, I put each set of items (with response options) into a table with the stem in the table heading. Each table was on a separate page. In both modes, I also tried including the reference to the instruction type into the items themselves, but it did not solve the problem. Instead, the items became burdensome to read.

Several formats of the tables were tried, improving students' attention to the referent. Further, modifications for formats of the pages with engagement items were made in order to further mitigate the problem of attribution. Aspects of the page format retained for field-testing was the use of paper of different color for each page with engagement items, as well as the use of titles with instruction types on the top of the pages. Additionally, different combinations of including and excluding referents from the items were tried, with the final decision being to include referents in the emotional items (as the same emotions are used across instruction types) but exclude from the behavioral and cognitive items (these items are typically worded in a way that makes the referent more explicit). Notably, the problem of attribution is not unique to my measure, as similar problems occurred with other measures, as well (e.g., Watt et al., 2008).

A number of problems at the item level were also identified. For example, terms, unfamiliar to participants, were found in some emotional items. Specifically, some students did not know the meaning of the words “sluggish” or “apathetic.” Other emotional items referred to emotions that students did not find applicable to their learning in class. Such emotions included feeling “content” and “comfortable.” These problematic emotions were removed from the survey and replaced with other emotions. Further, behavioral subscales initially included psychomotor items (e.g., “Followed the instructor along with your head and eyes” and “Turned to a student when he/she was speaking”). These items did not elicit expected responses. For example, students would not follow the instructor when writing, which does not suggest a lack of engagement. Therefore, these items were excluded. Another example of items that elicited an unintended meaning was an item about trying to determine if what the instructor is saying is worth paying attention to. Whereas some students interpreted the item as intended (i.e., distinguishing important information from unimportant), other students reported that they do not do that because they believed that everything the instructor says is worth paying attention to. With such an interpretation, the item was no longer a valid indicator of engagement and, therefore, was removed.

Finally, students’ verbal reports on Engagement Survey also provided an insight for the Instructional Time Form. In particular, students’ explanations of responses for individual and group work items revealed that these instruction types capture only the time when students do work on a task. Yet, the time when students decide not to work on a task was not captured anywhere in the instrument. To address this problem, a fifth category of deciding not to work on a task was added to the Instructional Time Form.

Conclusion. During cognitive interviews, I identified a variety of problems with the instrument. As a result, many revisions were made in an attempt to improve it. While I cannot say that all problems were eliminated during the process, it may be fair to say that the efforts to identify and address the problems likely resulted in the reduced number of problems or alleviated the severity of the problems. Thus, I provided some evidence for the assumption within the cognitive modeling component of the substantive aspect of Messick's model. In particular, I provided some evidence for the similarity between observed and expected response processes. Further, I also examined frequencies and descriptive statistics of the items administered during cognitive interviews. While these data were limited in scope, I found that some items were normally distributed, but others were highly negatively skewed. Some of the latter items were removed, whereas other items were retained for field-testing due to the need to preserve content validity or due to the possibility that the items may show different characteristics in a larger sample. Thus, I also provided some evidence for the assumption within the scale functioning component of the substantive aspect of Messick's model. In particular, I provided some evidence for the consistency between observed and expected item response characteristics.

Expert Reviews

Expert reviews were conducted to evaluate the assumption within the content aspect of Messick's model. This assumption states that items need to be relevant to and representative of the construct being measured.

Participants and recruitment. For expert reviews, I sought to recruit researchers with expertise in the area of student engagement and/or educational measurement. During the first round of expert reviews, three experts were recruited; during the second round, two experts were recruited. The second round of reviews differed from the first round mainly in the items, as they were revised based on the results from the first round and from cognitive interviews. Across rounds, four experts gave their permission to be identified: Dr. Jennifer Fredricks, Dr. Gwen Marchand, Dr. Benjamin Heddy, and Dr. Jacob Marszalek.

Materials and procedure. All experts received three documents in their invitation email: the consent form, the Expert Review Form (see Appendix F for the form used in the first round and Appendix G for the form used in the second round), and the student version of the survey. Experts were asked to email the signed consent form and the completed Expert Review Form back to me. Alternatively, experts were also given an option to participate in an interview if they preferred an interview to completing the form. However, no expert chose the interview option. The Expert Review Form included the description of the construct definition, the intended use of the instrument, and expert review instructions. It should be noted that the intended use of the instrument was reformulated as the study progressed and, therefore, has been changed since the expert reviews were conducted. However, the main idea that the instrument can be used by both researchers and practitioners remains. Experts were asked to rate clarity and relevance of each item in the Instructional Time Form and in the Engagement Survey; they were also asked to rate subscale representativeness for each engagement subscale. Rating scale included the following options: “1 = Not Acceptable (major modifications needed),” “2 = Below Expectations (some modifications needed),” “3 = Meets Expectations (no modifications needed but could be improved with minor changes),” and “4 = Exceeds Expectations (no modifications needed)” (Ramirez, 2016). Additionally, experts were asked to provide suggestions for item revisions or to make other comments. Feedback on response options was solicited as well. Finally, in the second round, experts were also asked to comment on the student version of the instrument.

Data analysis and results. An analysis of expert reviews includes both an analysis of ratings and a review of comments. In the analysis of experts’ ratings, I explored experts’ ratings of item clarity, item relevance, and subscale representativeness, looking for items or subscales that commonly received low ratings. Due to the small number of experts per round, no descriptive statistics for items and subscales were computed. However, I computed descriptive statistics of ratings for each expert. Further, I also examined interrater reliability via a variation of the kappa statistic that measures agreement among multiple raters (Fleiss, 1981). This kappa statistic ranges from negative one (indicating perfect disagreement among raters) to positive one (indicating perfect agreement), with zero indicating agreement not different from the chance alone. I calculated kappa for item relevance, item clarity, and subscale representativeness for each round of review using SAS Macro (Chen et al., 2005). For this analysis, ratings that were indicated as a range or as a decimal were excluded (i.e., treated as missing). Finally, reviewing experts' comments, I aimed to identify main problems with the instrument raised by the experts and suggestions they made to address these problems.

Ratings and comments from the first and second rounds of expert reviews are presented in Appendix H and Appendix I, respectively. Descriptive statistics of expert ratings for each rater are presented in Table 12. Kappa statistics are presented in Table 13. The majority of kappa values were low or even statistically non-significant, suggesting that raters did not tend to agree with each other. Despite the lack of agreement, raters, on average, rated item relevance and clarity above 3, i.e., "Meets expectations (no modifications needed but could be improved with minor changes)." Subscale representativeness was rated, on average, lower than 3 in the first round of reviews. However, there seemed to be an overall improvement from the first to the second round of reviews, with average ratings of item relevance, item clarity, and subscale representativeness increased.

Table 12. Descriptive statistics of expert ratings

	Item Relevance		Item Clarity		Subscale Representativeness	
	Mean	*SD*	Mean	*SD*	Mean	*SD*
Round 1						
Expert #1	3.42	0.70	3.28	0.71	2.70	0.48
Expert #2	3.17	0.80	3.22	0.54	2.82	0.40
Expert #3	3.59	0.67	3.66	0.69	-	-
Round 2						

	Item Relevance		Item Clarity		Subscale Representativeness	
Expert #4	3.80	0.47	3.79	0.45	3.42	0.67
Expert #5	3.99	0.12	3.91	0.28	4.00	0.00

Note. In Round 1, subscale representativeness was rated only by two experts. The number of maximum possible ratings for relevance was 66 for Round 1 and 69 for Round 2. The number of maximum possible ratings for clarity was 67 for Round 1 and 70 for Round 2. The number of maximum possible ratings for representativeness was 12 for both rounds.

Table 13. Kappa statistics for expert ratings

	Item Relevance		Item Clarity		Subscale Representativeness	
	Kappa	*SE*	Kappa	*SE*	Kappa	*SE*
Round 1 (3 experts)	0.16**	0.05	0.10*	0.05	0.46*	0.27
Round 2 (2 experts)	0.08	0.11	0.04	0.12	0.11	0.25

Note. * p < 0.05, ** p < 0.01. *SE* = standard error.

One problem commonly brought up by experts was the similarities between behavioral and cognitive items. For example, the item that intended to measure behavioral engagement in Group Work “Compared your and other students’ solutions/answers or ways of thinking about the task?” was suggested to be cognitive by Expert #1 and Expert #2. This item also had low ratings of relevance. To address the concern, I split the item into two, with one asking about comparing ways of thinking (cognitive) and another asking about matching answers, solutions, or approaches (behavioral). As another example, the item that intended to measure behavioral engagement in Individual Work “Tried different ways of solving or thinking about the task even if you already have an answer?” was also suggested to be cognitive by Expert

#5. Thus, I split this item into two, with one focused on thinking about different ways (cognitive) and another focused on writing down more than one way. Further, an item with low relevance ratings was the item about paying attention to what the instructor is explaining. This item was intended to indicate cognitive engagement in Lecture. However, all three experts in the first round suggested that the item seemed to rather be behavioral. As recommended by Expert #1, as well as in order to increase the variance of this item, I changed the wording to “giving full attention.”

Next, consistently with the evidence from cognitive interviews, Expert #1 and Expert #3 suggested including referents to instruction types in emotional items. The short forms of referents for emotional items were added in the second round of reviews, where issues with respect to these referents were not raised. Further, Expert #1 also suggested expanding the referent for Whole-Class Interaction items from “what is being said” to include the reference to interaction in order to emphasize the difference between these items and lecture items. I expanded the referent to “what is being said between your instructor and other students” in the field-testing version of the instrument.

The formatting of emotional items received other concerns. Specifically, Expert #1 and Expert #2 pointed that items about enjoyment are worded differently from other emotional items (e.g., “enjoyed listening to your instructor” vs. “felt interested” or “felt frustrated”). Further, Expert #5 pointed that emotional subscales were unbalanced in terms of valence and activation, with two items being positive activating, two items being negative activating, one item being positive deactivating, and one item being negative deactivating. I aimed to have one item of each in the final instrument. Deactivating items were easier to select based on cognitive interviews, as other tested deactivating items showed problems. However, activating items tended to work well and, therefore, were harder to select. I decided to keep six items for field-testing and select one item from each pair of activating items during the field-testing data analysis. Keeping two positive activating items (enjoyment and excitement) also allowed me to see whether the enjoyment item is problematic due to its unique wording and to replace it with the excitement item if needed.

More suggestions were made about the wording of specific items. For example, Expert #1 hypothesized that the term “posed” in the item about posing questions in class may be unfamiliar to students. I replaced the word with “asked.” In terms of subscale representativeness, several suggestions for emotions were made. Per the recommendation of Expert #1 and Expert #2, I included “anxious” in the emotional subscales. Expert #2 also suggested a revision for the item about verifying one’s work or answer on the task with the task instructions/question to checking that the work fits with the task instructions/question. I implemented this revision, developing the item “Checked that your work or answer on the task fits with the task instructions/question.”

While I tried to address many concerns and suggestions, I was not able to address all of them. For example, Expert #2 suggested differentiating referents of Whole-Class Interaction items between the instructor and other students. I tried doing that (and tested such items at the beginning of cognitive interviews), as I hypothesized that students might be engaged in whole-class interaction differently, depending on who is speaking at the moment. However, with such a split, the scale was getting too long. Therefore, I used the interaction as a single referent. Further, Expert #3 and Expert #5 expressed a concern about the critical thinking items, as critical thinking may be an unclear term and may solicit multiple interpretations. However, during cognitive interviews, I learned that students tended to interpreted critical thinking as deep thinking or as evaluating the information. Both interpretations fit with the intended meaning of thinking beyond some basic level.

Considering items that experts suggested for inclusion, I decided not to include some of them. For example, Expert #1 suggested seeking help as an indicator of behavioral engagement in Individual Work. I ultimately decided not to use this item due to concerns that help-seeking behaviors are likely to be conditional on the student's need for help. Thus, if a student did not need help, then reporting no help-seeking on the survey would indicate a lack of engagement for wrong reasons. For behavioral engagement in Group Work, Expert #1 suggested including items, such as helping to set group work rules, working on group documents, and locating resources to help with group tasks. I ultimately decided not to use such items because they imply a structured type of group work. However, I aimed for my instrument to be applicable for any type of group work, including both structured and unstructured.

Finally, I want to highlight the issue of frequency vs. quality of engagement, raised by Expert #1 and Expert #4. Expert #1 noted that one could argue that different items may reflect different quality of engagement. In my view, quality of engagement is represented in items. Some items should be more commonly endorsed than others. However, I did not include qualitatively different variations of the same item (e.g., taking brief notes vs. taking elaborated notes) because the lack of endorsement could mean either lack of engagement overall (not taking any notes) or taking notes of a different kind.

Conclusion. From expert reviews, I identified problematic items and subscales. Although I was not able to address all concerns and incorporate all suggestions provided by experts, I tried to address as many concerns and incorporate as many suggestions as I could. Further, quantitative ratings showed that relevance, clarity, and representativeness were adequate, even considering that the degree of agreement between experts was low. Thus, I provided some evidence for the assumptions of item relevance and subscale representativeness within the content aspect of Messick’s model.

Pretesting the Online Version of the Instrument

The instrument is intended to be administered in both paper-and-pencil and online formats. Thus, in addition to the paper-and-pencil survey, the online survey was also pretested. As items were already pretested via cognitive interviews described above, cognitive interviews conducted on the online survey had a different focus. Specifically, these interviews aimed to find problems with the online interface. Pretesting the online version of the survey, in addition to pretesting the paper-and-pencil version, was one way to evaluate an assumption about generalizability across formats within the generalizability aspect of Messick’s model. This assumption states that responses need to generalize across paper-and-pencil and online version of the survey. Pretesting both formats helps to identify and fix any problems specific to a format.

The online survey was very similar to the paper-and-pencil survey, as I tried to retain the formatting as much as possible (e.g., different colors for pages with engagement items within each instruction type, shading of alternate rows in tables with items, bolding or underlining where necessary, etc.). Shading of alternate rows in the online version was done with white rather than grey, as white seemed to be better for reading on the screen. The biggest difference between the paper-and-pencil and online versions was in the Instructional Time Form. First, in the online version, I was able to force participants to provide percentages of different instruction types that add up to 100%. Otherwise, the survey would show an error message. Second, if 0% was provided for a particular instruction type, then the page/block with engagement items for this instruction type was not administered. Another difference of the online survey from the paper-and-pencil survey was a forced response to the items so that no missing data occur from online data collection (with the exception of missing engagement blocks discussed above or not finished surveys).

Additionally, I included attention checks in the online version of the survey to be able to identify careless respondents. Careless, or insufficient effort, responding is “a response set in which the respondent answers a survey measure with low or little motivation to comply with survey instructions, correctly interpret item content, and provide accurate responses” (Huang et al., 2012, p. 100). Identifying and removing careless respondents increases data quality. Employing attention checks – specific items designed to catch careless responding – is one way of identifying careless respondents. Thus, attention checks were also pretested during cognitive interviews.

Participants and recruitment. Recruitment of participants for cognitive interviews conducted on the online version of the instrument was done in two steps. During the first step, I contacted instructors of potentially eligible classes and asked them to share an invitation to participate in my study with their students. As the interviews were conducted at the beginning of the semester, classes were selected from the previous semester. In the invitation email shared with students, there was a link to a sign-up form. During the second step, I contacted students who completed the sign-up form. Overall, two instructors shared the invitation email with their classes. Selecting students from the sign-up list, I invited students who signed up earlier, taking class into account. I aimed to have a representation of students from different classes. The classes were in two disciplines: Statistics and Civil Engineering. In total, six students from two classes were interviewed. The sample was diverse in terms of student demographic and background information (see Table 14). Participants' GPA, on average, was 3.18 (*SD* = 0.33). Students' average age was 20.17 (*SD* = 1.17), ranging from 18 to 21. All students were full-time students.

Table 14. Demographic and background information of cognitive interview participants (online version)

Characteristic	Frequency
Expected/actual grade in the course	
- A	3
- B	3
Major	
- Civil Engineering	4
- Environmental and Sustainability Studies	1
- Statistics	1

Characteristic	Frequency
Student classification	
- Sophomore	2
- Junior	4
Domicile	
- Domestic, in-state	4
- Domestic, out-of-state	2
Native language	
- Yes	5
- No	1
Gender	
- Male	4
- Female	2
Race/ethnicity	
- White	2
- Hispanic or Latinx	1
- Asian	1
- Mixed	2

Materials and procedure. All cognitive interviews were conducted on a university campus and lasted approximately 20-60 minutes. I was the only interviewer. All interviews were audio-recorded. Each student received a $25 Amazon gift card for their participation. At the beginning of the interview, I explained the procedure to the student. The procedure included two parts. First, students completed the online survey on their own device of their choice. The survey was set up on the Qualtrics platform and had two versions: one for personal computers (PCs) and one for mobile devices. Five students completed the survey on laptops, and one student completed the survey on a cell phone. Next, I conducted the interview. During the interview, I asked about students' experiences taking the survey, different aspects of the formatting (e.g., colors, font, etc.), attention checks, suggestions for survey improvement, and more. To facilitate the interview and help students remember formatting for different pages of the survey, I printed parts of pages and showed them during the interview.

To be able to identify careless respondents, I included attention checks in the online version of the survey. Although I did not use attention checks in the paper-and-pencil survey, I aim to use them in the paper-and-pencil format in the future. Two types of attention checks were employed. One type is the items constructed using the infrequency approach (e.g., Huang et al., 2015). This approach uses items, responses to which are expected to be the same or almost the same. One subtype of such items is instructed response items, for example "To monitor quality, please respond with a two for this item" (Meade & Craig, 2012). I used instructed response items for measures of additional constructs needed for validation, as these measures consisted of statements. Thus, an instructed response item, which is also a statement, did not seem to be out of place among items in these measures. Specifically, I used two instructed response items for the 29-item survey with measures of class-specific constructs: "To monitor quality, please select "Disagree" for this item" and "To monitor quality, please select "Strongly Disagree" for this item." For the 22-item survey with measures of general constructs, I used one instructed response item: "To monitor quality, please select "Somewhat Agree" for this item." For more details about these measures, see the section about field-testing.

In the Engagement Survey, I was not able to include instructed response items because engagement items were designed as questions rather than statements. Thus, for the Engagement Survey, I designed infrequency items as questions that students should not be likely to endorse. The tested infrequency items for each type of instruction are presented in Table 15. I aimed to select one infrequency item for each engagement survey page (i.e., instruction type). Specifically, I selected items that consistently elicited "never" or close to "never" responses.

Finally, the second type of attention checks is self-report measures of response quality (Meade & Craig, 2012). In particular, I used an indicator of a student's evaluation of the quality of the data they provided (Meade & Craig, 2012). The item was worded as follows: "For the study to produce accurate results, it is important to include data only from people who carefully read the questions and answered them thoughtfully. In your honest opinion, should we use your data in our analyses? You will receive credit for participating in this study no matter how you answer this question." Response options included "yes" and "no."

Data analysis and results. Data included students' verbal reports during cognitive interviews and their quantitative responses to the survey. For verbal reports, no specific analytic strategy was used beyond listening to the students. For quantitative responses to the survey, responses to attention checks were of a particular interest. Thus, only those responses were analyzed.

Overall, from students' verbal reports, I did not identify any technical difficulties or major problems with the formatting/interface. When asked about engagement items, designed to be infrequency items, students did not have negative feedback. Some students indicated their lack of engagement in response to these items strongly. Such reports suggest that the items work as intended, soliciting "never" responses. Responses to all infrequency engagement items are presented in Table 15. Not all infrequency items were administered to all participants, as some items were added later or replaced due to poor functioning. Additionally, some infrequency items were included in the survey but not administered to participants if the entire engagement block was not administered. Infrequency items selected for the use in field-testing included “Purposefully disrupted the class?” for Lecture, “Made a rude comment in front of the whole class?” for Whole-Class Interaction, “Cried working on the task in class alone?” for Individual Work, and “Refused to explain your thinking about the task to another student?” for Group Work. These items consistently solicited “Never” or “Almost Never” responses on a 7-point scale from “Never” to “Always.”

Table 15. Infrequency attention checks for the Engagement Survey tested during cognitive interviews (online version)

Attention Check	Student #1	Student #2	Student #3	Student #4	Student #5	Student #6
Lecture:						
Purposefully disrupted the class?	Never (1)	Never (1)	Never (1)	n/a	Never (1)	Almost Never (2)
Tried to distract other students when the instructor is explaining the material?		Never (1)	Almost never (2)	n/a	Never (1)	Rarely (3)
Whole-Class Interaction:						
Made fun of another student in front of the whole class?	Never (1)	Never (1)	Almost never (2)	Never (1)	Never (1)	Never (1)
Made a rude comment in front of the whole class?		Never (1)	Never (1)	Never (1)	Never (1)	Never (1)
Individual Work:						
Re-written task instructions word-for-word multiple times before starting working	Never (1)	Never (1)	Often (5)	n/a	n/a	n/a

Attention Check	Student #1	Student #2	Student #3	Student #4	Student #5	Student #6
on the task?						
Cried working on the task in class alone?		Never (1)	Never (1)	n/a	n/a	n/a
Group Work:						
Copied down solutions of everyone you talked to?	Sometimes (4)	Sometimes (4)	Often (5)			
Refused to explain your thinking about the task to another student?		Never (1)	Never (1)	Never (1)	Never (1)	Never (1)
Been rude to another student?				Never (1)	Never (1)	Never (1)

Note. If a cell is blank, the attention check was not included in the survey. If a cell is marked as "n/a," the attention check was included in the survey but was not administered to a student because the student indicated in the Instructional Time Form that the instruction type was not present.

Next, cognitive interviews showed that instructed response items functioned well. When asked about these items, students expressed an understanding of the reason why these items were included in the survey. Some students also noted their familiarity with such items based on their prior experience completing surveys. Finally, the item that measures self-reported data quality was also perceived well by students. Later, I thought that the word "credit" in this item might be misinterpreted as course credit rather than a gift card for study participation. Thus, for field-testing, the item was adjusted as follows: "For the study to produce accurate results, it is important to include data only from people who carefully read the questions and answered them thoughtfully. In your honest opinion, should we use your data in our analyses? You will be entered in a raffle with a chance to win one of five $50 gift cards (if you provided your email at the beginning of the survey) no matter how you answer this question."

Overall, cognitive interviews conducted to pretest the online version of the survey showed that the online survey was ready for use during field-testing. I did not find evidence for any aspects of online survey administration that could prevent generalizability across formats. Thus, I provided some evidence for the generalizability aspect of Messick's model. Further, during cognitive interviews, I also identified well-functioning attention checks.

Field-Testing

After the instrument was revised based on the data from cognitive interviews and expert reviews, a final version of the instrument ready for field-testing was produced. Field-testing included collection of the field-testing data and statistical analysis of the collected data.

Recruitment. For field-testing, I aimed to collect data from a variety of undergraduate mathematics-based classes over three semesters. To do that, I first recruited instructors who taught undergraduate mathematics-based classes that satisfied the eligibility criterion (see the Target Population section). When contacting the instructors, I asked them whether their class(es) met this criterion. Next, with the permission of the instructors, I recruited students enrolled in their classes.

Instructors were recruited mainly (but not exclusively) through personal connections and recommendations from other faculty members. My recruitment pool consisted of instructors teaching a broad range of mathematics-based classes in a number of disciplines. The courses ranged from low-level to high-level, with the former including foundational courses required for all students regardless of the major. No restrictions for class recruitment applied, i.e., classes taught by the same instructor were eligible to participate in the study. The original pool consisted of 58 classes taught by 31 instructors in the fall semester, 60 classes taught by 37 instructors in the spring semester, and 8 classes taught by 7 instructors in the summer. Thus, the total original pool consisted of 126 classes taught by 47 instructors. Out of this pool, data were collected from 24 classes taught by 16 instructors in the fall semester, 22 classes taught by 17 instructors in the spring semester, and 3 classes taught by 3 instructors in the summer. In total, students were from 49 classes taught by 27 instructors. An average of the number of classes per instructor was 1.81 (*SD* = 0.88), ranging from 1 to 4. Among these 49 classes, 13 classes received an online version of the survey. The other 36 classes received the paper-and-pencil version. The 49 classes, students from which participated in the study, varied in size. The average enrollment was 66.18 (*SD* = 48.40), ranging from 11 to 260. These disciplines included but were not limited to mathematics (e.g., pre-calculus, calculus, differential equations, discrete mathematics, geometry, etc.), electrical engineering (e.g., signals and systems, digital system design, etc.), mechanical engineering (e.g., statics, dynamics, etc.), civil engineering (e.g., remote sensing in civil engineering, etc.), physics, astronomy, and computer science (e.g., data structures, etc.). Student recruitment was conducted differently depending on the version of the survey. For the paper-and-pencil survey administration, I recruited students prior to administering the survey when I came to classes. For online survey administration, I asked instructors to share with their students my email with an invitation to participate in the study and with a survey link. In total, I collected 1499 surveys.

Measures. During field-testing, I measured student engagement as well as additional constructs needed for external validation. Additional constructs include course achievement, effort, persistence, interest, metacognitive strategies, and social efficacy with peers (measured at the course level), as well as intellect, preference for group work, and public speaking anxiety (measured at the person level, i.e., not specific to the course). Below, I describe measures of each construct. A full measure of student engagement (items grouped by hypothesized subscales) is presented in Appendix J. Full scales for other constructs (except achievement, for which no scales were used) are presented in Appendix K; response options for these scales ranged from 1 (“Strongly Disagree”) to 6 (“Strongly Agree”). It should be noted that the measures described in this section are the administered measures; for some measures, some items were excluded prior to developing composite scores (see Chapter Four for more information).

Student engagement. Student engagement – the primary construct of interest in this study – was measured by the proposed instrument. First, students completed the Instructional Time Form where they specified the percentages of in-class instructional time in the lecture section of their class, spent on lecture, whole-class interaction, individual work, group work, and the time when a student decides not to work on a task. Lecture was defined as “the time in class when your instructor explains the material without interacting with students, e.g., when your instructor lectures in a traditional sense, presents the material without asking questions along the way, etc.” Whole-class interaction was defined as “the time in class when your instructor interacts with students addressing the class as a whole, e.g., when your instructor asks questions to the whole class, does interactive lecture, holds a whole-class discussion, or when other students ask questions in front of the whole class, etc.” Individual work was defined as “the time in class when you work on a task without interacting with other students (excluding exams and formal quizzes), e.g., when you do the task on your own, start working on the task by yourself before turning to others, etc.” Finally, group work was defined as “the time in class when you interact with other students about a task, e.g., when you discuss the task with a neighbor, work in a group or pair, check your answers with people sitting nearby, etc.” Notably, students were presented only with definitions and not with the name of the instruction type these definitions describe. Students were also asked to verify that the percentages they specified add up to 100%.

Next, students completed four pages of the Engagement Survey. Each page consisted of items specific to a particular instruction type. Each page included a title to bring students’ attention to the instruction type. Each block of items also had a stem that described the instruction type. A block of Lecture items (16 items) has a title “Instructor’s presentations” and a stem “In class, when your instructor explains the material without interacting with students, how often have you…” A block of Whole-Class Interaction items (19 items) had a title “Instructor’s interactions with the class” and a stem “In class, when your instructor interacts with students addressing the class as a whole, how often have you…” A block of Individual Work items (18 items) had a title “Individual in-class work” and a stem “In class, when you work on a task without interacting with other students (excluding exams and formal quizzes), how often have you…" Finally, a block of Group Work items (18 items) had a title "Group in-class work" and a stem "In class, when you interact with other students about a task, how often have you…" In total, 71 items were field-tested. Response options included "Never," "Almost Never," "Rarely," "Sometimes," "Often," "Almost Always," and "Always."

Student achievement. To indicate student achievement, I used three types of measures that could be collected across different courses: final course grades, students' expected final course grades, and students' perceived learning in the course (i.e., the amount learned). Multiple measures (in contrast to a single measure) were used because each measure potentially provides unique information about achievement. Specifically, final course grades are considered to be objective measures of achievement; however, they may be unreliable (e.g., due to inconsistency in grading between and even within instructors), may weakly reflect the actual amount of learning (e.g., in the case when a student already had a solid knowledge of the course content prior to the beginning of the course), and/or may include irrelevant variance such as participation points (Richmond et al., 1987; Rovai & Barnum, 2003). Additionally, as course grade data were collected from instructors, I anticipated that some data would not be obtained due to instructors not willing to provide achievement data. In contrast to course grades, perceived learning is considered to be a subjective measure; therefore, it is inherently confounded with affect (Richmond et al., 1987). However, perceived learning does not have a source of error due to instructors' grading and may better reflect the actual learning in a given course. Finally, students' expected grades (used by, for example, Babcock, 2010) incorporate both course-grade components and students' beliefs on what grade they should receive. Expected grades may be the closest substitute for students' actual grades if the information about actual grades is missing.

Final course grades were measured by a proportion of attained points out of the maximum number of points possible (i.e., grades in percent), as well as by letters obtained via a transformation of grades in percent according to the instructors' grading scales. As mentioned above, these data were obtained from the instructors. Students' expected course grades were obtained from students via a self-report question (Babcock, 2010). Specifically, in the survey, students were asked: "What grade do you expect to get in this class?" Response options included five letter grades from "A" to "F." Perceived learning was measured using three self-report questions, two of which were adapted from Richmond et al. (1987); these questions were also included in the survey. The first question, adapted from Richmond et al. (1987), asked about a perceived learning gain, i.e., an actual amount of learning: "How much have you learned in this class?" Five response options were as follows: "Nothing," "Very little," "Some," "Quite a bit," and "A lot." The second question, also adapted from Richmond et al. (1987), was designed to take into account the maximum possible learning gain for a particular student, i.e., a potential amount of learning: "How much could you have learned in this class in the ideal circumstances?" Three response options were as follows: "As much as I learned," "Somewhat more than I learned," and "Much more than I learned." In addition to these two questions, I also included a question about students' prior knowledge of the material: "How much did you know about the course content before taking this class?" Five response options were as follows: “Nothing,” “Very little,” “Some,” “Quite a bit,” and “A lot.” The question was designed to complement the question about the potential amount of learning by accounting for prior knowledge.

Effort and persistence. To measure effort and persistence, I adapted the scales developed and validated by Elliot et al. (1999). The effort scale included two items: “I put a lot of effort into this class” and “I worked very hard in this class.” The persistence scale included four items; example items were “When I become confused about something I’m studying for this class, I go back and try to figure it out” and “In this class, I try to learn all of the testable material “inside and out,’’ even if it is boring.” In the study of Elliot et al. (1999), scores on the effort scale showed high internal consistency (Cronbach’s alpha was 0.93), and the scores on persistence scale showed good internal consistency (Cronbach’s alpha was 0.78).

Interest-feeling and interest-value. In this study, I used the construct of maintained situational interest (MSI), i.e., a form of interest where students “begin to forge a meaningful connection with the content of the material and realize its deeper significance” (Linnenbrink-Garcia et al., 2010, p. 648). MSI has two components: feeling-related and value-related. Linnenbrink-Garcia et al. (2010) developed and validated the subscales for these components over the course of three studies. To measure the feeling-related component of MSI, I adapted the final subscale from Study 3 (four items, for example: “What we are learning in [this class] is fascinating to me”). To measure the value-related component of MSI, I pulled and adapted items from each of the studies (six items, for example: “The things we are studying in [this class] are important to me”). In particular, I adapted three out of four items from the final version. I replaced the fourth item “What we are learning in [this class] can be applied to real life” with “I see how I can apply what we are learning in [this class] to real life” from Study 1 of Linnenbrink-Garcia et al. (2010). I made this replacement because with this scale I aimed to measure a student’s own interest; thus, I aimed the item to reflect whether a student believes they themselves can apply the class content to real life rather than whether a student believes the class content can be applied to real life in general, not necessarily by the student. Further, I also adapted two more items from Studies 1 and 2: “I find the content of [this class] personally meaningful” and “What we are learning in [this class] is important for my future goals.” I added these items because they are relevant to undergraduate students’ value of what they are learning in their classes. The items also seemed to behave fine in the studies of Linnenbrink-Garcia et al. (2010), although the authors noted that the second item was only marginally acceptable based on the modification indices. I hypothesize that it did not fit well with other value items because of the sample – middle and high school students. For undergraduate students, the item may be more relevant and, therefore, may exhibit a better fit. Scores on the feeling subscale showed high internal consistency in Study 3 of Linnenbrink-Garcia et al. (2010): Cronbach’s alpha was 0.92. For somewhat different sets of items measuring the value component of MSI, internal consistency in the studies of Linnenbrink-Garcia et al. (2010) was high: Cronbach’s alpha was 0.91, 0.88, and 0.88.

Metacognitive strategies. For the purposes of this study, I adopted the conceptualization of metacognitive strategies of Wolters (2004). According to Wolters

(2004), metacognitive strategies refer to “students’ use of planning, monitoring, and regulatory strategies when completing work for [the] class” (p. 240). To measure metacognitive strategies, I adapted the nine-item scale developed by Wolters (2004). Example items include “Before starting an assignment [for this class], I try to figure out the best way to do it” and “If what I am working on for [this class] is difficult to understand, I change the way I learn the material.” Scale scores showed high internal consistency in the study of Wolters (2004): Cronbach’s alpha was 0.78.

Intellect. To measure intellect, an aspect of the Openness/Intellect domain of personality within the Big Five model, I used the Intellect subscale of the Big Five Aspect Scales, developed and validated by DeYoung et al. (2007). To construct the subscale, DeYoung et al. (2007) selected items from the International Personality Item Pool (IPIP). The subscale includes ten items; example items are “I think quickly” and “I have difficulty understanding abstract ideas.” Scale scores showed good internal consistency in the study of DeYoung et al. (2007): Cronbach’s alpha was 0.79 in the original sample and 0.81 in the retest sample.

Social efficacy with peers. Social efficacy with peers refers to “students’ confidence that they could interact well with classmates” (Patrick et al., 2007b, p. 87). To measure social efficacy with peers, I adapted the four-item scale developed by Patrick et al. (2007b). Example items included “I can explain my point of view to other students in [this] class” and “I can work well with other students in [this] class.” Scale scores showed good internal consistency in the study of Patrick et al. (2007b): Cronbach’s alpha was 0.75.

Preference for group work. To measure preference for group work, I used the seven-item scale of Shaw et al. (2000). Example items were “When I have a choice, I try to work in a group instead of by myself” and “I prefer to work on a team rather than individual tasks.” Scale scores showed good internal consistency in the study of Shaw et al. (2000): Cronbach’s alpha was 0.88.

Public speaking anxiety. To measure public speaking anxiety, I adapted the Personal Report of Public Speaking Anxiety (PRPSA) of McCroskey (1970). Specifically, I selected five items that are most applicable to the context of my study. Further, I also replaced the reference of giving a speech to the reference of speaking in front of the whole class. Adapted example items are “My thoughts become confused and jumbled when I am speaking in front of the whole class” and “My heart beats very fast while I am speaking in front of the whole class.” Scores on the original 34-item scale showed high internal consistency of 0.94 in the study of McCroskey (1970).

Student demographics and class data. Finally, the survey included questions about student background and demographic characteristics. In particular, these questions were about the number of absences (0, 1-4, 5-10, More than 10), type of the course (required, elective, general education, other), current/intended major, GPA, student classification (freshman, sophomore, junior, senior, other), enrollment in the Honors Program (yes, no), status (full-time, part-time, other), domicile (domestic in-state, domestic out-of-state, international), English as a native language (yes, no, other), age, gender (male, female, other), and race/ethnicity (White, Black or African-American, Hispanic/Latinx, Asian, American Indian or Pacific Islander, Other). Additional information about the classes was obtained through PatriotWeb, e.g., course number and class size.

Procedure. For the paper-and-pencil survey administration, students completed the survey during class time at the end of the course. For the online administration, students completed the survey in their spare time at the end of the course or after the course was over. The survey consisted of the consent form, the Instructional Time Form, the Engagement Survey, the surveys of class-specific and general external constructs that were measured via multi-item scales, as well as the Demographic Survey. In the Engagement Survey, the four pages with engagement items (one page per instruction type) were assembled in different orders. In determining the order of engagement pages, I decided to keep a pair of Lecture and Whole-Class Interaction pages together and a pair of Individual and Group Work pages together because of the similar focus of instruction types (an instructor vs. students, respectively). Besides this constraint, the order of pages was varied. For example, a student, who received Form 1, first saw the Lecture page, then the Whole-Class Interaction page, then the Individual Work page, and then the Group Work page. After the Engagement Survey, students completed the surveys with multi-item measures of external constructs. The class-specific survey had two pages, and the general survey had one page. I also varied the order of these surveys. For example, a student, who received Form 1, first saw the survey with class-specific measures and then the survey with general measures. Finally, in every form, the Demographic Form was placed at the end of the survey. The Demographic form included questions about student achievement and perceptions of learning in the course, as well as their background and demographic information. In total, I assembled eight forms. The structure of each form is presented in Table 16. The student version of the survey is presented in Appendix L.

Table 16. Order of survey parts in different forms

Form	#1	#2	#3	#4	#5	#6
Form 1	L	W	I	G	AC	AG
Form 2	L	W	G	I	AG	AC
Form 3	W	L	I	G	AC	AG
Form 4	W	L	G	I	AG	AC
Form 5	I	G	L	W	AG	AC
Form 6	I	G	W	L	AG	AC
Form 7	G	I	L	W	AC	AG
Form 8	G	I	W	L	AC	AG

Note. L, W, I, and G refer to the pages of the Engagement Survey, where L is Lecture, W is Whole-Class Interaction, I is Individual Work, and G is Group Work. AC and AG refer to surveys with additional (external) class-specific and general measures, respectively.

After the course was over and final course grades were posted, I contacted instructors to request student achievement data for students who consented for this information to be requested from their instructors. The requested achievement data included student course grades in percent/points and letter grades. These data were obtained for 34 classes. For one of these classes, only letter grades were obtained. For two classes, grades in points/percent were in raw points (not percent); therefore, they were not used. For three classes, I converted grades in percent to letter grades myself based on the scale provided by the instructor.

Data cleaning. Data from paper-and-pencil surveys were transcribed. Data points were entered as missing if a student skipped an item, if a student selected more than one response option, or if it was not clear which response option was selected. In the Instructional Time Form, if data were entered in units other than percentages (e.g., minutes), I converted these data to percentages. If data were entered as a range (e.g., 10-15%), I entered a middle point of the range (e.g., 12.5%). Finally, if a student specified one joint percentage for more than one instruction type together, I divided the percentage by the number of joint instruction types. Most students, however, completed the Instructional Time Form as expected. Another note about data entry needs to be made about student majors and GPAs. For student majors, I entered majors in the commonly used wording, even if the wording of a particular student was somewhat different. I also entered majors without specific concentrations. For GPAs, if a student wrote a range, I entered a middle point of the range. Additionally, if a student wrote that they did not have a GPA, I treated the response as missing.

Further, prior to the analysis, I checked that the data were entered correctly. Specifically, the data were entered twice by three people (including me). Each item was entered by two different people. After the two data entries were completed, I compared the two data matrices using SAS. Data points that differed between the two entries were checked and corrected using the raw data. After the data were checked for correctness of entry, I cleaned the data. The total number of surveys was 1499. The data cleaning process included several steps. First, I removed all surveys from students who were younger than 18 years old (N = 7). Next, I removed a student who did not specify their name on the consent form. After that, I excluded two students who were not found on the instructors' rosters. This information was brought to my attention by instructors when I requested student achievement data. Further, I examined data on engagement items to see if some responses reflected “0” for some instruction types in the Instructional Time Form. Specifically, if students selected “Never” or responses close to “Never” (e.g., “Almost Never” or “Rarely”) for all items in a particular engagement page, I checked to see if the percentage of time, specified for the instruction type was zero. Some students also left comments to indicate that a particular engagement page was not applicable to them because the instruction type did not take place in their class. Thus, I replaced responses as missing if responses of “Never” or close to “Never” for all items on a particular engagement page corresponded to 0 for a particular instruction type in the Instructional Time Form. I also replaced such responses as missing even if the time for the instruction type in the Instructional Time Form was not zero, as I hypothesized the time in the Instruction Time Form could be specified as non-zero due to an error. Such replacements were made for Lecture pages of four surveys, for a Whole-Class Interaction page of one survey, for Individual Work pages of 21 surveys, and for Group Work pages of 13 surveys. In one survey, data may have been replaced with missing for more than one engagement page.

Next, I identified and removed careless respondents. For the paper-and-pencil surveys, I used the response pattern approach to identify careless respondents. In the response pattern approach, patterns of responses are used to indicate careless responding (Huang et al., 2012). One of the most commonly used methods within the response pattern approach is identifying long strings, i.e., a large set of consecutive responses of the same value. I considered a response pattern to be a long string if the entire survey page or a substantial part of the survey page had responses of the same value. I assumed a participant to be a careless respondent if two or more survey pages included long strings. One exception is the survey of class-specific items where all but one item was positive. Thus, it was plausible that students answered all but one item in the same way. If this one item was included in the long string, I considered it to be a potential error but did not assume a participant to be a careless respondent. Further, I assumed students whose response patterns had a particular shape (e.g., a z-shape) on any pages of the survey to be careless respondents. In total, I removed 27 paper-and-pencil surveys identified as completed by careless respondents.

For online surveys, I used responses to attention checks to identify careless respondents. Specifically, I assumed students to be careless respondents if, responding to the question about self-reported data quality, they reported that their data should not be used in the analyses. I also assumed students to be careless respondents if they failed more than two out of seven other attention checks (four infrequency engagement items and three instructed response items). Infrequency engagement items were marked as failed if a student responded with 3 (“Rarely”) or higher. Instructed response items were marked as failed if a given response was different from the instructed response. In total, I removed five online surveys identified as completed by careless respondents.

After removing careless respondents, I excluded students who identified themselves as graduate or non-degree students (N = 17), as my study focuses on undergraduate students. For students who selected the “Other” option for the student classification question and specified classifications that fall into one of the four undergraduate classifications (freshman, sophomore, junior, or senior), I recoded the responses. For students, who also selected the “Other” option but did not specify their classification, I replaced the responses with missing data but did not remove the surveys. For the question about the type of the course, some students selected more than one response option. If one of the options was “Required for major/minor,” I recoded the responses as “Required for major/minor.” If students selected both “Elective” and “General Education (Mason Core),” I recoded the responses as “General Education (Mason Core).” Among students who selected the “Other” response option, some indicated that the course was a pre-requisite for them. Thus, I coded these responses into a category of “Pre-requisite.” For the rest of the students who selected the “Other” response option, I recoded them into the existing categories based on their comments. For the question about student status (full-time or part-time), I coded responses for the summer courses into a separate, third category. For the question about English being a student’s native language, some students selected the “Other” response option. If these students indicated that they were bilingual, I coded their responses in the “Yes” category. If these students wrote their second language, I coded them in the “No” category. Finally, for the race/ethnicity question, I created a separate “Mixed race/ethnicity” category if a student selected more than one race or ethnicity.

Next, I identified surveys that were completed by the same students in different classes. I did that by comparing students' first and last names across cases. I found that 1135 students completed the survey once, 117 students completed it twice, 17 students completed it three times, four students completed it four times, and one student completed it five times. Then, I selected only one survey per student. The process included several steps. The first step included saving in a separate dataset surveys of students who completed the survey only once. The second step included investigating the number of zeros for the four instruction types in the Instructional Time Form among students who completed the survey more than once. For each student, I retained surveys with the number of zeros equal to the minimum number of zeros for this student across all surveys they completed. For example, if a student put 0% for Group Work in one class but put percentages greater than zero for all four instruction types in another class, then the survey from the second class was selected. If only one survey per student remained, then I saved it in a separate dataset. Two or more surveys that were completed by the same students were retained for the next step. The third step included investigating the number of missing item responses across all engagement pages for all remaining surveys. For each student, I retained surveys with the number of missing item responses on the Engagement Survey equal to the minimum number of missing item responses on the Engagement Survey for this student across all surveys they completed. For example, if a student, who completed two surveys and did not indicate zeros in the Instructional Time Form in both surveys, missed two engagement items in one survey but five engagement items in another survey, then the first survey was selected. If only one survey per student remained, then I saved these surveys in a separate dataset. Two or more surveys that belonged to the same students were retained for the next step. The fourth step involved selecting one survey per student out of the remaining pool of students, who completed more than one survey, based on the timing of survey administration. Specifically, a survey that was administered at the earliest time and date was selected. All surveys collected online in one semester were coded with one order number. There was one student who completed an online survey twice, but only one of these surveys remained after the previous steps. To create the dataset with all students where each student has only one survey, I combined the datasets saved during each of the four steps. In total, 1274 surveys were retained.

Further, some students, who completed paper-and-pencil surveys, filled out engagement pages for instruction types, for which they put zero on the Instructional Time Form. It is unclear what these students referred to when answering engagement items, considering that they indicated that the instruction type did not take place in their class. Thus, I replaced responses on engagement items with missing data for pages, the time for instruction types of which was indicated to be zero. For ten students, data in the Instructional Time Form were missing either because they did not answer the form or because their responses were in a form that I could not transform into percentages. Thus, for these students, I replaced all engagement data with missing data because it was unclear which instruction types occurred in their classes. I refer to the dataset before this replacement as the original dataset and to the dataset after this replacement as the reduced dataset. The reduced dataset was used for further analysis.

Next, I investigated the amount of missing data for each item using the original dataset. Descriptive statistics for missing data for each item block are presented in Table 17. Across all items on the survey (with the exception of the Instructional Time Form, where missing data was replaced with 0%), data were missing, on average, for 3.18% of surveys (*SD* = 1.92%), ranging from 1.18% to 21.90%. Among engagement items only, data were missing, on average, for 3.09% of surveys (*SD* = 1.41%), ranging from 1.18% to 6.12%. For additional (external) measures (both class-specific and general), data were missing, on average, for 2.94% (*SD* = 0.34%), ranging from 2.28% to 3.45%. For the Demographic Form, data were missing, on average, for 4.34% of surveys (*SD* = 4.72%), ranging from 2.51% to 21.90%. The largest amount of missing data was for the question about student GPA, which was not surprising. Some students did not have a GPA due to being in their first semester at GMU. Other students did not remember their GPA or preferred not to disclose it.

Table 17. Descriptive statistics for missing data for each survey block in the original dataset (N = 1274)

Statistic for the % of missing data	L	W	I	G	AC	AG	D
Mean	2.08	1.52	4.99	3.76	2.87	3.03	4.34
SD	0.20	0.19	0.32	0.22	0.37	0.27	4.72
Minimum	1.81	1.18	4.63	3.22	2.28	2.51	2.51
Maximum	2.51	1.88	6.12	4.00	3.45	3.38	21.90

Note. L, W, I, and G refer to the pages of the Engagement Survey, where L is Lecture, W is Whole-Class Interaction, I is Individual Work, and G is Group Work. AC and AG refer to surveys with additional (external) class-specific and general measures, respectively. D refers to the Demographic Form.

As some engagement data were block-missing (e.g., entire engagement pages were missing), I excluded these surveys in order to investigate the amount of data that were missing not due to the lack of applicability of the instruction type to the student's class. Additionally, I also excluded online surveys, as data on these surveys were missing only if engagement pages were not administered due to the lack of applicability of instruction types or if the survey was not finished. Descriptive statistics for missing data for each item block are presented in Table 18. Across all items on the survey (with the exception of the Instructional Time Form, where missing data was replaced with 0%), data were missing, on average, for 0.95% of surveys (*SD* = 1.40%), ranging from 0% to 15.78%. Among engagement items only, data were missing, on average, for 0.33% of surveys (*SD* = 0.20%), ranging from 0% to 1.33%. For additional (external) measures (both class-specific and general), data were missing, on average, for 1.33% (*SD* = 0.27%), ranging from 0.94% to 1.88%. For the Demographic Form, data were missing, on average, for 2.49% of surveys (*SD* = 3.57%), ranging from 1.10% to 15.78%.

Table 18. Descriptive statistics for missing data for each survey block in the data set with no engagement block-missing or online surveys (N = 911)

Statistic for the % of missing data	L	W	I	G	AC	AG	D
Mean	0.29	0.25	0.37	0.41	1.42	1.21	2.49
SD	0.18	0.16	0.27	0.16	0.30	0.15	3.57
Minimum	0.08	0.00	0.16	0.08	0.94	0.94	1.10
Maximum	0.78	0.55	1.33	0.63	1.88	1.65	15.78

Note. L, W, I, and G refer to the pages of the Engagement Survey, where L is Lecture, W is Whole-Class Interaction, I is Individual Work, and G is Group Work. AC and AG refer to surveys with additional (external) class-specific and general measures, respectively. D refers to the Demographic Form.

Next, I checked if data were missing completely at random (MCAR) via Little's MCAR test (Little, 1988). The null hypothesis of this test states that data are MCAR. Thus, a non-significant Chi-Square statistic used to evaluate the test suggests that the data are MCAR. In contrast, a statistically significant Chi Square statistic suggests that data are not MCAR. Using 148 items across the entire survey, I ran Little's MCAR test using the SAS macro for this test from the Applied Missing Data .com website (*Http://Www.Appliedmissingdata.Com/Littles-Mcar-Test.Sas*, n.d.). In particular, I ran the test on two datasets: the original dataset (N = 1274) and the dataset with no engagement block-missing or online surveys (N = 911). For both datasets, the null hypothesis was rejected. For the original dataset, Chi Square (32996) = 37938.55, $p < 0.001$. For the dataset with no engagement block-missing or online surveys, Chi Square (24396) = 28279.54, $p < 0.001$. Thus, the results suggest that the data are not MCAR.

Participants. Demographic and background information of the field-testing sample is presented in Table 19. Students' GPA, on average, was 3.31 (*SD* = 0.46), ranging from 1.5 to 4.00 (N = 995). Students' average age was 20.75 years (*SD* = 3.87), ranging from 18 to 70 (N = 1227). Notably, most students had only a few absences, suggesting that they knew about instruction in the course well and, therefore, providing some support for the validity of responses to the Instructional Time Form. Table 20 shows the number of classes per discipline and course level, as well as the total number of students per discipline and course level. Overall, participants were from 49 classes within nine disciplines. The average number of participants per class was 26.00 (*SD* = 20.33), ranging from 1 to 89.

Table 19. Demographic and background information of field-testing participants

Characteristic	Frequency	%
Number of absences	1242	
- 0	629	50.64
- 1-4	562	45.25
- 5-10	43	3.46
- More than 10	8	0.64
Type of course	1237	
- Required for major/minor	1014	81.97
- Elective	48	3.88
- General Education (Mason Core)	166	13.42
- Pre-Requisite	8	0.73
Major	1218	
- Bioengineering	25	2.05
- Biology	140	11.49
- Chemistry	21	1.72
- Civil Engineering	65	5.34
- Computer Engineering	94	7.72
- Computer Science	161	13.22
- Cyber Security Engineering	68	5.58
- Electrical Engineering	115	9.44
- Forensic Science	22	1.81
- Information Systems and Operations Management	23	1.89
- Information Technology	32	2.63
- Mathematics	24	1.97
- Mechanical Engineering	96	7.88
- Neuroscience	29	2.38
- Other STEM[1]	111	9.11
- Other non-STEM[2]	164	13.46
- Two or more majors where at least one major is STEM	28	2.30
Student classification	1129	
- Freshman	351	28.56
- Sophomore	290	23.60
- Junior	361	29.37
- Senior	227	18.47
Honors Program	1234	
- Yes	181	14.67
- No	1053	85.33
Status[3]	1195	
- Full-time	1119	93.64
- Part-time	76	6.36
Domicile	1237	
- Domestic, in-state	1059	85.61
- Domestic, out-of-state	128	10.35

Characteristic	Frequency	%
- International	50	4.04
Native language	1237	
- Yes	968	78.25
- No	269	21.75
Gender[4]	1229	
- Male	760	61.84
- Female	463	37.67
- Other	6	0.49
Race/ethnicity[4]	1210	
- White	476	39.34
- Black or African-American	124	10.25
- Hispanic / Latinx	103	8.51
- Asian	349	28.84
- American Indian or Pacific Islander	8	0.66
- Other	47	3.88
- Mixed race / ethnicity	103	8.51

Note. A major was listed if at least 20 students indicated having this major. [1]Other STEM majors included Accounting, Computational and Data Science, Computer Game Design, Earth Science, Medical Laboratory Science, Nursing, Physics, Statistics, Systems Engineering, etc. [2]Other non-STEM majors included Anthropology, Business Management, Communication, Community Health, Conflict Analysis and Resolution, Criminology, Dance, Economics, Education, English, Finance, Global Affairs, Government and International Politics, History, Kinesiology, Marketing, Philosophy, Psychology, Rehabilitation Science, Theatre, etc. [3]Students from summer classes were excluded. [4]Students who selected "Prefer not to answer" were excluded.

Table 20. Description of classes

Discipline	Number of classes (N = 49)		Total number of students (N = 1274)	
	Frequency	%	Frequency	%
Discipline				
Astronomy	5	10.20	157	12.32
Civil Engineering	3	6.12	41	3.22
Chemistry	2	4.08	10	0.78
Computer Science	6	12.24	153	12.01
Cyber Security Engineering	1	2.04	4	0.31
Electrical and Computer Engineering	6	12.24	187	14.68
Mathematics	19	38.78	431	33.83
Mechanical Engineering	3	6.12	50	3.92
Physics	4	8.16	241	18.92
Course level				
100	22	44.90	652	51.18

Discipline	Number of classes (N = 49)		Total number of students (N = 1274)	
	Frequency	%	Frequency	%
200	16	32.65	461	36.19
300	6	12.24	126	9.89
400	5	10.20	35	2.75

Data analysis plan. Via the analysis of field-testing data, I aimed to evaluate assumptions for the following inferences, as specified in the interpretation/use and validity arguments: the generalization inference, the theory-based inference for item scores, the scoring inference, and the theory-based inference for composite scores. For the generalization inference, I evaluated the assumption within the generalizability aspect of Messick's model. For the theory-based inference of item scores, I further evaluated the assumption within the scale functioning component of the substantive aspect; I also evaluated the assumption within the structural aspect of Messick's model. For the scoring inference, I evaluated the assumption of the plausibility of the scoring rule for subscales. Finally, for the theory-based inference of composite scores, I evaluated the assumption within the external aspect of Messick's model.

Generalizability across settings, forms, formats, and groups. Frequencies of students across settings, groups, forms, and formats were examined to evaluate the assumption of scores generalizability within the generalizability aspect of Messick's model. In particular, I investigated frequencies of surveys collected in the following settings: course disciplines, course levels, and course types (required, elective, general education, or pre-requisite). I also examined frequencies of student groups based on student classification (freshman, sophomore, junior, and senior), status (full-time or part-time), enrollment in the Honors Program, major, domicile (domestic in-state, domestic out-of-state, or international), native language (English vs. non-English), gender, and race/ethnicity. For evidence to provide initial support for the generalizability aspect, each setting or group category should have a fair representation in the sample. Finally, I explored frequencies of forms and formats (paper-and-pencil vs. online), as well as versions of the online survey (PC vs. mobile). Evidence of an approximately even number of forms, formats, and versions would provide initial support for generalizability across forms, formats, and versions.

Item characteristics. An examination of item characteristics was conducted to evaluate the assumption within the scale functioning component of the substantive aspect of Messick's model. The assumption states that item response characteristics need to be as expected. First, I expected engagement items to be of good quality. Items of good quality are those that are approximately normally distributed across the full range of the scale (i.e., no evidence of restriction of range). Thus, I screened all items for normality by examining descriptive statistics (means, standard deviations, skewness, and kurtosis). I recoded the three negative emotions (frustrated, anxious, and bored) prior to the analysis.

Second, I expected patterns of relationships between the items within subscales to be positive and moderate. These relationships also need to be statistically significant within subscales if they were to indicate a common factor. Further, correlations of items designed to indicate different engagement dimensions within an instruction type should be lower than correlations within subscales for this instruction type. Finally, correlations of items designed to indicate different instruction types within an engagement dimension should be lower than correlations within subscales for this engagement dimension. Thus, I analyzed a correlation matrix to examine relationships between the items. Correlations were computed in SAS, version 9.4. It is important to note that no adjustment for clustering was made; instead, I used a more conservative alpha level of 0.01.

Third, the items in the Instructional Time Form also need to be of good quality. As one concern may be related to the ability of students, who completed paper-and-pencil surveys, to add to 100%, I investigated the sum of the responses to the Instructional Time Form. Second, I explored student agreement in the percentages they specified for different instruction types via multilevel reliability. I expected a high agreement for Lecture and Whole-Class Interaction but not a high agreement for Individual Work and Group Work, as Individual Work and Group Work are specific not only to a particular class but also to a particular student.

Internal structure. ESEM was used to evaluate the assumption within the structural aspect of Messick's model. This assumption states that an internal structure of the instrument needs to be determined. Thus, I used ESEM to identify the internal structure. ESEM combines Exploratory Factor Analysis (EFA; a method commonly used to identify factors that underlie the internal structure of the instrument) and Structural Equation Modeling (SEM). For the purposes of this study, the main advantage of ESEM over EFA is the ability of ESEM to accommodate correlated errors. Due to the nature of my instrument, some items have similar wording, which may lead to a systematic error in measurement. The similarity in wording occurred for behaviors, cognitive processes, and emotions that are applicable to multiple instruction types. For example, critical thinking applies to all instruction types (e.g., critical thinking about what the instructor says during lecture and critical thinking about the task solution during individual work). To account for this systematic error, I correlated errors of all item pairs with similar wording.

I ran ESEM analyses in MPLUS (Version 8), using TYPE = COMPLEX to account for data clustering. Although classes were also clustered within instructors, the number of classes per instructor was small (1.81 with *SD* = 0.88, ranging from 1 to 4). Thus, only clustering of students within classes was taken into account. For estimation, I used the maximum likelihood estimator with robust standard errors (MLR). MLR treats data as continuous and corrects standard errors for non-normality. For rotation, I used oblique Geomin. An oblique method was chosen over orthogonal because the former allows for factor correlations. For my instrument, I hypothesized that factors are correlated. Geomin was chosen because it minimizes variable complexity in an attempt to identify a simple structure (Sass & Schmitt, 2010). In particular, Geomin provides a solution with minimal cross-loadings. However, its downside is the potential to produce overestimated factor correlations.

ESEM analyses were run on four sets of items, each with a different combination of activating emotions. As a reminder, two positive activating emotions (enjoyment and excitement) and two negative activating emotions (frustration and anxiety) were field-tested. However, the instrument is designed to have one positive activating emotion and one negative activating emotion. Thus, I tested four sets in order to select one. Each set had 63 items out of field-tested 71 items. The full 71-item set was not tested because (1) it does not represent the intended structure and (2) more activating than deactivating emotional items may produce a distorted picture of the internal structure. Further, as this work is exploratory, I investigated models with a different number of factors. In particular, I tested 10 ESEM models, from a 3-factor model (as the minimum number of factors was hypothesized to be three given three designed dimensions) to a 12-factor model (as the instrument was hypothesized to have 12 subscales).

To evaluate and compare the models, I used two approaches. First, I evaluated and compared model fit. To evaluate model fit, I examined the following tests and indices: the Satorra-Bentler Scaled Chi-Square test of exact fit, the Root Mean Square Error of Approximation (RMSEA) with a 90% confidence interval, the Bentler Comparative Fit Index (CFI), the Tucker Lewis Index (TLI), the Standardized Root Mean Square Residual (SRMR), the Akaike information criterion (AIC), the Bayesian information criterion (BIC), and the sample-size adjusted BIC. I also investigated local misfit via Lagrange Multiplier Statistics. RMSEA less than 0.06, SRMR less than 0.08, CFI above 0.95 (Hu & Bentler, 1999), and the lack of significant local misfit suggest good model fit. In general, the closer RMSEA and SRMR are to zero, the closer CFI and TLI are to one, the closer Chi Square is to a critical value (when Chi Square is larger than the critical value), and the smaller AIC, BIC, and sample size adjusted BIC are, the better model fit is.

Second, I investigated which model is the most interpretable from the theoretical point of view. In a good model, items loading on a particular factor reflect this factor, items do not cross-load when no cross-loading is hypothesized, and no factors have loadings that are not substantial. Thus, I explored item quality. In particular, I examined factor loadings. Factor loadings on factors that they reflect need to be statistically significant. In EFA, pattern coefficients should be greater than 0.32 (Comrey & Lee, 1992; Tabachnick & Fidell, 2007). I used this rule of thumb to guide my evaluation of ESEM loadings. Further, I investigated the amount of variance explained (R^2) in each item by the factors. In EFA, communalities, i.e., the amount of variance in an item that is explained by the factors, should be greater than 0.50 (Meyers et al., 2012). I used this rule of thumb to guide my evaluation of item quality of ESEM, as well. Items with low loadings and/or explained variance were considered for removal. Finally, I evaluated factor correlations. I hypothesized factor correlations to be moderate, as I expected engagement subscales to be related. Overall, I used the interpretability approach to determine the number of factors, as model fit in ESEM typically improves with more factors. However, models with a large number of factors might not be interpretable. To compare different sets of items, I used both the interpretability approach and the model fit approach.

Creating composite engagement scores. Subscale (or factor) composite scores intended to indicate the levels of particular engagement dimensions in particular instruction types. To create composite engagement scores for each subscale, I first evaluated the assumption of the plausibility of the scoring rule. As a reminder, the designed scoring rule for subscales was computing an average of items in each subscale. To evaluate the plausibility of this rule, the analytic plan included testing whether the items within each factor in the final ESEM model are parallel (McNeish & Wolf, 2020).

In particular, I planned to test whether loadings and error variances of items that indicate one factor are the same. Next, I intended to create engagement dimension composites, instruction type composites, and global engagement composites via the scoring rules specified in the scoring inference. These scoring rules were assumed to be reasonable and did not require empirical evidence. Specifically, I intended to compute engagement dimension composite scores via summing subscale composite scores across instruction types, weighted by the amount of instruction, and dividing the sums by the total amount of time spent on the four types of instruction. Next, I intended to compute instruction type composite scores via averaging subscale composite scores across engagement dimensions. Finally, I intended to compute global engagement composite scores via averaging engagement dimension composite scores.

Additionally, I examined descriptive statistics and correlations for engagement composite scores to further evaluate the assumption within the scale functioning component of the substantive aspect of Messick’s model. This assumption states that the characteristics of composite scores need to be as expected. In particular, I screened composites for normality by examining descriptive statistics (means, standard deviations, skewness, and kurtosis). Correlations were computed in SAS, version 9.4. It is important to note that no adjustment for clustering was made; instead, I used a more conservative alpha level of 0.01. I expected correlations between engagement composites to be positive and moderate. Such correlations would indicate that engagement composites are proxies of related but distinct constructs.

Correlational and regression analyses. Correlational and regression analyses were used to evaluate the assumption within the external aspect of Messick's model. The assumption states that expected relationships with other relevant constructs need to be demonstrated. Thus, first, a series of correlational analyses were performed. Specifically, all engagement composite scores were correlated with composite scores of external constructs, i.e., effort, persistence, feelings-related and value-related components of interest, metacognitive strategies, intellect, social efficacy with peers, preference for group work, and public speaking anxiety. Second, multiple and simple regression analyses were performed. Specifically, I conducted four sets of simple or multiple regression analyses where I regressed achievement on the four types of composite engagement scores, i.e., factor composite scores, engagement dimension composite scores, instruction type composite scores, and global engagement composite scores. I describe the analysis and expected relationships in more detail below. In each regression analysis, I used four achievement scores: actual grade in percent, actual letter grade, expected letter grade, and perceived amount of learning.

Preparing additional constructs for the analysis. To prepare scores on additional (external) measures for the analyses, I first recoded negatively worded items. Then, I tested measurement models to evaluate internal structures of additional constructs, which were measured via multi-item scales, and to improve the internal structures if needed by removing problematic items. Measurement models were run via CFA. I ran separate CFA models for each construct as opposed to one CFA model with all constructs included because I did not intend to simultaneously use multiple additional constructs in the analysis. However, two CFA models included two constructs. One CFA model included effort and persistence because the effort scale had only two items and, therefore, could not be modeled separately. Besides, effort and persistence are conceptually similar, warranting an investigation of whether they can be distinguished empirically. Another CFA model included the feeling and value components of interest. The two constructs were modeled together because they are two dimensions of a larger construct. Model fit was evaluated in the same way as for ESEM engagement models. After I established internal structures of additional measures, I checked the internal consistency of scores for each construct. Specifically, I computed Cronbach's alpha. Cronbach's alpha of 0.70-0.90 indicates good internal consistency (Nunnally, 1978; Streiner, 2003). Next, I created composite scores for each construct via averaging the corresponding items. For achievement, no scales were used. I used student actual course grades in percent, actual letter grades, and expected letter grades as separate dependent variables. For perceived learning, I used actual amount of leaning as a dependent variable, while including potential amount of learning and prior knowledge as control variables. Actual letter grades were recoded into an 11-point scale where "F" was 1 and "A" was 11. Expected letter grades were measured on a 5-point scale and were recoded so that 1 indicated "F" and 5 indicated "A." Finally, scores on all external variables were screened for normality via an examination of descriptive statistics (means, standard deviations, skewness, and kurtosis). They were also examined for restriction of range. Further, actual grades in percent were investigated for outliers, considering the large range of this variable.

Correlational analyses. Before conducting correlational analyses, I examined all relationships for linearity. Specifically, linearity was investigated through bivariate scatterplots. For correlational analyses, I computed Pearson correlation coefficients to measure the strength and direction of the relationships between engagement composite scores with composite scores of effort, persistence, feelings-related and value-related components of interest, metacognitive strategies, intellect, social efficacy with peers, and preference for group work. The analysis was conducted in SAS, version 9.4. It is important to note that no adjustment for clustering was made; instead, I used a more conservative alpha level of 0.01. I expected positive moderate correlations of engagement with effort, persistence, interest components, and metacognitive strategies. Prior research used effort and persistence to indicate behavioral engagement, feelings and values to indicate emotional engagement, and metacognitive strategies to indicate cognitive engagement. Thus, it was particularly important to show correlations moderate in magnitude for effort and persistence with behavioral engagement, feelings and values with emotional engagement, and metacognitive strategies with cognitive engagement. Moderate relationships suggest that engagement dimensions, though related to these constructs, are yet distinct from them, thus providing evidence of discriminant validity for the instrument. Correlations of engagement composite scores with intellect are expected to be positive and low-to-moderate. Such correlations suggest that the instrument measures engagement rather than intellect. Thus, relationships with intellect are also expected to provide evidence of discriminant validity. Social efficacy with peers is expected to correlate positively with group work composite scores but is not expected to correlate with engagement in other instruction types. Preference for group work is expected to be correlated positively with group work composite scores, negatively with individual work composite scores, and non-significantly with lecture and whole-class interaction composite scores. Finally, public speaking anxiety is expected to be correlated negatively with whole-class interaction composite scores, and non-significantly correlated with engagement in other types of instruction. Results of the correlational analyses between engagement and social efficacy with peers, group work, and public speaking anxiety are expected to provide validity evidence specifically for instruction type composite scores.

Regression analyses. Before conducting regression analyses, I investigated relationships between achievement and engagement composites for linearity. Specifically, linearity was examined through bivariate scatterplots. Further, the assumption of homoscedasticity was evaluated by examining scatterplots of residuals plotted against predicted values and each predictor. Next, I examined zero-order correlations with engagement composite scores. Correlations were computed in SAS, version 9.4. It is important to note that no adjustment for clustering was made; instead, I used a more conservative alpha level of 0.01. Finally, for each type of engagement composite scores, I conducted four regression analyses that differed in the measurement of achievement. Four measures of course achievement were used: student actual grades in percent, student actual letter grades, student expected grades, and actual perceived learning. Regressions with actual perceived learning as an outcome also controlled for perceived potential learning and perceived prior knowledge. Additionally, all regressions with subscale composite scores or instruction types composite scores controlled for the amount of time spent on each instruction type. As not all students specified percentages to add to 100%, I recoded the amounts of time for each instruction type as missing if the total was not 100%. The time when a student decided not to work on a task was used as a reference instruction type. Regressions with dimension composite scores or global engagement composite scores controlled only for the time when a student decided not to work on a task because other times were incorporated in engagement composites through weighting. Regression analyses were conducted in MPLUS (Version 8). TYPE = COMPLEX was used to account for student clustering within classes. The MLR estimator was employed to correct standard errors for non-normality. In each analysis, I examined regression coefficients. I expected the relationships to be either positive or statistically non-significant, based on the results from prior research.

Internal consistency. I planned to examine the assumption of internal consistency within the generalizability aspect of Messick's model via Cronbach's alpha. Specifically, I planned to investigate internal consistency for each engagement factor. Recommendations for evaluating Cronbach's alpha vary between researchers, purposes of research, and stages of research. For example, Nunnally (1967) recommended the following reliability standards: 0.50 – 0.60 for the early stages of research, 0.80 for basic research, and 0.90 as the minimum for clinical purposes, with the desired value of 0.95. Later, Nunnally (1978) increased the standard for early stages to 0.70. Other researchers disagree with the "the higher, the better" rule because very high values indicate redundancy more than homogeneity (e.g., Boyle, 1991; Streiner, 2003). In this case, reliability is achieved at the expense of validity – the phenomenon known as the "attenuation paradox" (e.g., Boyle, 1991; Clark & Watson, 1996; Loevinger, 1954). While highly redundant items result in high correlations leading to high internal consistency, they also do not provide more information about the construct than a single item, leading to very narrowly defined constructs and potential construct underrepresentation (Clark & Watson, 1996). For this reason, Streiner (2003) recommended a maximum Cronbach's alpha of 0.90. Boyle (1991) claims that for measuring broadly defined constructs in the non-ability areas of motivation, personality, and mood states, having moderate to low item homogeneity is actually preferred.

Chapter Four

In this chapter, I present results of the field-testing phase of the study. First, I examined frequencies of students across settings, groups, forms, and formats to evaluate the assumption of scores generalizability within the generalizability aspect of Messick's model. Second, I investigated item characteristics to evaluate the assumption within the scale functioning component of the substantive aspect of Messick's model. Third, I explored the internal structure of the instrument to evaluate the assumption within the structural aspect of Messick's model. Fourth, I evaluated the assumption of the plausibility of the subscale scoring rule in order to develop subscale composite scores. Fifth, after composite scores of all types (subscale, dimension, instruction type, and global composite scores) were produced, I conducted correlational and regression analyses to evaluate the assumption within the external aspect of Messick's model. Finally, I investigated the assumption of internal consistency within the generalizability aspect of Messick's model.

Generalizability Across Settings, Forms, Formats, and Groups

Frequencies across settings and groups are presented in Table 19 and Table 20 in Chapter Three. Results show that for the majority of students, the course was required for their major or minor. Most classes were also in the lower division. Further, while many classes within disciplines were present in the sample, this representation was not equal across disciplines. The largest number of classes was in Mathematics. In terms of student classification, all groups (freshmen, sophomores, juniors, and seniors) were presented in the sample, although the number of seniors was the lowest. Students also varied in majors, with the most represented majors being Biology and Computer Science. The number of students with non-STEM majors was less than one-seventh of the sample. Next, most students were full-time, in-state, native English speakers, and not in the Honors Program. The sample also had more male participants than female. Further, all races/ethnicities were represented in the sample, although the number of American Indian and Pacific Islander students was very low. Most students were White or Asian. In terms of forms, each form was completed by approximately the same number of students (see Table 21). In terms of formats, the majority of surveys were completed in the paper-and-pencil format (95.13%), while the number of online surveys completed on a mobile device or on a PC was very small (1.49% and 3.38%, respectively).

Table 21. Frequencies of forms

Form	Frequency	%
Form 1	156	12.27
Form 2	172	13.53
Form 3	161	12.67
Form 4	162	12.75
Form 5	156	12.27
Form 6	155	12.20
Form 7	151	11.88
Form 8	158	12.43

Note. N = 1271.

Overall, with the analysis of frequencies across settings, groups, forms, and formats, I examined the score generalizability assumption within the generalizability aspect of Messick's model. This assumption states that scores are generalizable and reliable regardless of the setting, group membership, form, or format. An initial step in evaluating this assumption was to investigate whether the field-testing sample included a variety of classes from different settings, all forms, all formats, and students from a variety of backgrounds. In general, I found support for this assumption. However, within some settings and groups, some categories had low numbers of students. Further, while all forms were approximately equally distributed in the sample, the format was not. There was a very small number of students who completed the survey online.

Item Characteristics

Item descriptive statistics and frequencies are presented in Appendix M. Items differed in the means, with some items having lower means than other items. Items with lower means (i.e., more "difficult" items) were less commonly endorsed by students, while items with higher means (i.e., "easier" items) were more commonly endorsed by students. In other words, more "difficult" items may reflect more difficult ways to engage in class, while "easier" items may reflect easier ways to engage in class. As a reminder, response options ranged from 1 to 7. The items with the lowest means were items about active behavioral engagement in Whole-Class Interaction: "Volunteered to answer your instructor's questions in front of the whole class" (M = 3.04, *SD* = 1.65), "Shared your ideas or thoughts with the whole class" (M = 2.85, *SD* = 1.52), and "Asked questions to your instructor in front of the whole class" (M = 2.90, *SD* = 1.64). Other behavioral and cognitive items with lower means were items about drawing own pictures and writing own remarks or comments in Lecture (M = 3.77, *SD* = 1.58 and M = 3.76, *SD* = 1.70, respectively) and Whole-Class Interaction (M = 3.77, *SD* = 1.70 and M = 3.21, *SD* = 1.62, respectively); taking notes in Whole-Class Interaction (M = 4.29, *SD* = 1.73) and Group Work (M = 3.73, *SD* = 1.67), writing down in detail your task solution or thinking about the task (M = 4.51, *SD* = 1.46), thinking about different ways of solving or answering the task even you already have an answer (M = 4.35, *SD* = 1.61), writing down more than one way of solving or of thinking about the task even if you already have an answer (M = 3.63, *SD* = 1.56) in Individual Work, and putting something into own words in all instruction types (M = 4.47, *SD* = 1.37 for Lecture, M = 4.31, *SD* = 1.31 for Whole-Class Interaction, M = 4.52, *SD* = 1.53 for Individual Work, and M = 4.47, *SD* = 1.38). In contrast, some behavioral and cognitive items with higher means were items about listening (M = 5.88, *SD* = 1.08) and reading (M = 5.85, *SD* = 1.23) in Lecture, listening (M = 5.62, *SD* = 1.13) and answering in your head or thinking about questions your instructor asks the class (M = 5.36, *SD* = 1.18) in Whole-Class Interaction, recalling from memory the content needed to solve/answer the task (M = 5.53, *SD* = 1.11) and keeping the task instructions/question in mind while solving or answering the task (M = 5.51, *SD* = 1.15) in Individual Work, as well as checking with other students to see if your answers, solutions, or approaches match theirs (M = 5.56, *SD* = 1.29) and giving full attention (M = 5.40, *SD* = 1.11) in Group Work. Among emotional engagement items, items about excitement had the lowest means in all types of instructions (M = 3.69, *SD* = 1.49 in Lecture, M = 3.83, *SD* = 1.47 in Whole-Class Interaction, M = 3.46, *SD* = 1.42 in

Individual Work, and M = 4.29, *SD* = 1.50 in Group Work). In contrast, items about feeling calm was among items with the highest means in all types of instruction (M = 5.49, *SD* = 1.37 in Lecture, M = 5.62, *SD* = 1.28 in Whole-Class Interaction, M = 4.85, *SD* = 1.42 in Individual Work, and M = 5.54, *SD* = 1.31 in Group Work). Thus, for example, in Lecture, the item about drawing own pictures or writing own remarks or comments were more “difficult” (i.e., less commonly endorsed) than the items about reading or listening to the instructor. As another example, in Individual Work, thinking about different ways of solving or answering the task even you already have an answer and writing down more than one way of solving or of thinking about the task even if you already have an answer were more “difficult” (i.e., less commonly endorsed) than recalling from memory the content needed to solve/answer the task or keeping the task instructions/question in mind while solving or answering the task.

In terms of skewness, most items (particularly those with higher means) were negatively skewed, with the item about reading in Lecture being the most skewed (skewness = -1.22). Items with particularly low means were not skewed (e.g., the item about writing remarks or comments in Lecture, skewness = 0) or positively skewed (e.g., asking questions to your instructor in front of the whole class, skewness = 0.46). Finally, items also differed in kurtosis (i.e., peakedness). Some items had an intermediate peak (i.e., were mesokurtic), e.g., an item about critical thinking in Lecture (kurtosis = 0.03). Among the items with a sharper peak (i.e., leptokurtic items), the most peaked was the item about reading in Lecture (kurtosis = 1.80). Among the flat (i.e., platykurtic) items, the flattest was the recoded item about feeling anxious in Individual Work (kurtosis = -

0.79). In sum, many of my items were not normally distributed. However, particularly problematic was only one item about reading in Lecture, as it had a very high mean (M = 5.85), was very skewed (skewness = -1.22), and had a sharp peak (kurtosis = 1.80). Other non-normal items had adequate characteristics.

Item correlations within each instruction type are presented in Appendix N. Descriptive statistics of item correlations by a set of item pairs are presented in Table 22. An investigation of correlations within each hypothesized subscale showed that most correlations were positive, with the exception of two statistically non-significant correlations between two emotions in Whole-Interaction (feeling anxious and feeling excited) and two behaviors in Individual Work (writing down more than one way of solving or of thinking about the task even if you already have an answer and looking at your notes or other resources). As seen from Table 22, average item correlations within hypothesized subscales were weak-to-moderate. Particularly low average correlations were for behavioral subscales in Lecture, Whole-Class Interaction, and Individual Work. These results may be explained by the range of items in terms of item difficulty in these subscales. Thus, while average item correlations within hypothesized subscales fell somewhat short from the ideal correlations (moderate positive), they were nevertheless adequate. In particular, they showed that an increase in some behaviors, cognitive processes, and emotions within each instruction type likely means an increase in others, although their levels may differ. The low-to-moderate magnitude of correlations suggests that behaviors, cognitive processes, and emotions within subscales were sufficiently distinct and not redundant. However, some exceptions existed. For example, in Whole-

Class Interaction, the items about volunteering to answer your instructor's questions in front of the whole class and sharing your ideas or thoughts with the whole class may be very similar (correlation = 0.746).

Although items within subscales may be rather distinct, they were nevertheless more similar within subscales then outside of the subscales. In particular, average correlations of item pairs that indicate different engagement dimensions within instruction types were lower than average correlations of items within subscales for each instruction type (see Table 22). For example, an average correlation of item pairs that indicated different engagement dimensions within Lecture was 0.218, while average correlations within Lecture subscales were 0.313 (behavioral), 0.422 (cognitive) and 0.430 (emotional). Further, average correlations of item pairs that indicate engagement in different instruction types within engagement dimensions were also lower than correlations of items within subscales for each instruction type (see Table 22). For example, an average correlation of item pairs that indicate engagement in different instruction types within behavioral engagement was 0.181, while average correlations within behavioral engagement subscales were 0.313 (Lecture), 0.299 (Whole-Class Interaction), 0.207 (Individual Work), and 0.480 (Group Work). However, some high correlations for item pairs outside of the subscales were also observed. The highest correlation was found for two Group Work items: the item about comparing your and other students' ways of thinking about the task (a cognitive item) and sharing your thinking about the task with other students (a behavioral item). The correlation between these items was 0.703. In sum, the exploration of item correlations showed that item correlation patterns were as expected, with most items correlating positively within hypothesized subscales and correlating more strongly within subscales than outside of them. However, some correlations within subscales were lower than ideally desired for items designed to measure a common factor.

Table 22. Descriptive statistics of item correlations by sets of item pairs

Item set	Mean	*SD*
Correlations of items within a subscale		
Behavioral Engagement in Lecture (LB)	0.313	0.110
Cognitive Engagement in Lecture (LC)	0.422	0.075
Emotional Engagement in Lecture (LE)	0.430	0.150
Behavioral Engagement in Whole-Class Interaction (WB)	0.299	0.195
Cognitive Engagement in Whole-Class Interaction (WC)	0.393	0.127
Emotional Engagement in Whole-Class Interaction (WE)	0.393	0.169
Behavioral Engagement in Individual Work (IB)	0.207	0.096
Cognitive Engagement in Individual Work (IC)	0.364	0.112
Emotional Engagement in Individual Work (IE)	0.408	0.150
Behavioral Engagement in Group Work (GB)	0.480	0.115
Cognitive Engagement in Group Work (GC)	0.480	0.084
Emotional Engagement in Group Work (GE)	0.411	0.127
Correlations of item pairs that indicate different engagement dimensions within instruction types		
Lecture (L)	0.218	0.155
Whole-Class Interaction (W)	0.187	0.151
Individual Work (I)	0.158	0.146
Group Work (G)	0.328	0.150
Correlations of item pairs that indicate engagement in different instruction types within engagement dimensions		
Behavioral (B)	0.181	0.099
Cognitive (C)	0.273	0.084
Emotional (E)	0.235	0.156

For the Instructional Time Form, among students who completed a paper-and-pencil version of the survey (N = 1212), most students (N = 1124 or 92.74%) correctly specified times spent on the instruction types to add to 100%. Further, I also computed multilevel reliability of the specified percentages, i.e., the consistency with which students specified the percentages within their classes. For this analysis, I excluded ten students, for whom data on the Instructional Time Form were missing. I also excluded seven classes that had less than five students. To compute multilevel reliability, I used formulas from| Raudenbush and Bryk (2002). Multilevel reliability for each class was computed as:

$$\lambda_j = \frac{\tau_{00}}{\tau_{00} + \frac{\sigma^2}{n_j}}$$

where τ_{00} is between-class variance, σ^2 is within-class variance, and n_j is class size. After multilevel reliability coefficients were computed for each class, I averaged them across classes to produce average multilevel reliability. I computed average multilevel reliability for each of the four instruction types, as well as for the teacher-centered instruction types together (Lecture and Whole-Class Interaction) and for student-centered instruction types together (Individual Work and Group Work, also including the time not working on a task). The results are presented in Table 23. Results for Lecture suggest that student-reported percentage of time spent on Lecture is a quite reliable indicator of the actual percentage of time spent on Lecture. For Whole-Class Interaction, the student-reported percentage of time is a less reliable indicator of the actual percentage. A further investigation of multilevel reliability coefficients for each class revealed that classes with lower multilevel reliability for Whole-Class Interaction also had lower multilevel reliability for Lecture and Individual Work, suggesting that students may differentiate between these instruction types differently. This result may also be explained by a relatively low number of students in these classes (5-20). For Lecture and Whole-Class Interaction together, the percentage appears to be reliable. The least reliable percentage was for Individual Work, which is not surprising because this instruction type was student-specific rather than class-specific. On that note, it is quite surprising to see that the percentage of time spent on Group Work is quite reliable. This finding may be explained by students' high compliance with instructors' directions or encouragement to work in groups. The overall student-centered time (Individual Work and Group Work, also including the time not working on a task) is also reliable. In sum, I found that students tended to correctly add up percentages of class time to 100%. I also found that they tended to agree on the percentages for teacher-centered and student-centered parts of their classes, although there was less agreement for Whole-Class Interaction and Individual Work.

Table 23. Multilevel reliability for responses to the Instructional Time Form

Statistic	L	W	LW	I	G	IGN
Mean	0.917	0.753	0.934	0.597	0.951	0.924
SD	0.060	0.132	0.050	0.159	0.038	0.056
Min	0.738	0.414	0.784	0.241	0.834	0.756
Max	0.980	0.926	0.985	0.850	0.989	0.982

Note. L = Lecture, W = Whole-Class Interaction, LW = Lecture and Whole-Class Interaction, I = Individual Work, G = Group Work, IGN = Individual Work, Group Work, and Not working on a task. Number of classes is 42.

Overall, with the analysis of item characteristics, I examined the assumption within the scale functioning component of the substantive aspect of Messick's model. The assumption states that item response characteristics need to be as expected. In particular, the analysis of engagement items' descriptive statistics showed that some engagement items were not normally distributed, although most of them did not deviate from the normal distribution substantially. I also found that items varied in their means, with some items being more difficult for students to endorse (i.e., items with lower means). Further, the analysis of engagement items' correlations showed that most items within their hypothesized subscales were positively correlated. Engagement items were also correlated more strongly within the hypothesized subscales than outside of them. However, some within-subscale correlations were not as high as desired for items designed to indicate a common factor. Finally, the analysis of responses to the Instructional Time Form showed that students tended to add up to 100% when specifying percentages of time spent on different types of instruction. Additionally, students in the same classes tended to agree on the percentages for teacher-centered and student-centered parts of their classes, although there was less agreement for Whole-Class Interaction and Individual Work. Thus, I provided some evidence for the assumption within the scale functioning component of the substantive aspect of Messick's model.

Internal Structure

For each set of items, the most interpretable solution from the theoretical viewpoint was a 7-factor structure. A note needs to be made about using TYPE = COMPLEX in my ESEM analyses. In these analyses, I have more free parameters than clusters. This characteristic of my ESEM models presents a problem, the effect of which on results has not been well studied (L. Muthen, personal communication, April 17, 2020). I approached the problem in the following way: standard errors and, hence, p values should be interpreted with caution. However, TYPE = COMPLEX does not affect parameter estimates themselves.

In general, results of the 7-factor model were similar across the four item sets. The model fit for each item set is presented in Table 24. Neither model fit the data well, although the fit for the set with excitement and frustration may be the best, and the fit for the set with enjoyment and frustration may be the worst. RMSEA and SRMR supported a good model fit. In contrast, CFI and TLI did not. However, smaller than desired estimates for CFI and TLI may be explained by mostly low-to-moderate correlations between items, which made CFI and TLI harder to detect the improvement of the 7-factor models over null models. Results for the item set with excitement and frustration are presented in Table 25 and Table 26. Results for the other three sets of items can be found in Appendix O. It should be noted that these results should be approached with caution due to the lack of model fit.

Table 24. Model fit statistics for 7-factor models

Model	AIC	BIC	Sample-size adjusted BIC	Scaled Chi Square	RMSEA and 95% CI	CFI	TLI	SRMR
Enjoyed & Frustrated	221127.44	224226.65	222311.24	4440.62	0.040 [0.039, 0.041]	0.899	0.867	0.029
Enjoyed & Anxious	222090.45	225189.67	223274.26	4829.63	0.042 [0.041, 0.044]	0.884	0.846	0.031

Excited & Frustrated	221186.40	224285.61	222370.20	4159.56	0.038 [0.037, 0.039]	0.908	0.879	0.028
Excited & Anxious	222075.30	225174.52	223259.11	4476.86	0.040 [0.039, 0.042]	0.895	0.861	0.031

Note: df = 1476. Results of all Chi Square tests are statistically significant (p < 0.0001).

In general, across item sets, behavioral and cognitive engagement in each instruction type tended to constitute separate factors: Factor 1 for Lecture, Factor 2 for Whole-Class Interaction, Factor 4 for Individual Work, and Factor 5 for Group Work. Importantly, behavioral and cognitive engagement in Whole-Class Interaction included only passive behavioral items. Active behavioral engagement items within Whole-Class Interaction formed a separate factor (Factor 3). Emotional engagement items tended to load on their own factor (Factor 6). Low-to-moderate factor correlations suggest that these factors were distinct.

Finally, there was a factor, which I refer to as the “Difficulty” factor. Positive loadings on this factor come from cross-loadings of “difficult” items (i.e., items with low means). Negative loadings, which were typically low in magnitude, came from very “easy” items (i.e., items with very high means). Other items did not load on the “Difficulty” factor statistically significantly. Important to note that all items cross-loaded on the “Difficulty” factor except items for the active behavioral engagement in Whole-Class Interaction. This factor is very distinct from other factors with no substantial cross-loadings of its items on any other factor. Further, a notable observation about the “Difficulty” factor is as follows: the lower the means were, the higher the item loadings on the “Difficulty” factor tended to be (and the more attenuated their loadings on the substantive factors tended to be). In fact, correlations between loadings on the “Difficulty” factor (excluding items for active behavioral engagement in Whole-Class Interaction) and item means was -0.553 for the item set with Enjoyed and Frustrated, -0.782 for the item set with Excited and Frustrated, -0.836 for the item set with Enjoyed and Anxious, and -0.848 for the item set with Excited and Anxious.

Specific behavioral and cognitive items that cross-loaded on the “Difficulty” factor tended to be similar across item sets, although small differences in statistical significance were observed. Among behavioral items, “difficult” items that tended to load positively and substantially on the “Difficulty” factor were the items about drawing own pictures, writing own remarks or comments, and taking notes in Lecture and Whole-Class Interaction, writing down in detail your task solution or thinking about the task and writing more than one way of solving or of thinking about the task in Individual Work, and taking notes in Group Work. An “easy” item that tended to load negatively was the item about listening in Whole-Class Interaction. Among cognitive items, “difficult” items that tended to load positively and substantially on the “Difficulty” factor were the items about putting information into own words in all types of instruction. “Easy” items that tended to load negatively were the items about answering in your head or thinking about questions your instructor asks the class in Whole-Class Interaction, as well as about keeping the task instructions/question in mind while solving or answering the task and recalling from memory the content needed to solve or/answer the task in Individual Work. Finally, among emotional items, positive loadings on the “Difficulty” factor were frequently observed for feeling excited, whereas negative loadings on the “Difficulty” factor were frequently observed for feeling calm.

In addition to cross-loadings on the “Difficulty” factor, other cross-loadings were also observed. Some of the most salient cross-loadings appeared for some behavioral and cognitive items in Lecture and Whole-Class Interaction that cross-loaded on the behavioral and cognitive engagement in the Individual Work factor. Examples of Lecture items were the items about reading, listening, identifying, connecting, critical thinking, and putting into own words. The last three items even tended to load on the behavioral and cognitive engagement in the Individual Work factor stronger than on the behavioral and cognitive engagement in the Lecture factor. Among Whole-Class Interaction items, a particularly salient cross-loading item was the item about answering in your head or thinking about questions your instructor asks the class. This item also tended to cross-load on the behavioral and cognitive engagement in Individual Work factor stronger than on the behavioral and cognitive engagement in the Whole-Class Interaction factor. Other cross-loadings were observed for Whole-Class Interaction behavioral and cognitive items that tended to load not only on the behavioral and cognitive engagement in the Whole-Class Interaction factor but also on the behavioral and cognitive engagement in Lecture factor. Items about listening and paying attention in Whole-Class Interaction tended to have the largest cross-loadings. The discussed cross-loadings are not illogical, as they reflect individual behaviors and cognitive processes that are characteristic of these instruction types. Yet, the loadings of these items with other behavioral and cognitive items within their instruction type may suggest their conceptual similarity more so with the items of the same instruction type than with a different instruction type.

Another pattern of cross-loadings was observed for emotional items. In general, emotional items, regardless of the instruction type, formed their own single factor (Factor 6). However, not all emotional items loaded strongly or even significantly on this factor. Some emotional items cross-loaded on the behavioral and cognitive factors in their respective instruction types. Among emotional Lecture items, the most salient cross-loading items tended to be enjoyment, excitement, and boredom. Among emotional Whole-Class Interaction items, the most salient cross-loading items tended to be enjoyment and excitement. The item about excitement even tended to load more strongly on the behavioral and cognitive engagement in the Whole-Class Interaction factor than on the emotional factor. Among emotional Individual Work items, the most salient cross-loading items tended to be enjoyment, excitement, and calm. In some item sets, these items even loaded more strongly on the behavioral and cognitive engagement in Individual Work than on the emotional factor. Finally, among emotional Group Work items, the most salient cross-loading items tended to be enjoyment, excitement, calm, and anxiety. Enjoyment, excitement, and calm tended to load substantially more strongly on the behavioral and cognitive engagement in Group Work than on the emotional factor. These cross-loadings of emotional items may suggest that although emotional items, in general, formed their own single factor, they may also be specific to instruction types. In particular, these cross-loadings may indicate a conceptual similarity of emotional items with behavioral and cognitive items within the same instruction type. However, there may also be some disconfirming evidence for this conclusion. Specifically, some emotional items also cross-loaded on factors that represented behavioral and cognitive engagement not in their corresponding instruction types. For example, items about frustration in Group Work tended to cross-load on the Individual Work factor, and items about boredom in Group Work tended to cross-load on the Individual Work factor and on the Lecture factor. Items about anxiety in Individual Work and Whole-Class Interaction tended to have particularly strong cross-loadings on the Group Work factor. Further, items about frustration in Whole-Class Interaction tended to cross-load on the Lecture factor.

An examination of the amount of variance explained by items showed variability. On average, R^2 was 0.463 (*SD* = 0.133) for the item set with Enjoyed and Frustrated, 0.460 (*SD* = 0.131) for the item set with Excited and Frustrated, 0.449 (*SD* = 0.130) for the item set with Enjoyed and Anxious, and 0.442 (*SD* = 0.132) for the item set with Excited and Anxious. Particularly low R^2 was observed for the item about looking at your notes or other resources in Individual Work (R^2 = 0.072) and the item about answering in your head or thinking about questions your instructor asks the class in Whole-Class Interaction (R^2 = 0.248). These items also had very small loadings on their primary factors. Thus, these results suggest that these items might not be conceptually similar to other items within their respective factors and might not indicate their respective constructs.

More problems with the models were revealed during an examination of local misfit. Investigating Lagrange Multiplier Statistics, I found several very high, out of pattern values for error correlations. Most of these high values were for emotional items. In the set with enjoyment and frustration, high values were 84.64 for calm and enjoyment in Individual Work and 98.70 for frustration in Group Work and calm in Individual Work. In the set with enjoyment and anxiety, high values were 124.46 for boredom and enjoyment in Lecture, 158.53 for calm in Individual Work and anxiety in Lecture, 94.54 for boredom in Individual Work and enjoyment in Whole-Class Interaction, and 153.22 for anxiety and calm in Group Work. In the set with excitement and frustration, a high value was 103.75 for frustration in Group Work and calm in Individual Work. In the set with excitement and anxiety, high values were 157.16 for calm in Individual Work and anxiety in Lecture and 154.34 for calm and anxiety in Group Work. This substantial local misfit for emotional items may suggest that emotional items do not fit well together to represent a single factor. Among behavioral and cognitive items, particularly large values were observed for two pairs of items. One pair is putting in own words and re-reading the task before trying to solve or answer it in Individual Work (values ranging from 92.03 to 99.80 across item sets). Another pair is comparing your and other students' ways of thinking about the task and sharing your thinking about the task with other students (values ranging from 113.43 to 132.43 across item sets). This local misfit might be due to the zero-order correlations between the items within each pair being larger than any other zero-order correlations between these items and other items within the same instruction type. Thus, the items within these pairs may be more similar than the similarity accounted for by their factors.

Table 25. Factor loadings and R^2 for the 7-factor model with Excited and Frustrated

Item (abbreviated)	Factor 1: LB+LC	Factor 2: WB $_{\text{passive}}$ + WC	Factor 3: WB $_{\text{active}}$	Factor 4: IB+IC	Factor 5: GB + GC	Factor 6: E	Factor 7: “Diffi-culty”	R^2
LB7_read	**0.552**	-0.049	-0.022	**0.205**	0.012	0.004	-0.043	**0.387**
	0.053	0.044	0.021	**0.063**	0.026	0.028	0.037	**0.039**
LB10_listen	**0.705**	0.044	0.010	**0.254**	-0.047	0.005	-0.029	**0.683**
	0.048	0.035	0.019	**0.072**	0.026	0.026	0.030	**0.021**
LB2_notes	**0.608**	-0.052	0.003	-0.018	0.052	**-0.120**	**0.167**	**0.393**
	0.049	0.042	0.040	0.069	0.041	**0.049**	**0.074**	**0.040**
LB5_pictures	**0.199**	0.042	0.007	0.121	0.030	-0.088	**0.330**	**0.235**
	0.063	0.054	0.044	0.063	0.034	0.049	**0.066**	**0.035**
LB13_remarks	**0.170**	-0.027	0.047	**0.170**	**0.088**	-0.024	**0.459**	**0.354**
	0.070	0.049	0.031	**0.080**	**0.042**	0.047	**0.063**	**0.050**
LC3_attention	**0.720**	-0.013	**0.046**	0.113	-0.002	0.065	**0.134**	**0.636**
	0.043	0.034	**0.023**	0.069	0.031	0.038	**0.052**	**0.025**
LC6_identify	**0.453**	0.085	0.002	**0.329**	-0.020	**-0.049**	-0.039	**0.441**
	0.055	0.074	0.033	**0.067**	0.032	**0.024**	0.030	**0.031**
LC15_connect	**0.260**	0.112	0.056	**0.443**	0.000	**0.085**	-0.022	**0.453**
	0.055	0.070	0.035	**0.045**	0.029	**0.033**	0.045	**0.027**
LC12_critical	**0.286**	0.068	0.051	**0.419**	-0.042	0.058	**0.210**	**0.469**
	0.064	0.079	0.035	**0.050**	0.039	0.036	**0.047**	**0.036**
LC9_ownwords	0.137	0.087	0.043	**0.328**	0.026	**-0.097**	**0.247**	**0.315**
	0.075	0.077	0.036	**0.068**	0.050	**0.031**	**0.065**	**0.031**
LE11_excited	**0.169**	0.116	0.040	0.029	-0.051	**0.395**	**0.465**	**0.479**
	0.042	0.062	0.028	0.041	0.035	**0.045**	**0.070**	**0.036**
LE14_calm	**0.105**	0.024	0.019	**0.152**	-0.029	**0.566**	**-0.169**	**0.462**
	0.047	0.046	0.023	**0.043**	0.032	**0.035**	**0.065**	**0.032**
LE4_frustrated_rec	0.006	-0.053	-0.027	0.008	0.006	**0.784**	**0.153**	**0.600**
	0.033	0.038	0.023	0.029	0.030	**0.029**	**0.077**	**0.040**
LE8_bored_rec	**0.341**	0.064	-0.053	-0.063	-0.007	**0.582**	0.382	**0.636**
	0.051	0.069	0.033	0.042	0.026	**0.052**	0.104	**0.052**
WB6_volunteer	**0.066**	0.000	**0.872**	-0.010	0.011	0.003	-0.018	**0.759**
	0.027	0.019	**0.023**	0.019	0.022	0.019	0.023	**0.034**
WB14_shared	-0.034	**0.098**	**0.827**	-0.013	**0.047**	-0.008	0.016	**0.738**
	0.026	**0.036**	**0.018**	0.020	**0.021**	0.016	0.017	**0.023**
WB18_asked	**0.054**	0.039	**0.734**	0.011	0.044	**-0.089**	0.043	**0.587**
	0.026	0.029	**0.026**	0.021	0.025	**0.034**	0.025	**0.032**
WB9_listen	**0.250**	**0.571**	-0.004	-0.004	0.058	**0.110**	**-0.133**	**0.558**
	0.063	**0.051**	0.023	0.057	0.038	**0.036**	**0.059**	**0.032**
WB3_notes	**0.117**	**0.495**	-0.007	-0.059	0.017	**-0.142**	**0.184**	**0.392**
	0.045	**0.063**	0.033	0.047	0.044	**0.030**	**0.071**	**0.033**
WB11_pictures	-0.023	**0.417**	0.003	0.050	-0.020	-0.006	**0.385**	**0.413**
	0.036	**0.035**	0.025	0.048	0.032	0.029	**0.056**	**0.037**
WB16_remarks	-0.017	**0.349**	0.052	0.037	-0.002	**-0.118**	**0.395**	**0.395**
	0.055	**0.056**	0.034	0.057	0.037	**0.042**	**0.062**	**0.035**

Item (abbreviated)	Factor 1: LB+LC	Factor 2: WB $_{\text{passive}}$ + WC	Factor 3: WB $_{\text{active}}$	Factor 4: IB+IC	Factor 5: GB + GC	Factor 6: E	Factor 7: “Diffi-culty”	R^2
WC13_answeredhead	**0.149**	**0.183**	-0.006	**0.243**	**0.106**	0.046	**-0.110**	**0.250**
	0.053	**0.057**	0.033	**0.055**	**0.041**	0.033	**0.047**	**0.025**
WC19_attention	**0.328**	**0.404**	0.020	0.010	0.069	**0.115**	0.051	**0.493**
	0.049	**0.063**	0.020	0.037	0.042	**0.026**	0.059	**0.025**
WC7_identify	0.024	**0.650**	**0.044**	**0.183**	-0.023	-0.039	-0.015	**0.556**
	0.037	**0.052**	**0.022**	**0.087**	0.026	0.027	0.023	**0.028**
WC10_connect	0.066	**0.648**	-0.011	**0.173**	0.004	0.043	-0.015	**0.592**
	0.046	**0.066**	0.022	**0.070**	0.026	0.031	0.027	**0.033**
WC4_critical	-0.035	**0.654**	0.000	**0.133**	-0.004	0.030	**0.120**	**0.553**
	0.034	**0.040**	0.025	**0.066**	0.026	0.026	**0.042**	**0.026**
WC2_ownwords	-0.061	**0.422**	0.001	0.106	0.030	**-0.086**	**0.154**	**0.280**
	0.043	**0.051**	0.031	0.070	0.045	**0.034**	**0.056**	**0.028**
WE5_excited	-0.005	**0.324**	**0.091**	-0.063	0.036	**0.289**	**0.377**	**0.413**
	0.044	**0.067**	**0.035**	0.046	0.025	**0.046**	**0.084**	**0.034**
WE8_calm	0.036	**0.150**	**0.069**	0.095	0.046	**0.513**	**-0.221**	**0.421**
	0.039	**0.040**	**0.029**	0.061	0.030	**0.045**	**0.052**	**0.031**
WE15_frustrated_rec	**0.220**	0.105	-0.028	**-0.123**	**0.099**	**0.564**	**0.322**	**0.513**
	0.051	0.075	0.021	**0.045**	**0.036**	**0.050**	**0.115**	**0.047**
WE1_bored_rec	-0.056	0.032	-0.001	0.005	-0.028	**0.727**	-0.121	**0.560**
	0.045	0.040	0.030	0.036	0.036	**0.036**	0.067	**0.028**
IB15_reread	0.023	0.006	**-0.110**	**0.454**	**0.186**	-0.042	0.099	**0.320**
	0.039	0.038	**0.033**	**0.038**	**0.049**	0.034	0.068	**0.029**
IB11_looked	0.037	0.036	**-0.099**	**0.113**	**0.139**	**-0.095**	0.042	**0.072**
	0.056	0.071	**0.035**	**0.053**	**0.054**	**0.044**	0.057	**0.018**
IB7_checked	0.116	0.026	-0.005	**0.529**	**0.119**	0.017	-0.044	**0.412**
	0.064	0.052	0.026	**0.046**	**0.059**	0.033	0.043	**0.030**
IB2_write	**0.127**	-0.011	0.008	**0.312**	0.079	-0.068	**0.261**	**0.259**
	0.051	0.040	0.041	**0.069**	0.045	0.040	**0.061**	**0.033**
IB17_wrotedifways	-0.051	-0.069	0.061	**0.372**	0.030	-0.023	**0.496**	**0.417**
	0.050	0.051	0.037	**0.076**	0.030	0.032	**0.072**	**0.037**
IC3_recall	**0.153**	0.083	**-0.090**	**0.399**	**0.100**	**-0.107**	**-0.169**	**0.301**
	0.067	0.043	**0.036**	**0.059**	**0.050**	**0.036**	**0.041**	**0.038**
IC13_keepinmind	0.104	-0.001	**-0.070**	**0.646**	**0.116**	0.039	-0.118	**0.537**
	0.059	0.034	**0.031**	**0.043**	**0.049**	0.033	0.067	**0.037**
IC9_why	0.018	0.024	0.021	**0.624**	0.020	**0.045**	0.124	**0.473**
	0.038	0.051	0.027	**0.033**	0.039	**0.021**	0.076	**0.034**
IC6_thoughtdifways	-0.091	0.029	**0.071**	**0.508**	-0.017	0.007	**0.238**	**0.352**
	0.058	0.062	**0.036**	**0.056**	0.044	0.029	**0.081**	**0.050**
IC5_identify	0.033	0.054	-0.070	**0.578**	**0.122**	-0.041	-0.040	**0.426**
	0.058	0.052	0.037	**0.041**	**0.059**	0.035	0.058	**0.034**
IC12_critical	0.045	0.049	-0.018	**0.671**	0.069	0.041	0.101	**0.579**
	0.038	0.031	0.021	**0.034**	0.043	0.029	0.078	**0.031**
IC16_ownwords	**-0.145**	0.030	**-0.065**	**0.418**	**0.130**	-0.013	**0.322**	**0.339**
	0.048	0.034	**0.030**	**0.061**	**0.051**	0.036	**0.078**	**0.036**

Item (abbreviated)	Factor 1: LB+LC	Factor 2: WB $_{passive}$ + WC	Factor 3: WB $_{active}$	Factor 4: IB+IC	Factor 5: GB + GC	Factor 6: E	Factor 7: “Difficulty”	R^2
IE14_excited	-0.013	-0.078	**0.133**	**0.316**	**-0.133**	**0.223**	**0.347**	**0.308**
	0.037	0.046	**0.039**	**0.051**	**0.045**	**0.036**	**0.076**	**0.037**
IE10_calm	-0.074	-0.033	**0.121**	**0.351**	**-0.169**	**0.438**	-0.137	**0.400**
	0.052	0.050	**0.039**	**0.056**	**0.044**	**0.043**	0.100	**0.036**
IE18_frustrated_rec	-0.005	**-0.112**	-0.033	0.013	**0.101**	**0.712**	**-0.178**	**0.586**
	0.027	**0.045**	0.027	0.022	**0.034**	**0.041**	**0.066**	**0.030**
IE4_bored_rec	0.020	0.017	**-0.079**	**-0.095**	**0.122**	**0.754**	-0.001	**0.551**
	0.033	0.047	**0.028**	**0.041**	**0.038**	**0.028**	0.075	**0.028**
GB3_asked	**0.081**	0.003	0.019	**-0.155**	**0.781**	**-0.072**	-0.066	**0.585**
	0.034	0.031	0.023	**0.036**	**0.030**	**0.023**	0.051	**0.035**
GB7_justified	-0.028	0.007	**0.057**	**0.207**	**0.543**	**0.075**	-0.012	**0.436**
	0.044	0.042	**0.026**	**0.040**	**0.043**	**0.030**	0.037	**0.033**
GB10_checked	0.043	-0.043	0.016	-0.014	**0.710**	-0.042	**-0.154**	**0.492**
	0.042	0.041	0.029	0.040	**0.029**	0.028	**0.053**	**0.036**
GB16_shared	-0.031	-0.070	**0.157**	0.088	**0.738**	**0.074**	-0.003	**0.602**
	0.035	0.041	**0.031**	0.045	**0.029**	**0.029**	0.040	**0.026**
GB13_notes	0.023	0.066	-0.055	-0.022	**0.425**	**-0.124**	**0.334**	**0.385**
	0.055	0.053	0.036	0.066	**0.035**	**0.043**	**0.059**	**0.042**
GC9_attention	**0.185**	0.037	-0.033	0.065	**0.507**	0.057	-0.040	**0.395**
	0.050	0.044	0.025	0.044	**0.048**	0.035	0.039	**0.029**
GC17_compared	**-0.079**	-0.038	**0.061**	**0.138**	**0.702**	0.018	**0.094**	**0.576**
	0.033	0.042	**0.027**	**0.046**	**0.031**	0.025	**0.048**	**0.035**
GC12_use	-0.013	**0.076**	**-0.069**	-0.058	**0.661**	**-0.090**	**-0.129**	**0.457**
	0.031	**0.037**	**0.029**	0.037	**0.024**	**0.031**	**0.052**	**0.031**
GC2_identify	0.104	0.017	-0.002	0.031	**0.693**	0.017	0.017	**0.561**
	0.057	0.047	0.025	0.033	**0.030**	0.024	0.039	**0.024**
GC15_connect	**-0.056**	0.069	0.006	**0.101**	**0.714**	0.016	**0.069**	**0.625**
	0.026	0.015	0.018	**0.046**	**0.026**	0.019	**0.035**	**0.023**
GC4_critical	0.042	0.008	0.001	0.030	**0.731**	-0.019	**0.077**	**0.586**
	0.055	0.043	0.025	0.037	**0.031**	0.023	**0.028**	**0.029**
GC6_ownwords	**-0.091**	0.074	-0.003	**0.157**	**0.499**	-0.033	**0.145**	**0.398**
	0.043	0.053	0.030	**0.044**	**0.034**	0.026	**0.040**	**0.034**
GE1_excited	0.000	0.017	**0.112**	-0.080	**0.549**	**0.116**	**0.175**	**0.373**
	0.046	0.042	**0.028**	0.047	**0.028**	**0.033**	**0.047**	**0.028**
GE11_calm	-0.006	0.007	**0.056**	0.040	**0.469**	**0.300**	**-0.245**	**0.369**
	0.043	0.051	**0.026**	0.045	**0.042**	**0.046**	**0.048**	**0.031**
GE18_frustrated_rec	**-0.108**	**-0.083**	0.021	**0.164**	**-0.100**	**0.676**	0.008	**0.513**
	0.044	**0.036**	0.035	**0.044**	**0.033**	**0.027**	0.061	**0.031**
GE8_bored_rec	**0.176**	-0.055	-0.035	**0.153**	**-0.070**	**0.411**	**0.153**	**0.273**
	0.054	0.063	0.038	**0.045**	**0.035**	**0.041**	**0.072**	**0.033**

Note. For each factor, standardized loadings are presented. Standard errors are presented in the second line. Highlighted in yellow are loadings for items that represent a substantive factor. In bold are statistically significant loadings ($p < 0.05$).

Table 26. Factor correlations for the 7-factor model with Excited and Frustrated

	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	Factor 6	Factor 7
Factor 1: LB+LC	-						
Factor 2: WB $_{passive}$ + WC	**0.458** **(0.055)**	-					
Factor 3: WB $_{active}$	0.021 (0.046)	**0.193** **(0.040)**	-				
Factor 4: IB+IC	**0.307** **(0.060)**	**0.393** **(0.069)**	**0.180** **(0.044)**	-			
Factor 5: GB + GC	**0.248** **(0.047)**	**0.441** **(0.049)**	**0.134** **(0.033)**	**0.367** **(0.051)**	-		
Factor 6: E	**0.167** **(0.038)**	0.081 (0.052)	**0.103** **(0.043)**	**0.172** **(0.056)**	-0.055 (0.051)	-	
Factor 7: “Difficulty”	**0.088** **(0.044)**	**0.270** **(0.046)**	**0.235** **(0.032)**	**0.111** **(0.049)**	**0.122** **(0.037)**	**-0.117** **(0.036)**	-

Note. In bold are statistically significant loadings (p < 0.05). Standard errors are presented in parentheses.

In sum, ESEM results provided support for the separation of behavioral and cognitive engagement from emotional engagement. The results also showed that behavioral and cognitive engagement were not distinct dimensions of engagement as measured by my instrument. Further, the results provided support for instructional specificity of behavioral and cognitive engagement. Thus, overall, the results demonstrated the potential for combining multidimensionality and instructional specificity in engagement measurement in STEM classes. However, the emergence of the “Difficulty” factor, as well as model fit problems and the presence of cross-loadings (especially for the emotional items), suggest that the 7-factor ESEM model might not represent the internal structure of the instrument, despite being the most interpretable model.

Notably, I do not view the “Difficulty” factor as a methodological factor. Rather, I view it as a factor that represents ways to engage in class that may require more effort or motivation. If I were to ignore the “Difficulty” factor, the most “difficult” items may seem as “bad” items due to their low loadings on their substantive factors. Yet, as these items have lower means and often larger variances, they may be able to estimate students’ engagement levels more precisely and, hence, better differentiate between engagement levels of different students. Removing “difficult” items would effectively shrink the content validity of the instrument to “easy” engagement. One possible explanation for the emergence of the “Difficulty” factor is an inappropriate application of a linear factor analysis to items with different “difficulty.” “Difficulty” factors have been known to occur in such applications because “more factors than content would demand are required to reduce the residuals to a random pattern” (Gibson, 1960, p. 381). In other words, multiple factors are needed to account for the relationships among items and achieve local independence; that is, the number of factors will be larger than what would be expected content-wise. The problem can be demonstrated with my engagement items. Specifically, items with similar “difficulty” showed a linear relationship, i.e., one can see that if a student engaged in one way, they are likely to also be engaged in another way. For example, Figure 2 presents a scatterplot with two “easy” items: paying attention and trying connect the information with prior knowledge in Whole-Class Interaction. The scatterplot shows that students who pay attention are also likely to try to connect the information with prior knowledge and vice versa. Yet, when items with different “difficulty” are considered, one can see that engaging in an “easier” way does not necessarily mean also engaging in a more “difficult” way. For example, Figure 3 presents a scatterplot for note-taking (a more “difficult” item) and paying attention (an “easier” item) in Whole-Class Interaction. In the scatterplot, one can see that some students are highly attentive but at the same time do not take notes (see the lower right corner of the scatterplot). Fabrigar amd Wegener (2011) explained that nonlinearity occurs when the ability of indicators to differentiate between people with diffirent levels of the construct is not constant across the levels of the construct. In my case, more “difficult” items are able to better differentiate between highly engaged students, whereas “easier” items are able to better differentiate between not very engaged students. Thus, a potential solution to the “Difficulty” factor is an application of a nonlinear factor analysis.

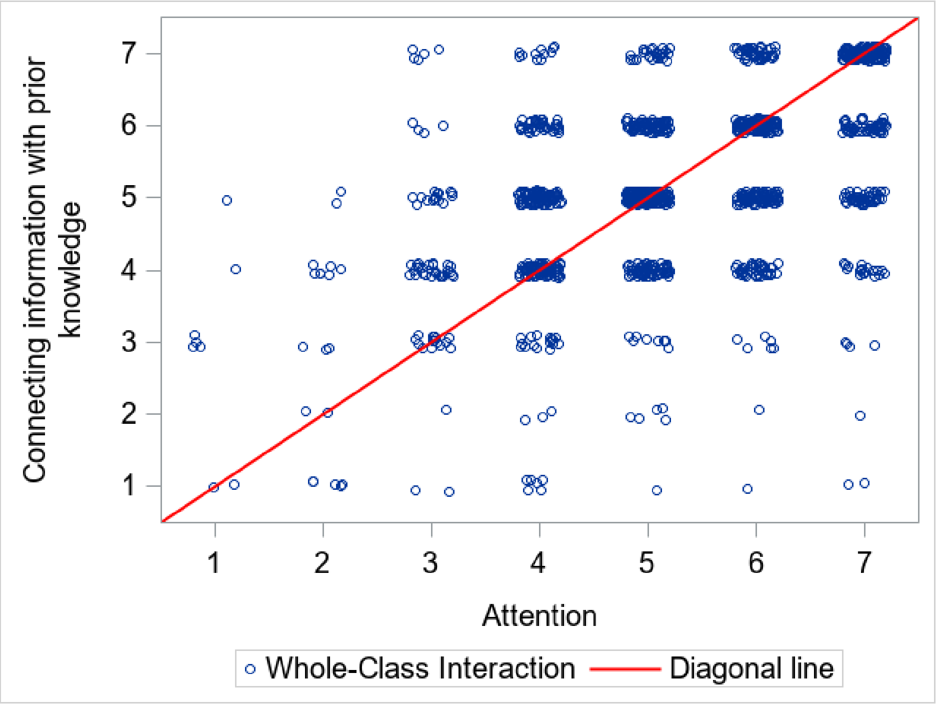


Figure 2. Scatterplot for connecting information with prior knowledge and paying attention in Whole-Class Interaction

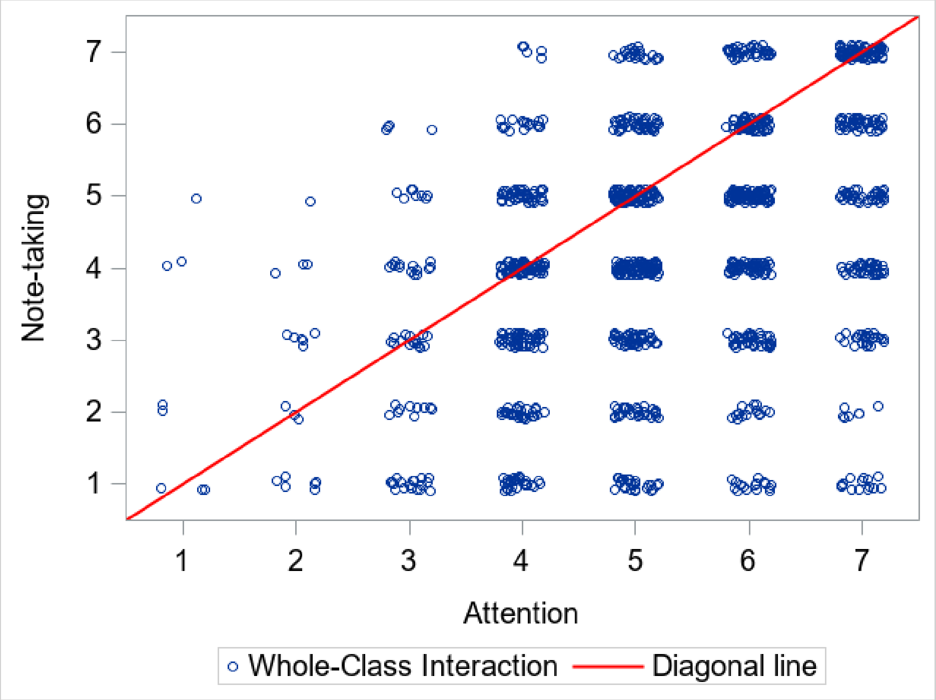


Figure 3. Scatterplot for attention and note-taking in Whole-Class Interaction

While modeling nonlinear relationships between the indicators and latent variables may solve the problem of the "Difficulty" factor, it might not solve all problems with the model lack of fit, especially the lack of fit caused by emotional items. Some emotional items, with their cross-loadings and lack of local fit, may seem as "bad" items. Yet, the observed results may indicate that the emotions are not similar enough to represent a single factor rather than that they are not good indicators of emotional engagement. Different emotions may contribute uniquely to emotional engagement, leading to a more precise estimation of emotional engagement levels and, hence, better differentiation between engagement levels of different students. Removing some emotions would also affect content validity, leading to an imbalance between activating and deactivating emotions, as well as between positive and negative emotions. The problems with the 7-factor ESEM model (the “Difficulty” factor, the lack of fit, and the presense of cross-loadings) prompted me to reconsider my approach to engagement measurement.

Approaches to measurement. Two measurement approaches that are often contrasted are reflective and formative measurement. I will describe each approach briefly in general terms first before applying them to engagement measurement.

Reflective measurement. The approach I initially adopted was reflective measurement, which has been commonly used in engagement measurement and measurement of psychological constructs more broadly. In reflective measurement, a construct – and a latent variable, which is a “stand-in” for a construct in a model (Bollen & Diamantopoulos, 2017) – is assumed to be real, i.e., it exists independently of its measurement (e.g., Borsboom et al., 2003). An example of a reflective construct is depressed affect (Bainter & Bollen, 2014). Reflective indicators have conceptual unity, as they represent one construct (Bollen & Bauldry, 2011). Conceptually, a latent variable in reflective measurement is “whatever a set of reflective indicators have in common” (Rhemtulla et al., 2020, p. 31). In other words, only shared variance among indicators is of interest.

Next, changes in the latent variable cause changes in reflective indicators (e.g., Borsboom et al., 2003). In a reflective model, indicators are regressed on the latent variable (e.g., Borsboom et al., 2003). Regression coefficients are structural coefficients that represent an expected change in the indicator when the latent variable changes by one unit (Bollen & Bauldry, 2011). Error terms are specific to indicators; the error terms encompass all other influences on the indicators besides the latent variable (Bollen & Diamantopoulos, 2017). Reflective models assume local independence, i.e., indicators are independent, conditioning on the latent variable (e.g., Borsboom et al., 2003). Reflective indicators are also referred to as effect indicators (e.g., Bainter & Bollen, 2014).

Further, reflective indicators (with similar validity and reliability) within a single dimension are interchangeable, i.e., removing an indicator would not influence the relationship of other indicators with the latent variable (Bollen & Bauldry, 2011). Additionally, reflective indicators should be positively correlated if they are positively related to the latent variable (Bollen & Bauldry, 2011). Finally, reflective models are identified if there are at least three indicators, there are no error correlations, and there is a scale for the latent variable (e.g., Edwards, 2011). Many commonly used methods, such as Cronbach's alpha and EFA, were developed for reflective measurement (Bollen & Diamantopoulos, 2017).

Formative measurement. An alternative approach is formative measurement. Here, a construct is not assumed to be real, and it does not need to exist independently of its measurement (e.g., Borsboom et al., 2003). While several ontological stances are compatible with formative measurement, the most applicable to construct measurement is the constructivist stance. Within the constructivist stance, a construct is developed by a person (e.g., Borsboom et al., 2003). However, Bollen and Diamantopoulos (2017) argue that a construct does exist independently from its indicators because a construct is defined first; then, it guides the selection of indicators. Examples of formative constructs are SES (e.g., Borsboom et al., 2003) and the degree of social interaction (Bainter & Bollen, 2014). Formative indicators, similarly to reflective indicators, have conceptual unity, as they represent one construct (Bollen & Bauldry, 2011). However, differently from reflective models, unique parts of indicators are of interest; they are viewed as a part of the construct (Rhemtulla et al., 2020). Modeling a formative construct as reflective may result in removing a part of the construct under the disguise of correcting for measurement error (Rhemtulla et al., 2020).

Next, also differently from reflective models, changes in formative indicators lead to changes in the latent variable (e.g., Borsboom et al., 2003). The causal nature of the relationships between indicators and the latent variable in a formative model has been a subject to debate. For example, Lee and Chamberlain (2016) argued that formative indicators cannot cause the construct because in order for causes to cause effects, (1) effects need to be real and (2) causes and effects need to be distinct. Thus, instead of causing the construct, formative indicators form it (specifically, they form a composite). Some researchers also argue that formative measurement is not measurement per se because formative indicators are not measures of a construct; rather, they are measures of attributes that construct a composite (Markus & Borsboom, 2013). Next, it is important to note that in order for the level of the formative construct to change, it is enough to change the level of a single indicator (Bollen & Bauldry, 2011).

Further, in a formative model, a latent variable is regressed on indicators (e.g., Borsboom et al., 2003). Regression coefficients are structural coefficients that represent an expected change in the latent variable when an indicator changes by one unit, controlling for other indicators (Bollen & Bauldry, 2011). Indicators are also correlated with each other (e.g., Borsboom et al., 2003). An error term is specific to the latent variable; the error term encompasses all other influences on the latent variable besides the indicators and is uncorrelated with the indicators (Bollen & Diamantopoulos, 2017). When the error term is not included, we have a composite variable instead of a latent variable; the composite variable is completely determined by its indicators (Bainter & Bollen, 2014). Formative indicators are also referred to as causal indicators (e.g., Bainter & Bollen, 2014).

Further, differently from reflective indicators, formative indicators are not interchangeable. Removing an indicator would change the meaning of the construct as well as bias weights of the remaining indicators in a latent variable model if the removed indicator was correlated with them (Bollen & Bauldry, 2011). Thus, in case of formative measurement, a census of indicators (i.e., a set of all indicators that form the construct) is needed (Bollen & Lennox, 1991). Additionally, formative indicators are not expected to have a specific pattern of correlations; they can also be uncorrelated (Bollen & Bauldry, 2011). While reflective indicators correlate to the extent they are caused by the same construct, formative indicators correlate to the extent they have common antecedents (Fabrigar & Wegener, 2011). Finally, in contrast to reflective models, formative models are not identified on their own; for the model to be identified, at least two outcomes (or reflective measures) are required (e.g., Borsboom et al., 2003).

Case for formative measurement of engagement. The first aspect, in which formative and reflective measurement perspectives differ, is the nature of the construct.

From the perspective of reflective measurement, student engagement is viewed as an internal quality of a student. This internal quality exists independently from the specific ways of engagement that researchers measure (e.g., note-taking and paying attention). Student engagement is assumed to be manifested in these ways. From the perspective of formative measurement, the meaning of student engagement is constructed by researchers. In particular, researchers specify ways of engagement (such as note-taking or paying attention) in accordance with the engagement conceptualization that they developed. These ways of engagement form the meaning of student engagement. Considering my conceptualization of student engagement, it seems more plausible that student engagement is constructed rather than real.

The second aspect, in which formative and reflective measurement perspectives differ, is the direction of causality. From the perspective of reflective measurement, students' internal levels of engagement lead them to have particular levels of specific ways of engagement. In other words, students take notes and pay attention because they are engaged. From the perspective of formative measurement, students' levels of specific ways of engagement lead researchers to develop students' levels of engagement. In other words, students are said to be engaged because they take notes and pay attention. It seems that the perspective of formative measurement on the direction of causality is more plausible than the perspective of reflective measurement.

The third aspect, in which formative and reflective measurement perspectives differ, is interchangeability of indicators. From the perspective of reflective measurement, different ways of engagement are assumed to be interchangeable. It should not matter whether students are asked about note-taking or paying attention. Only what is common between the ways of engagement is of interest; everything else is not of relevance to the construct. From the perspective of formative measurement, different ways of engagement are not assumed to be interchangeable, as each way of engagement contributes unique information about the construct of student engagement. For this reason, excluding note-taking or paying attention would change the meaning of the engagement construct. It does not seem plausible to view ways of engagement as interchangeable, as reflective measurement assumes. In fact, when we are interested in student engagement, we are likely to be interested in a variety of ways, in which students are engaged, in order to capture the full picture of student engagement in a particular setting. It seems plausible that adding or removing some ways of engagement would change what we mean by student engagement.

Besides the differences between reflective and formative measurement, there is one aspect that is characteristic of both measurement approaches. This aspect is conceptual unity. It can be argued that engagement indicators do have conceptual unity in order to represent a construct of student engagement. Items within each hypothesized subscale were conceptually similar in the sense that they represented behaviors, cognitive processes, or emotions in particular instruction types.

The problems encountered in the ESEM analysis are also interpreted differently, depending on which measurement approach is considered. One of the problems was the lack of model fit and the presence of cross-loadings. From the perspective of reflective measurement, poor model fit and cross-loadings mean that poor indicators were used.

Thus, if note-taking does not load cleanly on one factor, then it should be removed. Removing poor indicators is assumed to improve measurement and not affect the meaning of the engagement construct. From the perspective of formative measurement, different ways of engagemet are not expected to fit together. Formative constructs, modeled as reflective constructs, may show misfit in a factor analysis, as a factor analysis is designed to detect an internal structure of reflective constructs (Bollen & Diamantopoulos, 2017). Formative indicators are not expected to fit in a reflective model because such indicators are not effects of common causes. Thus, removing, for example, note-taking on the basis of poor fit would lead to the underrepresentation of the engagement construct. In other words, if engagement is viewed as a formative construct, then the lack of fit in ESEM is not a problem. In fact, it is expected.

Another problem with ESEM was the emergence of the “Difficulty” factor. As discussed above, from the perspective of reflective measurement, the emergence of the “Difficulty” factor is likely to mean that a linear analysis was used inappropriately; instead, a nonlinear analysis should have been used if items had different difficulty. However, this solution assumes that specific ways of engagement are reflective indicators of the engagement construct. Notably, in reflective measurement in our field, it is not typical for items to be of different “difficulty” and, hence, require nonlinear methods. In contrast, formative measurement does not require employing data-driven methods where artificial factors like “Difficulty” factors can emerge. If constructs were specified by researchers, then only these constructs would be developed. It should be noted that data-driven methods for formative measurement exist; however, they are not always appropriate. In terms of item “difficulty,” a range is desired, as it leads to a better representation of the formative construct. For example, note-taking (being a more “difficult” item) helps differentiate between more engaged students, whereas paying attention (being an “easier” item) helps differentiate between less engaged students. Thus, the “Difficulty” factor is also not a problem for formative measurement. In fairness, the “Difficulty” factor is not necessarily a problem for reflective measurement, either, if appropriate methods are used.

In sum, formative measurement seems to be more plausible than reflective measurement when the construct being measured is student engagement. Additionally, the problems with the ESEM analyses can be explained from the perspective of formative measurement. Considering the conceptual arguments for formative measurement of student engagement and against reflective measurement of student engagement, I re-conceptualized student engagement as a formative construct.

Developing formative constructs. Measures of formative constructs, similarly to reflective constructs, are developed by researchers in accordance with the conceptualization of a construct. Yet, the internal structure can also be evaluated empirically. In reflective measurement, factor analyses are commonly used for this purpose. In formative measurement, it is common to fit an SEM model with at least two outcomes (or reflective indicators). However, this method has been largely critisized as being highly prone to interpretational confounding. According to Burt (1976), interpretational confounding is “the assignment of empirical meaning to an unobserved variable which is other than the meaning assigned to it by an individual a priori to estimating unknown parameters. Inferences based on the unobserved variable then become ambiguous and need not be consistent across separate models" (p. 4). Differences in item weights between formative models with different outcomes (or reflective indicators) serve as evidence of interpretational confounding (Bollen & Diamantopoulos, 2017). Howell et al. (2007) showed that when different outcomes are used, weights and, hence, the meaning of the formative construct change. Howell and Breivik (2016) further showed that the meaning of the construct is completely driven by the reflective side of the model (i.e., by the outcomes). In particular, they demonstrated that removing any or all formative indicators from the model does not have any effect on the reflective side (i.e., on the coefficients of paths from the latent variable to outcomes). Alternative methods, such as Principal Components Analysis (PCA), could be used to empirically explore the internal structure. Differently from EFA or ESEM, PCA works with all variance in indicators, i.e., not only with the shared variance.

For my instrument, I decided not to employ the SEM method due to the problem of interpretational confounding. Additionally, modeling formative constructs as latent variables seems counterproductive. It is expected that the census of indicators would minimize the disturbance variance and maximize the explained variance in the latent formative construct. With a small amount of variance remaining to be explained, identifying facilitators of the construct would likely be problematic. As one of the intended uses of the engagement instrument is the use in research that aims to identify facilitators of engagement, the latent variable approach to formative modeling might not be ideal. I also decided not to employ the PCA method because it is somewhat similar to an EFA, as both methods maximize explained variance in indicators through an Eigenvalue decomposition and, therefore, may produce similar results (Markus & Borsboom, 2013). Thus, I expected that PCA would likely produce somewhat similar results to the ESEM analysis. With that said, nonlinear PCA may potentially be promising for exploring the internal structure as it could account for potential nonlinear relationships between indicators and constructs and, hence, resolve the problem of the "Difficulty" factor while at the same time being appropriate for formative measurement. Therefore, nonlinear PCA could be explored as a viable option in the future validation work. In particular, it may be useful for testing whether items within one instruction type and dimension are more similar to each other than to items in different instruction types and dimensions. Indeed, while formative indicators are not expected to be strongly related, it is reasonable to expect that items within one instruction type and dimension would be more strongly related than items in different instruction types and dimensions.

In this study, I developed formative constructs based on the theoretical considerations (i.e., the correspondence between the items and the conceptualizations of the constructs) as well as based on the available results from the ESEM analysis. In developing formative constructs, the idea of conceptual unity plays an important role. Theoretically, I designed the instrument in a way that each subscale (each engagement dimension within each instruction type) has conceptual unity. However, there is a question about whether the subscales are conceptually distinct enough to stand on their own. To answer this question, I turned to the ESEM results. The ESEM results showed that behavioral and cognitive engagement are conceptually similar and cannot be distinguished empirically. Yet, behavioral and cognitive engagement can be differentiated between different instruction types. Further, ESEM results also showed that active behavioral engagement in Whole-Class Interaction is distinct from behavioral and cognitive engagement in Whole-Class Interaction, where behaviors were passive. For emotional engagement, ESEM results suggested a single factor. However, the lack of fit suggests that emotional items might not be conceptually similar enough to represent one factor. Further, some emotional items tended to cross-load on behavioral and cognitive factors in the corresponding instruction types. These observations may suggest that emotional engagement may be differentiated between different instruction types. Thus, I developed nine formative engagement constructs: two dimensions (behavioral/cognitive and emotional) in four instruction types (lecture, whole-class interaction, individual work, and group work), and active behavioral engagement in Whole-Class Interaction. Behaviors included in behavioral/cognitive engagement in Whole-Class Interaction were only passive. Active behavioral engagement in Whole-Class Interaction can be further combined with passive behavioral/cognitive engagement in Whole-Class Interaction on the basis of the same instruction type. A single behavioral/cognitive engagement construct for each instruction type is needed for developing behavioral/cognitive engagement scores at the class level where behavioral/cognitive engagement in each instruction type will be weighted by the amount of class time spent on the corresponding instruction type.

The next step is selecting indicators for each construct. First, I turned back to the ESEM results again to see if there are items that are not conceptually similar to any items in the measure. These items are the items that do not have substantial loadings on any factor and, thus, have a very low R^2. As discussed in the ESEM results, there are two such items: the item about looking at your notes or other resources in Individual Work and the item about answering in your head or thinking about questions your instructor asks the class in Whole-Class Interaction. Thus, I excluded these items from the measure. Further, in formative measurement, redundancy in indicators should be avoided in order not to give more weight to a particular behavior, cognitive process, or emotion. Thus, I examined correlations between items designed to indicate the same construct. The items about listening and paying attention in both Lecture and Whole-Class Interaction had fairly high correlations (r = 0.672 and r = 0.584, respectively). I retained the items about paying attention because they had somewhat lower means and somewhat higher variances. Thus, these items may provide more information about the constructs. Another pair of highly correlated items were in Group Work: the item about sharing your thinking about the task with other students and the item about comparing your and other students' ways of thinking about the task (r = 0.703). I retained the item about comparing for the same reason as for retaining the item about paying attention. Finally, I needed to select a set of emotional items to represent emotional engagement constructs. For an activating positive emotion, I selected excitement over enjoyment because excitement was consistently more "difficult" across instruction types and had lower correlations with other emotional items. Thus, I hypothesized that excitement might provide more information about student emotional engagement. The decision between the two activating negative emotions was made at the stage of creating composite scores.

Conclusion. Overall, with the ESEM analysis and the following re-conceptualization of student engagement as a formative construct, I examined the assumption within the structural aspect of Messick's model. The assumption states that the internal structure of the instrument needs to be determined. Thus, I developed nine first-order (i.e., subscale) formative engagement constructs: two dimensions (behavioral/cognitive and emotional) in four instruction types (Lecture, Whole-Class Interaction, Individual Work, and Group Work), and active behavioral engagement in Whole-Class Interaction. The ESEM analysis provides some evidence for this structure. However, although the ESEM analysis was capable of identifying some sets of items that have conceptual unity in this study, ESEM is not designed for formative constructs and, therefore, generally is not capable of identifying the internal structure of such constructs. Thus, I also used the theoretical considerations, i.e., the information about what the instrument was designed to measure, to further inform the internal structure. In sum, while the internal structure for the instrument was developed, it should be approached with caution and interpreted as tentative because the evidence for this structure is limited.

Creating Composite Engagement Scores

Initially, I aimed to create composite scores for each subscale by averaging corresponding items. I aimed to examine the plausibility of this scoring method by testing whether the items within each factor in the final ESEM model are parallel, i.e., whether they have equal loadings and error variances (McNeish & Wolf, 2020). However, this method is only suitable for reflective measures with established internal structures. As I re-conceptualized student engagement as a formative construct, for which the ESEM model did not represent the internal structure well, testing parallel models became not appropriate. Further, considering that items differed in “difficulty,” I also reconsidered my approach to creating composite engagement scores. In particular, I considered both not weighted and weighted approaches. The not weighted approach to create composite scores is to average the items. This approach was initially planned. It gives each item the same weight and does not take into account item “difficulty.” The weighted approach to creating composites is to compute a weighted average where different items can have different weights.

As items that indicate subscale engagement constructs have different “difficulty” (with the exception of the active behavioral engagement in Whole-Class Interaction), I hypothesized that giving the same weight to such items would not be appropriate. In particular, I hypothesized that more “difficult” items should have larger weights than “easier” items. In other words, engaging in more difficult ways should be recognized more than engaging in easier ways; similarly, not engaging in more difficult ways should be penalized more than not engaging in easier ways. I expected that students who engage in “difficult” ways are likely to also engage in “easy” ways; yet, engaging in “easy” ways does not necessarily mean that students also engage in “difficult” ways. Thus, giving the same weight to all items is likely to consistently result in higher engagement levels, compared to giving larger weights to more “difficult” items. For example, a student with a high score on attention and a low score on note-taking would have a higher engagement score if the two items were given the same weight, compared to giving different weights. Weighted scores would also allow for more appropriate comparisons between students.

To illustrate, one student may have a high score on attention and a high score on note-taking, another student may have a high score on attention but a low score on note-taking, and a third student may have low scores on both. If unweighted averages were computed, the differences in engagement between these students would be the same. However, if the averages were weighted, the difference between the first and second students' engagement levels is larger than the difference between the second and third students' engagement levels. Further, in some situations, it is also possible, albeit less likely, for some students to have high scores on "difficult" items and low scores on "easy" items. For example, a student with a high score on calm and a low score on excitement is less engaged than a student with a low score on calm and a high score on excitement. The difference in engagement levels of these students could not be detected if unweighted average scores were used.

Weights that items have in forming latent variables or composites can be empirically estimated or specified in advance (Bollen & Diamantopoulos, 2017). An empirical estimation of weights involves fitting an SEM model or conducting PCA, which I discussed above. In contrast, some researchers suggested using composites with fixed weights specified by the researcher (Howell et al., 2007; N. Lee & Cadogan, 2013; Rhemtulla et al., 2015). In this study, I decided to specify fixed weights that would follow the expected pattern. As item "difficulty" is reflected in item means, I computed weights in such a way that they reflect item means. In particular, I started by compiling lists of items and their means, with separate lists for each subscale that measures a construct. As I aimed for items with lower means to have larger weights, I flipped the means before computing weights. The flipping of the means was done by subtracting item means from eight; this operation is equivalent to recoding items. Next, I computed an average of the flipped means for each subscale. Weights in each subscale were computed by dividing flipped item means by the average of the flipped means in this subscale. Thus, weights larger than one indicated above average item "difficulty," whereas weights smaller than one indicated below average item "difficulty." For emotional subscales, I examined whether the same weights for each item can be used across instruction types. Overall, differences in weights for the same emotional items in different instruction types were relatively similar (see Table 27 for an example of the item set with excitement and frustration). Thus, to make scoring easier and aid to the comparability of emotional engagement across instruction types, I decided to specify the same weights for the same items across instruction types. Next, I decided to select the set with excitement and frustration over the set with excitement and anxiety because the pattern of weights across instruction types seemed to be more stable in the former set. Finally, for all subscales, I fixed weights in such a way that they can be divided by 0.05 without a remainder, yet keeping the average weight of the subscale to one. Final weights for each subscale are presented in Appendix P. A note needs to be made about the active behavioral engagement in Whole-Class Interaction. As items within this subscale all had approximately the same "difficulty," these items were unit-weighted when calculating composite scores for this subscale.

Table 27. Descriptive statistics for weights of emotional items across instruction types (the set with excitement and frustration)

Item	Mean	*SD*	Min	Max	Range
Excited	1.27	0.06	1.22	1.35	0.13
Calm	0.79	0.06	0.72	0.84	0.12
Not Frustrated	0.89	0.06	0.81	0.94	0.14
Not Bored	1.05	0.05	0.99	1.10	0.10

While I have a theoretical argument for using the weighted approach to create composite engagement scores instead of the not weighted approach, there is a question about whether the superiority of the weighted approach can be supported empirically. To examine this question, I compared both weighted and not weighted composite scores. Creating weighted composite scores for each subscale was conducted in the following way. First, I created weighted item scores by multiplying original item scores by the item weight. Second, I summed weighted item scores. Third, I summed weights of the items, scores on which a student had. Fourth, I divided the sum of weighted item scores by the sum of weights to develop a weighted composite score. I employed steps 2-4 instead of simply averaging weighted item scores because the latter would not be appropriate if missing data are present. Not weighted composites scores for each subscale were created by averaging the subscale items. Finally, composites for active behavioral engagement in Whole-Class Interaction were created by averaging the subscale items.

Engagement dimension composite scores, instruction type composites scores, and global engagement composite scores were created in a way similar to what was planned. Yet, some deviations from the plan occurred. First, I planned to have separate behavioral and cognitive constructs in each instruction type. Instead, I have one behavioral/cognitive construct in each instruction type (with the exception of Whole-Class Interaction). Second, for Whole-Class Interaction, I have two constructs, the differentiation between which was also not planned: behavioral (passive)/cognitive engagement and active behavioral engagement. However, for computing dimension composite scores, I needed to have one composite for behaviors and cognitive processes in Whole-Class Interaction in order to be able to weigh this composite by the amount of time spent on this instruction type. To compute composite scores for a single construct of behavioral/cognitive engagement in Whole-Class Interaction, I averaged behavioral (passive)/cognitive engagement and active behavioral engagement. To distinguish the eight composite scores where behavioral (passive)/cognitive engagement and active behavioral engagement in Whole-Class Interaction was combined in a single composite from the nine subscale composites, I refer to the former as 2x4 composites (two dimensions in four types of instruction).

Next, two dimension composites – behavioral/cognitive engagement and emotional engagement – were created in the following way. First, within each dimension, I created weighted 2x4 composite scores by multiplying 2x4 composite scores by the amount of time spent on the corresponding instruction types. Second, within each dimension, I summed these weighted 2x4 composite scores. Third, within each dimension, I summed the amounts of time spent on all types of instruction, on which a student had 2x4 composite scores. Fourth, within each dimension, I divided the sums of these weighted 2x4 composite scores by the sum of the amounts of time spent on all types of instruction. Further, four instruction type composites – engagement in Lecture, engagement in Whole-Class Interaction, engagement in Individual Work, and engagement in Group Work – were created by averaging 2x4 composite scores within each instruction type. Global engagement composite scores were created by averaging engagement dimension composite scores.

Descriptive statistics for weighted and not weighted composite engagement scores are presented in Table 28. In general, both weighted and not weighted composite engagement scores were approximately normally distributed. As expected, means for weighted composites tended to be lower than for not weighted composites. Further, the weighted approach tended to result in somewhat larger variances; distributions of weighted composites also tended to be less negatively skewed. To graphically compare weighted and not distributions, I present histograms for Behavioral (passive)/Cognitive Engagement in Whole-Class Interaction (see Figure 4) and for Emotional Engagement in Whole-Class Interaction (see Figure 5). The histograms show that in weighted distributions, there tend to be more students with lower scores and fewer students with higher scores, compared to not weighted distributions. However, the observed differences between weighted and not weighted distributions are rather small. In fact, for each subscale, correlations between weighted and not weighted composites were greater than 0.99. Example scatterplots for Behavioral (passive)/Cognitive Engagement in Whole-Class Interaction (see Figure 6) and for Emotional Engagement in Whole-Class Interaction (see Figure 7) show that weighted and not weighted composite scores tended to be similar, with not weighted scores typically being higher than weighted scores. Thus, the empirical difference between the weighted and not weighted approaches is minimal.

Table 28. Descriptive statistics for weighted and not weighted composite engagement scores

Composite	Mean	STD	Skewness	Kurtosis
Behavioral/Cognitive Engagement in Lecture (N = 1143)				
Weighted	4.72	0.92	-0.23	0.23
Not weighted	4.87	0.89	-0.30	0.23
Emotional Engagement in Lecture (N = 1143)				
Weighted	4.43	1.11	-0.13	0.03
Not weighted	4.54	1.10	-0.19	0.05
Behavioral (passive)/Cognitive Engagement in Whole-Class Interaction (N = 1238)				
Weighted	4.33	1.02	-0.11	0.19
Not weighted	4.45	0.99	-0.18	0.28
Emotional Engagement in Whole-Class Interaction (N = 1238)				
Weighted	4.79	1.04	-0.18	0.00
Not weighted	4.91	1.03	-0.27	0.02
Behavioral/Cognitive Engagement in Individual Work (N = 1049)				
Weighted	4.84	0.86	-0.03	0.48
Not weighted	4.96	0.83	-0.10	0.54
Emotional Engagement in Individual Work (N = 1049)				
Weighted	4.18	1.06	-0.03	0.18
Not weighted	4.27	1.07	-0.06	0.16
Behavioral/Cognitive Engagement in Group Work (N = 1193)				
Weighted	4.87	0.94	-0.43	1.02
Not weighted	4.95	0.93	-0.48	1.06
Emotional Engagement in Group Work (N = 1194)				
Weighted	4.96	1.05	-0.29	0.26
Not weighted	5.04	1.03	-0.33	0.26
Active Behavioral Engagement in Whole-Class Interaction (N = 1237)				
Weighted	-	-	-	-
Not weighted	2.93	1.42	0.34	-0.58
Behavioral/Cognitive Engagement in Whole-Class				

Composite	Mean	STD	Skewness	Kurtosis
Interaction (N = 1238)				
Weighted	3.63	0.99	0.20	-0.03
Not weighted	3.69	0.98	0.20	-0.02
Engagement in Lecture (N = 1143)				
Weighted	4.57	0.83	-0.02	0.12
Not weighted	4.70	0.82	-0.07	0.12
Engagement in Whole-Class Interaction (N = 1238)				
Weighted	4.21	0.82	0.13	0.13
Not weighted	4.30	0.81	0.11	0.12
Engagement in Individual Work (N = 1049)				
Weighted	4.51	0.76	0.10	0.44
Not weighted	4.61	0.75	0.09	0.42
Engagement in Group Work (N = 1194)				
Weighted	4.92	0.86	-0.19	0.23
Not weighted	5.00	0.85	-0.21	0.18
Behavioral/Cognitive Engagement (N = 1261)				
Weighted	4.47	0.80	-0.13	0.45
Not weighted	4.57	0.79	-0.16	0.45
Emotional Engagement (N = 1261)				
Weighted	4.68	0.94	-0.22	0.50
Not weighted	4.78	0.93	-0.29	0.55
Global Engagement (N = 1261)				
Weighted	4.58	0.71	0.01	0.47
Not weighted	4.68	0.69	-0.02	0.48

A final note needs to be made about the comparability of engagement composites. Having the same items and the same weights for emotional items across instruction types provides some support for the comparability of emotional engagement levels across instruction types. However, behavioral/cognitive subscales had at least partially different items; the distribution of weights across subscales also differed. Thus, comparisons of behavioral/cognitive engagement levels across instruction types should be approached with caution, as adding or removing “easy” or “difficult” items would change subscale means.

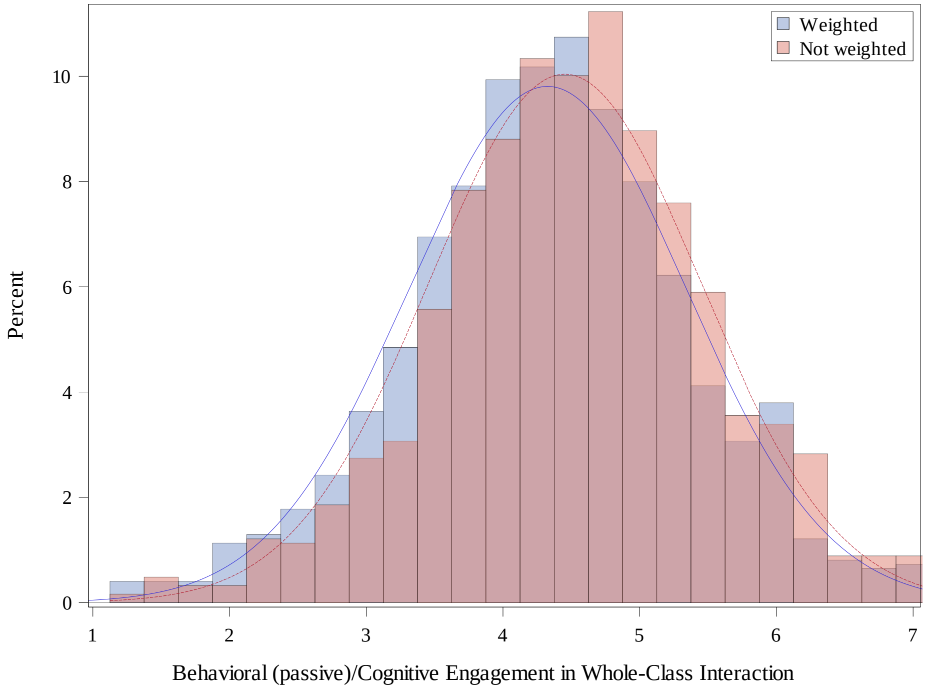


Figure 4. Histogram for Behavioral (passive)/Cognitive Engagement in Whole-Class Interaction

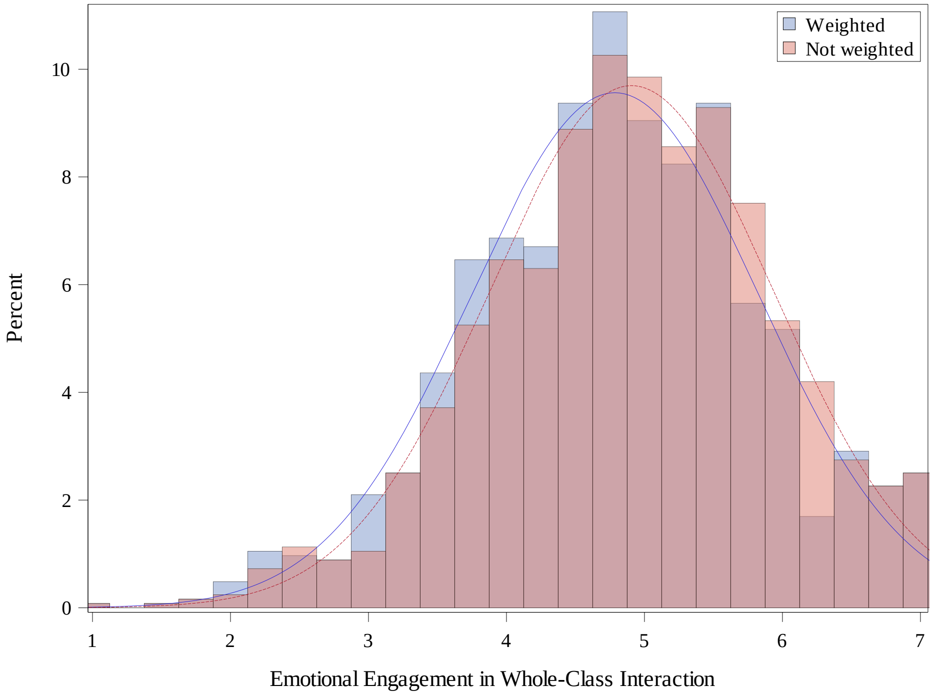


Figure 5. Histogram for Emotional Engagement in Whole-Class Interaction

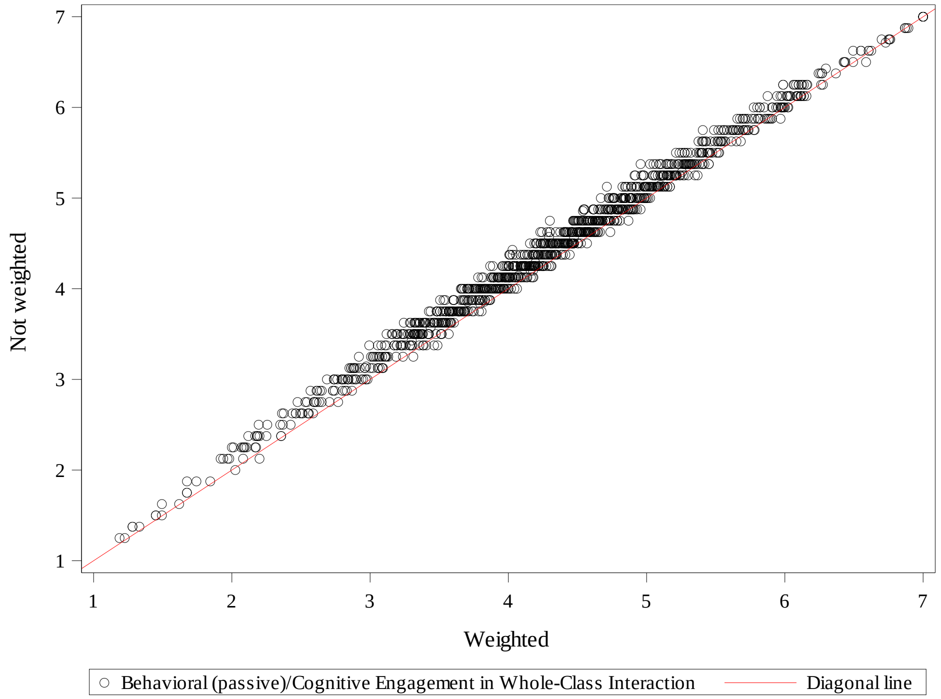


Figure 6. Scatterplot for Behavioral (passive)/Cognitive Engagement in Whole-Class Interaction

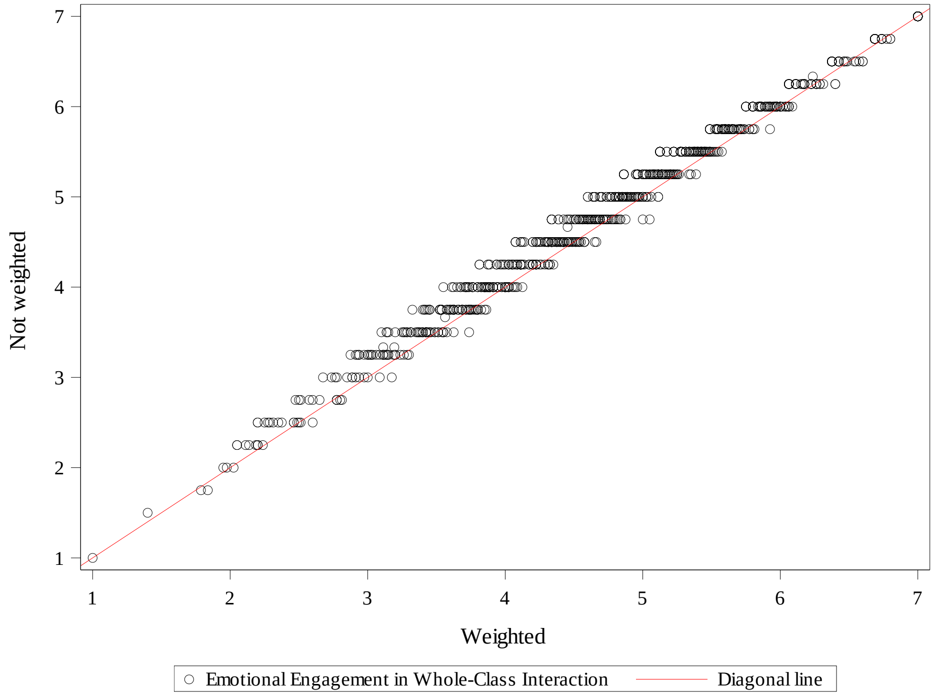


Figure 7. Scatterplot for Emotional Engagement in Whole-Class Interaction

Next, I examined correlations between engagement composites, using both weighted and not weighted approaches. For the not weighted approach, correlations between subscale engagement composites are presented in Table 29, and correlations between dimension composites as well as instruction type composites are presented in Table 30. For the weighted approach, correlations are presented in Appendix Q. Overall, correlations of weighted composites and correlations of not weighted composites were similar, with some correlations being higher for weighted composites and other correlations being higher for not weighted composites. On average, in absolute value, the difference between correlations of weighted composites and correlations of not weighted composites was 0.008 (*SD* = 0.005), with a maximum difference of 0.022 for the correlation between behavioral/cognitive and emotional engagement. Thus, due to the very small difference in correlations between the two approaches, I continue the discussion of correlations with a reference to the correlations of not weighted composites. Almost all correlations were positive and statistically significant. Non-statistically significant correlations were observed between Behavioral/Cognitive Engagement in Group Work and Emotional Engagement in Individual Work, as well as between Behavioral/Cognitive Engagement in Group Work and Emotional Engagement in Lecture. Overall, correlations between subscale constructs tended to be low-to-moderate in magnitude, suggesting that these constructs are empirically distinct. Higher correlations were observed within an engagement dimension (behavioral/cognitive or emotional) across instruction types. The highest correlations were between Emotional Engagement in Lecture and Emotional Engagement in Whole-Class Interaction (r =

0.738) and between Behavioral/Cognitive Engagement in Lecture and Behavioral (passive)/Cognitive Engagement in Whole-Class Interaction (r = 0.679), suggesting that Lecture and Whole-Class Interaction were the most similar instruction types. This result can also be seen from the instruction type correlations, where the largest correlation was between Engagement in Lecture and Engagement in Whole-Class Interaction (r = 0.684). The correlation between dimensions (behavioral/cognitive and emotional) was small (r = 0.287).

Table 29. Correlations between not weighted subscale engagement composites

	LBC	LE	WBC passive	WE	IBC	IE	GBC	GE
LBC								
LE	**0.342**							
WBC passive	**0.679**	**0.267**						
WE	**0.290**	**0.738**	**0.315**					
IBC	**0.584**	**0.243**	**0.522**	**0.266**				
IE	**0.185**	**0.595**	**0.085**	**0.457**	**0.241**			
GBC	**0.408**	0.060	**0.466**	**0.160**	**0.499**	-0.033		
GE	**0.210**	**0.317**	**0.243**	**0.456**	**0.282**	**0.261**	**0.496**	
WB active	**0.234**	**0.137**	**0.285**	**0.178**	**0.218**	**0.149**	**0.196**	**0.163**

Note. LBC = Behavioral/Cognitive Engagement in Lecture; LE = Emotional Engagement in Lecture; WBC passive = Behavioral (passive)/Cognitive Engagement in Whole-Class Interaction; WE = Emotional Engagement in Whole-Class Interaction; IBC = Behavioral/Cognitive Engagement in Individual Work; IE = Emotional Engagement in Individual Work; GBC = Behavioral/Cognitive Engagement in Group Work; GE = Emotional Engagement in Group Work; WB active = Active Behavioral Engagement in Whole-Class Interaction. Statistically significant correlations ($p < 0.01$) are in bold.

Table 30. Correlations between not weighted dimension and instruction type engagement composites

	L	W	I	BC
Instruction type composites:				
Engagement in Lecture (L)				
Engagement in Whole-Class Interaction (W)	**0.684**			
Engagement in Individual Work	**0.626**	**0.512**		
Engagement in Group Work (G)	**0.343**	**0.449**	**0.344**	
Dimension composites				
Behavioral/Cognitive Engagement (BC)				
Emotional Engagement (E)				**0.287**

Note. Statistically significant correlations (*p* < 0.01) are in bold.

Overall, I developed and evaluated the scoring approaches for subscale engagement composites. Plausibility of the scoring method is the assumption within the scoring inference. In particular, I examined the weighted and not weighted approaches. An exploration of composites' descriptive statistics and correlations showed that the two approaches provide similar results. As expected, composite means within the weighted approach were somewhat lower than composite means within the not weighted approach. The weighted approach also tended to produce composite scores with somewhat higher standard deviations and somewhat lower skewness (in absolute values). The weighted approach also has a theoretical rationale. However, give the minimal differences between the two approaches and the greater parsimony of the not weighted approach, I would recommend using the not weighted approach. It should be noted that when using the not weighted approach, it is important to remember that composite scores are likely to be slightly inflated. The matter may need to be re-visited in the future if the instrument undergoes substantial revisions. Yet, in the following analyses in this study, I used not weighted composite scores. Additionally, it is also important to note that comparisons of behavioral/cognitive composites across instruction types should be approached with caution due to the differences in items and in distributions of their weights. Further, the correlations of engagement composites within the weighted and not weighted approaches were similar. Low-to-moderate correlations between composites suggest that the constructs are empirically distinct. In sum, I provided some evidence for the plausibility of the scoring approaches, thus supporting the scoring inference. Additionally, I also examined descriptive statistics and correlations of engagement composites. In general, distributions of composite scores were approximately normal. Mostly low-to-moderate correlations suggest that the constructs are empirically distinct. However, some correlations were on the higher end of acceptable. Thus, I provided some further evidence for the scale functioning component within the substantive aspect of Messick's model.

Preparing External Constructs for Analysis

In this section, I discuss the preparation of external constructs for analysis. The preparation included an investigation of the internal structure and internal consistency for the constructs that were measured via multi-item scales. Further, I discuss the characteristics of scores for all external constructs.

Internal structure and internal consistency. Investigations of the internal structure and internal consistency were conducted for the following constructs, measured via multi-item scales: effort, persistence, feeling- and value-related components of interest, metacognitive strategies, intellect, social efficacy with peers, preference for group work, and public speaking anxiety. The internal structures were improved when original scales indicated problems.

Feeling and value components of interest. A CFA model for two factors – feeling and value components of interest – did not show good fit: scaled Chi Square (34) = 610.740, p < 0.0001, RMSEA = 0.117 with 95%CI [0.109, 0.125], SRMR = 0.046, CFI = 0.951, TLI = 0.935, AIC = 32126.295, and BIC = 32285.328. An examination of modification indices revealed that the Interest-Value item "I find the content of this class personally meaningful" tended to cross-load on the Interest-Feeling factor. Thus, the item was removed. Further, errors of the following two items tended to correlate: "What we are studying in this class is useful for me to know" and "What we are learning in this class is important for my future goals." The result is not necessarily surprising because both items are about usefulness; thus, they are more similar to each other than to other items. Therefore, I removed the item about future goals, as it also tended to cross-load on the Interest-Feeling factor. A further examination of modification indices did not reveal extremely high values. The fit of the final model was adequate: Chi Square (19) = 166.515, p < 0.0001, RMSEA = 0.079 with 95%CI [0.068, 0.090], SRMR = 0.025, CFI = 0.983, TLI = 0.975, AIC = 25353.949, and BIC = 25482.181. Loadings on the 4-item Interest-Feeling factor ranged from 0.878 to 0.931, with a mean of 0.905 (*SD* = 0.022). Loadings on the 4-item Value-Interest factor ranged from 0.661 to 0.849, with a mean of 0.794 (*SD* = 0.089). The correlation between the two interest components was 0.852. Cronbach's alphas for Interest-Feeling and Interest-Value were 0.947 and 0.867, respectively.

Effort and persistence. A CFA model for Effort and Persistence did not show good fit: scaled Chi Square (8) = 172.947, p < 0.0001, RMSEA = 0.128 with 95%CI [0.112, 0.145], SRMR = 0.043, CFI = 0.958, TLI = 0.920, AIC = 18796.604, and BIC = 18894.076. An examination of modification indices revealed that the Persistence item "In this class, regardless of whether or not I like the material, I work my hardest to learn it" tends to cross-load on the Effort factor. This result is not surprising given that this item is similar to the Effort items in that they all are about hard work. Therefore, I removed the item. A further examination of modification indices did not reveal extremely high values. The fit of the final model was good: scaled Chi Square (4) = 21.355, p = 0.0003, RMSEA = 0.059 with 95%CI [0.036, 0.085], SRMR = 0.021, CFI = 0.995, TLI = 0.987, AIC = 16001.790, and BIC = 16083.872. Loadings on the 2-item Effort factor were 0.879 and 0.935; loadings on the 3-item Persistence factor were 0.600, 0.675, and 0.829. The correlation between the factors was 0.664. Cronbach's alphas for Effort and Persistence were 0.902 and 0.727, respectively.

Metacognitive strategies. A CFA model for Metacognitive Strategies did not show good fit: scaled Chi Square (27) = 167.609, p < 0.0001, RMSEA = 0.065 with 95%CI [0.055, 0.074], SRMR = 0.038, CFI = 0.939, TLI = 0.919, AIC = 32443.878, and BIC = 32582.391. An examination of standardized loadings revealed that one item ("In this class, I start my assignments without really planning out what I want to get done") had a low loading (0.286). This result may be due to this item being the only recoded item in the scale. Thus, I removed this item. Further, an examination of modification indices revealed that errors of the following two items tended to correlate: "I try to change the way I study for this class to fit the type of material I am trying to learn" and "If what I am working on for this class is difficult to understand, I change the way I learn the material." This result may be due to the items being more similar to each other (both are about changing ways of learning) than to other items in the scale. As the first item had somewhat lower standardized loading than the second item, I removed the first item. A further examination of modification indices did not reveal extremely high values. The fit of the final model was adequate: scaled Chi Square (14) = 51.500, $p < 0.0001$, RMSEA = 0.046 with 95%CI [0.033, 0.060], SRMR = 0.024, CFI = 0.980, TLI = 0.970, AIC = 24606.824, and BIC = 24714.523. Loadings on the 7-item factor ranged from 0.565 to 0.672, with a mean of 0.629 (*SD* = 0.033). Cronbach's alpha was 0.821.

Social efficacy with peers. A CFA model for Social Efficacy with Peers showed good fit: scaled Chi Square (2) = 7.864, $p = 0.0196$, RMSEA = 0.048 with 95%CI [0.017, 0.086], SRMR = 0.015, CFI = 0.994, and TLI = 0.982. An examination of modification indices also did not reveal substantial sources of the misfit. Loadings on the 4-item factor ranged from 0.458 to 0.760 with a mean of 0.655 (*SD* = 0.135). Cronbach's alpha was 0.735.

Preference for group work. A CFA model for Preference for Group Work did not show good fit: scaled Chi Square (14) = 162.257, $p < 0.0001$, RMSEA = 0.092 with 95%CI [0.080, 0.105], SRMR = 0.034, CFI = 0.963, TLI = 0.945, AIC = 23501.098, and BIC = 23608.712. An examination of modification indices revealed that errors of the following two items tended to correlate: "I like to interact with others when working on projects" and "I personally enjoy working with others." This result may be because the two items are more similar to each other (both are about feelings toward group work) than to other items in the scale. Thus, I removed the first item because it was more specific than other items (referred to work on projects than work in general) and had a standardized loading lower than the second item. Further, there was a high modification index for the two recoded items. As the items were the only recoded items in the scale, the modification index may reflect reverse-coding. Thus, I did not remove any of these two items. The fit of the final model was adequate: scaled Chi Square (9) = 80.882, p < 0.0001, RMSEA = 0.080 with 95%CI [0.065, 0.097], SRMR = 0.031, CFI = 0.980, TLI = 0.967, AIC = 20411.718, and BIC = 20503.959. Loadings on the 6-item factor ranged from 0.648 to 0.912 with a mean of 0.795 (*SD* = 0.109). Cronbach's alpha was 0.913.

Intellect. A CFA model for Intellect did not show good fit: scaled Chi Square (35) = 361.909, p < 0.0001, RMSEA = 0.087 with 95%CI [0.079, 0.095], SRMR = 0.049, CFI = 0.906, TLI = 0.880, AIC = 35678.783, and BIC = 35832.517. An examination of standardized loadings revealed that one item ("I avoid philosophical discussions") had a particularly low loading. Thus, I removed this item. An examination of modification indices did not reveal extremely high values. The fit of the final model was adequate: scaled Chi Square (27) = 271.420, p < 0.0001, RMSEA = 0.085 with 95%CI [0.076, 0.095], SRMR = 0.042, CFI = 0.926, TLI = 0.901, AIC = 31403.823, and BIC = 31542.184. Loadings on the 9-item factor ranged from 0.377 to 0.805, with a mean of 0.626 (*SD* = 0.152). Cronbach's alpha was 0.849.

Public speaking anxiety. A CFA model for Public Speaking Anxiety did not show good fit: scaled Chi Square (5) = 120.400, p < 0.0001, RMSEA = 0.136 with

95%CI [0.116, 0.158], SRMR = 0.036, CFI = 0.956, TLI = 0.913, AIC = 18806.256, and BIC = 18883.123. An examination of modification indices revealed that errors of the following two items tended to correlate: “I breathe faster just before I need to speak in front of the whole class” and “My heart beats very fast while I am speaking in front of the whole class.” This result may be because the two items are more conceptually similar (focus on physiological reactions) than other items in the scale. I removed the first item because it had a lower standardized loading than the second item. The fit of the final model was adequate: scaled Chi Square (2) = 36.303, p < 0.0001, RMSEA = 0.118 with 95%CI [0.086, 0.152], SRMR = 0.016, CFI = 0.984, TLI = 0.952, AIC = 15051.946, and BIC = 15113.430. A further examination of modification indices did not reveal extremely high values. Loadings on the 4-item factor ranged from 0.810 to 0.876, with a mean of 0.850 (*SD* = 0.031). Cronbach’s alpha was 0.911.

Characteristics of external variables. Composite scores for external constructs that were measured via multi-item scales were developed by averaging their corresponding items. Investigating potential outliers for Actual Grade in Percent, I found that four students had very low grades (below 30%) while other students had grades above 40%. Thus, I replaced the grades for four students with missing data. Further, one student had a grade of 111% while other students had grades below 104%. Thus, this grade was also recoded as missing. Descriptive statistics for all external variables are presented in Table 31. All but three variables were positively skewed. Three variables – Public Speaking Anxiety, Perceived Potential Learning, and Perceived Prior Knowledge – had small positive skewness. Variables also differed in kurtosis (i.e., peakedness).

Among the variables with a sharper peak (i.e., leptokurtic items), the most peaked was Actual Grade in Percent (kurtosis = 0.70). Among the flat (i.e., platykurtic) variables, the flattest was Perceived Potential Learning (kurtosis = -0.99). Thus, deviations from normality were not substantial. No variables appeared to have a restricted range. In sum, the external variables showed acceptable characteristics.

Table 31. Descriptive statistics for additional constructs

Construct	N	Mean	*SD*	Min	Max	Skewness	Kurtosis
Effort	1246	4.62	1.08	1.00	6.00	-0.70	0.21
Interest-Feeling	1247	4.00	1.27	1.00	6.00	-0.44	-0.41
Interest-Value	1248	4.31	1.11	1.00	6.00	-0.62	0.06
Persistence	1247	4.61	0.87	1.00	6.00	-0.52	0.27
Metacognitive Strategies	1247	4.33	0.79	1.14	6.00	-0.41	0.50
Social Efficacy with Peers	1248	4.75	0.78	1.25	6.00	-0.64	0.64
Preference for Group Work	1242	3.80	1.13	1.00	6.00	-0.33	-0.31
Intellect	1242	4.11	0.80	1.00	6.00	-0.21	0.16
Public Speaking Anxiety	1241	3.44	1.38	1.00	6.00	0.06	-0.92
Expected Grade	1214	4.18	0.76	1.00	5.00	-0.58	-0.14
Actual Letter Grade	898	7.50	2.60	1.00	11.00	-0.64	-0.49
Actual Grade in Percent	817	83.19	10.26	41.37	103.80	-0.77	0.70
Perceived Learning	1242	3.88	0.86	1.00	5.00	-0.48	-0.08
Perceived Potential Learning	1236	1.93	0.71	1.00	3.00	0.11	-0.99
Perceived Prior Knowledge	1238	2.71	1.04	1.00	5.00	0.27	-0.31

Correlational Analyses

Correlational analysis was conducted between engagement composites and composites of the following external variables: effort, persistence, feelings-related and value-related components of interest, metacognitive strategies, intellect, social efficacy with peers, preference for group work, and public speaking anxiety. Prior to conducting correlational analyses, I examined relationships between engagement composites and these external variables. Most bivariate scatterplots did not show evidence of non-linear relationships. However, for some of the more "difficult" engagement composites (e.g., active behavioral engagement in Group Work), the relationships showed non-linearity due to the differences in "difficulty" between variables. Thus, the results of the correlational analysis should be approaches with caution.

Results for correlational analyses are presented in Table 32. As expected, correlations with effort, persistence, feelings-related and value-related components of interest, metacognitive strategies, and intellect were positive in sign and low-to-moderate in magnitude. Further, as expected, effort and persistence were correlated more strongly with behavioral/cognitive subscale and dimension composites than with emotional subscale and dimension composites. Effort had the highest correlation with behavioral/cognitive engagement in Lecture, whereas persistence had the highest correlations with behavioral/cognitive engagement in Lecture and in Individual Work. Next, also as expected, the feeling and value components of interest were correlated more strongly with emotional subscale and dimension composites than with behavioral/cognitive subscale and dimension composites. The highest correlations were observed for emotional engagement in Lecture. Also as expected, metacognitive strategies were correlated higher with behavioral/cognitive subscale and dimension composites than with emotional subscale and dimension composites. Metacognitive strategies had the highest correlation with behavioral/cognitive engagement in Individual Work. This result is logical since the measure of metacognitive strategies focuses on the work with assignments and other individual studying. Intellect was generally weakly related to engagement, with the exception of a moderate correlation with engagement in Individual Work. This result also seems to be logical because it is possible that for students with higher levels of intellect, engagement in Individual Work may be easier, leading to higher levels of engagement.

Further, as expected, social efficacy with peers and preference for group work were most strongly correlated with Group Work subscale and instruction type composites. Although, contrary to expectations, social efficacy with peers was significantly (positively) related to behavioral/cognitive and emotional engagement in other types of instructions, the correlations were rather weak. Preference for group work was, as expected, negatively related to engagement in Individual Work, particularly with Emotional Engagement in Individual Work. It also did not seem to be correlated with behavioral/cognitive engagement in Individual Work. Also as expected, preference for group work was not correlated significantly with Lecture subscale and instruction type composites. Contrary to expectations, it was significantly (positively) correlated with Whole-Class Interaction subscale and instruction type composites; however, these correlations were weak. Finally, as expected, public speaking anxiety was negatively related to active behavioral engagement in Whole-Class Interaction. Public speaking anxiety was not significantly correlated with behavioral/cognitive engagement in Lecture, Whole-Class Interaction (passive), and Individual Work, and minimally (negatively) correlated with behavioral/cognitive engagement in Group Work, as these engagement types do not require speaking in front of the whole class. Contrary to expectations, public speaking anxiety was negatively related to emotional subscale and dimension composites; yet, these correlations were rather weak. This result may be explained by overall class-level anxiety, where a student may fear being asked to speak in front of the whole class in any instruction type.

Table 32. Correlations between engagement composites and external composites

Engagement Composites	Effort	Persistence	Interest-Feeling	Interest-Value	Metacognitive Strategies	Intellect	Social Efficacy with Peers	Preference for Group Work	Public Speaking Anxiety
Subscale composites:									
LBC	**0.450**	**0.459**	**0.240**	**0.305**	**0.517**	**0.174**	**0.242**	0.049	-0.044
LE	**0.167**	**0.280**	**0.543**	**0.441**	**0.305**	**0.216**	**0.163**	-0.027	**-0.157**
WBC (passive)	**0.380**	**0.369**	**0.238**	**0.278**	**0.473**	**0.090**	**0.279**	**0.144**	-0.047
WB (active)	**0.136**	**0.195**	**0.230**	**0.213**	**0.254**	**0.219**	**0.302**	**0.141**	**-0.346**
WE	**0.218**	**0.259**	**0.468**	**0.391**	**0.333**	**0.188**	**0.257**	**0.079**	**-0.141**
IBC	**0.366**	**0.465**	**0.292**	**0.345**	**0.580**	**0.248**	**0.321**	**0.085**	-0.056
IE	0.074	**0.268**	**0.453**	**0.414**	**0.314**	**0.376**	**0.111**	**-0.259**	**-0.209**
GBC	**0.280**	**0.279**	**0.123**	**0.137**	**0.336**	**0.095**	**0.482**	**0.365**	**-0.084**
GE	**0.172**	**0.247**	**0.224**	**0.217**	**0.284**	**0.174**	**0.556**	**0.399**	**-0.222**
Instruction type composites:									
L	**0.358**	**0.439**	**0.495**	**0.462**	**0.487**	**0.240**	**0.241**	0.008	**-0.130**
W	**0.316**	**0.365**	**0.474**	**0.429**	**0.470**	**0.244**	**0.383**	**0.157**	**-0.257**
I	**0.256**	**0.448**	**0.483**	**0.485**	**0.544**	**0.404**	**0.256**	**-0.137**	**-0.179**
G	**0.258**	**0.303**	**0.204**	**0.207**	**0.357**	**0.157**	**0.601**	**0.442**	**-0.181**
Dimension									

composites:									
BC	**0.382**	**0.414**	**0.236**	**0.257**	**0.488**	**0.211**	**0.406**	**0.188**	**-0.170**
E	**0.193**	**0.293**	**0.500**	**0.418**	**0.356**	**0.257**	**0.351**	**0.122**	**-0.200**
Global composite	**0.348**	**0.434**	**0.472**	**0.429**	**0.519**	**0.293**	**0.468**	**0.190**	**-0.232**

Note. LBC = Behavioral/Cognitive Engagement in Lecture; LE = Emotional Engagement in Lecture; WBC (passive) = Behavioral (passive)/Cognitive Engagement in Whole-Class Interaction; WB (active) = Active Behavioral Engagement in Whole-Class Interaction; WE = Emotional Engagement in Whole-Class Interaction; IBC = Behavioral/Cognitive Engagement in Individual Work; IE = Emotional Engagement in Individual Work; GBC = Behavioral/Cognitive Engagement in Group Work; GE = Emotional Engagement in Group Work; L = Engagement in Lecture; W = Engagement in Whole-Class Interaction; I = Engagement in Individual Work; G = Engagement in Group Work; BC = Behavioral/Cognitive Engagement; E = Emotional Engagement; Global = Global Engagement. Statistically significant correlations ($p < 0.01$) are in bold.

In sum, I found that correlations between engagement composites and external composites ranged from statistically non-significant to moderate in magnitude, providing support for discriminant validity. The majority of correlations appeared as expected. Most importantly, effort and persistence (indicators of behavioral engagement in prior research), as well as metacognitive strategies (an indicator of cognitive engagement in prior research) had low-to-moderate positive correlations with behavioral/cognitive subscale and dimension composites. These correlations suggest that these engagement subscales and dimensions are distinct from effort, persistence, and metacognitive strategies. Further, feeling and value components of interest (indicators of emotional engagement in prior research) had low-to-moderate correlations with emotional subscale and dimension composites. These correlations suggest that these engagement subscales and dimensions are distinct from feeling and value components of interest. Next, intellect also showed low-to-moderate correlations, providing further evidence for discriminant validity. Additionally, correlations with social efficacy with peers, preference for group work, and public speaking anxiety provided discriminant evidence for engagement in different instruction types, particularly the Group Work, Individual Work, and active Whole-Class Interaction subscales.

Regression Analyses

Regression analysis was conducted with four achievement outcomes: actual grades in percent, actual letter grades, expected grades, and perceived learning. For each outcome, four regressions were run with different sets of predictors: (1) the subscale composites, (2) the instruction type composites, (3) the dimension composites, and (4) the global engagement composite. Regressions with perceived learning as an outcome also controlled for the perceived potential learning and the perceived prior knowledge. When perceived potential learning and perceived prior knowledge were regressed on perceived (actual) learning, the results showed that both predictors had significant negative effects. In particular, standardized coefficients for perceived potential learning and perceived prior knowledge were -0.171 (*SE* = 0.040) and -0.125 (*SE* = 0.056), respectively. Additionally, all regressions with subscale composite scores or instruction types composite scores controlled for the amount of time spent on each instruction type. The time when a student decided not to work on a task was used as a reference instruction type. Regressions with dimension composite scores or global composite scores controlled only for the time when a student decided not to work on a task because other times were incorporated in the engagement composites through weighting.

The effects of the time spent on different class parts are presented in Table 33. Compared to the effect of the time spent on Lecture, the effect of deciding not to work on a task was lower for actual grades in percent and for perceived learning but approximately the same for actual letter grades and expected grades. The effect of the time spent on Whole-Class Interaction was higher for perceived learning but approximately the same for actual grades and expected grades. The effect of the time spent on Individual Work was lower for actual grades in percent but approximately the same for actual letter grades, expected grades, and perceived learning. The time spent on Group Work did not seem to be related to achievement more or less than the time spent on Lecture.

Table 33. Results for the regression analysis with achievement as an outcome and types of instruction as predictors

	Actual Grades in Percent	Actual Letter Grades	Expected Grades	Perceived Learning*
Whole-Class Interaction	-0.068 (0.038)	0.016 (0.048)	0.063 (0.035)	**0.139 (0.044)**
Individual Work	**-0.062 (0.031)**	-0.054 (0.050)	-0.041 (0.036)	0.045 (0.033)
Group Work	-0.100 (0.070)	-0.063 (0.050)	0.035 (0.059)	-0.030 (0.056)
Deciding not to work on a task	**-0.121 (0.050)**	-0.053 (0.040)	-0.031 (0.040)	**-0.121 (0.035)**

Note. Lecture is the reference instruction type. Standardized regression coefficients are presented. Standard errors are in parentheses. * Controlled for Perceived Potential Learning and Perceived Prior Knowledge.

Prior to conducting regression analyses, I examined relationships between engagement composites and achievement variables for linearity. Most bivariate scatterplots did not show evidence of non-linear relationships. For some of the more “difficult” engagement composites (e.g., active behavioral engagement in Group Work), some of the relationships showed non-linearity due to the differences in “difficulty” between variables. However, the deviations from non-linearity were not substantial. Finally, I evaluated the assumption of homoscedasticity by examining scatterplots of residuals plotted against predicted values and each predictor. No particularly anomalous patterns were detected.

Zero-order correlations are presented in Table 34. All engagement composites were consistently positively related to perceived learning, with correlations ranging from low to moderate in magnitude. For other achievement variables, correlations were positive and low in magnitude or statistically non-significant. Correlations with expected grades tended to be somewhat higher than actual grades (in percent or letter) but lower than perceived learning. Among subscale composites, the highest correlations tended to be for emotional engagement in Lecture, Whole-Class Interaction, and Individual Work. Behavioral/cognitive engagement in all types of instruction, as well as Emotional engagement in Whole-Class Interaction and Group Work, tended to be correlated with actual or expected grades weakly or statistically non-significantly. Among instruction type composites, correlations with perceived learning were stronger for engagement in Lecture, Whole-Class Interaction, and Individual Work, compared to engagement in Group Work. Engagement in Group Work was also not correlated significantly with actual or expected grades, and engagement in Lecture was not correlated significantly with actual grades. Among dimension composites, correlations with perceived learning were stronger for emotional engagement than for behavioral/cognitive engagement. Additionally, behavioral/cognitive engagement was not correlated significantly with other achievement variables. Finally, global engagement was positively related to all achievement variables except actual letter grades, with the strongest correlation being for perceived learning.

Table 34. Correlations between engagement composites and achievement variables

	Actual Grade in Percent	Actual Letter Grade	Expected Grade	Perceived Learning
Subscale composites:				
LBC	-0.052	-0.033	0.024	**0.346**
LE	**0.103**	**0.102**	**0.278**	**0.441**
WBC (passive)	-0.014	-0.005	0.037	**0.303**
WB (active)	**0.151**	**0.112**	**0.167**	**0.162**
WE	0.088	0.061	**0.214**	**0.401**
IBC	-0.013	-0.013	**0.084**	**0.304**
IE	**0.229**	**0.220**	**0.338**	**0.346**
GBC	-0.051	-0.035	-0.009	**0.144**
GE	0.058	0.056	**0.110**	**0.226**
Instruction type composites:				
L	0.035	0.048	**0.200**	**0.485**
W	**0.116**	0.085	**0.221**	**0.421**
I	**0.157**	**0.150**	**0.285**	**0.414**
G	0.007	0.015	0.062	**0.216**
Dimension composites:				
BC	0.019	-0.016	0.047	**0.258**
E	**0.124**	**0.119**	**0.272**	**0.434**
Global composite	**0.091**	0.068	**0.210**	**0.441**

Note. LBC = Behavioral/Cognitive Engagement in Lecture; LE = Emotional Engagement in Lecture; WBC (passive) = Behavioral (passive)/Cognitive Engagement in Whole-Class Interaction; WB (active) = Active Behavioral Engagement in Whole-Class Interaction; WE = Emotional Engagement in Whole-Class Interaction; IBC = Behavioral/Cognitive Engagement in Individual Work; IE = Emotional Engagement in Individual Work; GBC = Behavioral/Cognitive Engagement in Group Work; GE = Emotional Engagement in Group Work; L = Engagement in Lecture; W = Engagement in Whole-Class Interaction; I = Engagement in Individual Work; G = Engagement in Group Work; BC = Behavioral/Cognitive Engagement; E = Emotional Engagement; Global = Global Engagement. Statistically significant correlations ($p < 0.01$) are in bold.

The results of the regression analyses are presented in Table 35. Subscale composites varied in their effects on outcomes. Most effects were positive or statistically non-significant, although negative effects were also observed. Yet, statistically significant effects were low in magnitude, with no standardized effect being greater than 0.3. In particular, when actual grades (in percent or letter) were used as an outcome, Active Behavioral Engagement in Whole-Class Interaction and Emotional Engagement in Individual Work were positive predictors. When actual grades in letter but not in percent were used as an outcome, Behavioral/Cognitive Engagement in Group Work was a negative predictor. When expected grades were used as an outcome, statistically significant predictors were Emotional Engagement in Lecture, active behavioral engagement in Whole-Class Interaction, and Emotional Engagement in Individual Work. All three predictors were positive. When perceived learning was used as an outcome, in addition to these three predictors, three more predictors appeared to be statistically significant (and positive): Behavioral/Cognitive Engagement in Lecture, Behavioral/Cognitive Engagement (passive) in Whole-Class Interaction, and Behavioral/Cognitive Engagement in Individual Work. It should be noted that Emotional Engagement in Lecture and Emotional Engagement in Whole-Class Interaction were both moderately correlated with perceived learning (r = 0.441 and r = 0.401, respectively). However, due to the relatively high correlation between the two composites (r = 0.738), Emotional Engagement in Whole-Class Interaction was no longer significantly related to perceived learning, when Emotional Engagement in Lecture was also accounted for. Overall, Active Behavioral Engagement in Whole-Class Interaction was the strongest predictor of actual grades, Emotional Engagement in Individual Work was the strongest predictor of expected grades, and Emotional Engagement in Lecture was the strongest predictor of perceived learning.

Looking across outcomes, one can see that two subscale composites – Active Behavioral Engagement in Whole-Class Interaction and Emotional Engagement in Individual Work – positively predicted all achievement variables. Active Behavioral Engagement in Whole-Class Interaction had the strongest relationship with actual grades in percent, whereas Emotional Engagement in Individual Work had the strongest relationship with expected grades. Emotional Engagement in Lecture was a positive predictor of two outcomes: expected grades and perceived learning. The effect of Emotional Engagement in Lecture was stronger for perceived learning than for expected grades. Three subscale composites – Behavioral/Cognitive Engagement in Lecture, Behavioral/Cognitive (passive) Engagement in Whole-Class Interaction, and Behavioral/Cognitive Engagement in Individual Work – were statistically significant (and positive) predictors only of perceived learning. Further, two subscale composites – Emotional Engagement in Whole-Class Interaction and Emotional Engagement in Group Work – were not statistically significant predictors of any achievement variables. Finally, Behavioral/Cognitive Engagement in Group Work was a negative predictor of actual letter grades. It should be noted that Behavioral/Cognitive Engagement in Group Work was not significantly correlated with actual letter grades but became a negative predictor once other subscale composites were accounted for.

Similarly to subscale composites, instruction type composites varied in their effects on outcomes. Most effects were positive or statistically non-significant, although negative effects were also observed. Yet, statistically significant effects were low in magnitude, with no standardized effect being greater than 0.3. In particular, when actual grades in percent were used as an outcome, engagement in Whole-Class Interaction and Individual Work were positive predictors of actual grades in percent, and engagement in Group Work was a negative predictor of actual grades in percent. When actual letter grades were used as an outcome, only engagement in Whole-Class Interaction and Individual Work were statistically significant (and positive) predictors. With expected grades as outcomes, engagement in Whole-Class Interaction and engagement in Individual Work were positive predictors, and engagement in Group Work was a negative predictor. However, with perceived learning as an outcome, three composites were statistically significant (and positive): engagement in Lecture, Whole-Class Interaction, and Individual Work. Overall, engagement in Whole-Class Interaction was the strongest predictor of actual grades in percent, engagement in Individual Work was the strongest predictor of expected grades, and engagement in Lecture was the strongest predictor of perceived learning. Engagement in Whole-Class Interaction and Individual Work predicted actual grades in percent in a similar way.

Looking across outcomes, one can see that Engagement in Whole-Class Interaction and Individual Work were statistically significant (and positive) predictors of all achievement variables. Engagement in Whole-Class Interaction had the strongest relationship with actual grades in percent and the weakest relationship with actual letter grades. Engagement in Individual Work had the strongest relationship with expected grades and the weakest relationship with actual letter grades. Engagement in Group Work was related negatively to two achievement variables: actual grades in percent and expected grades. The magnitude of the relationships was approximately the same. Again, it should be noted that Behavioral/Cognitive Engagement in Group Work was not significantly correlated with actual grades in percent and expected grades but became a negative predictor once other instruction type composites were accounted for. Finally, engagement in Lecture was a significant (and positive) predictor only of perceived learning.

Dimension composites had positive or statistically non-significant effects on outcomes. Yet, positive effects were relatively weak, with no standardized effect being greater than 0.4. In particular, both behavioral/cognitive and emotional engagement positively predicted perceived learning. However, only emotional engagement was a statistically significant (and positive) predictor of actual and expected grades. Emotional engagement had the strongest relationship with perceived learning and the weakest relationships with actual grades. For all outcomes, emotional engagement was the strongest predictor. Finally, global engagement was a statistically significant positive predictor of all achievement variables except actual letter grades. Global engagement had the strongest effect on perceived learning (0.433).

Table 35. Results for regression analysis with achievement as an outcome and engagement composites as predictors

Engagement Composites	Actual Grade in Percent	Actual Letter Grade	Expected Grade	Perceived Learning**
Subscale composites*:				
LBC	-0.126 (0.083)	-0.082 (0.075)	-0.096 (0.058)	**0.106 (0.046)**
LE	0.057 (0.066)	0.060 (0.044)	**0.158 (0.047)**	**0.210 (0.050)**
WBC (passive)	0.048 (0.061)	0.043 (0.066)	0.004 (0.062)	**0.067 (0.033)**
WB (active)	**0.183 (0.047)**	**0.159 (0.056)**	**0.151 (0.034)**	**0.088 (0.032)**
WE	-0.012 (0.083)	-0.058 (0.062)	0.018 (0.056)	0.054 (0.044)
IBC	-0.029 (0.072)	-0.012 (0.061)	0.034 (0.051)	**0.087 (0.039)**
IE	**0.165 (0.051)**	**0.149 (0.046)**	**0.227 (0.056)**	**0.120 (0.041)**
GBC	-0.079 (0.048)	**-0.100 (0.045)**	-0.068 (0.043)	-0.027 (0.047)
GE	0.007 (0.048)	0.044 (0.061)	0.014 (0.036)	0.069 (0.049)
Instruction type composites*:				
L	-0.130 (0.075)	-0.086 (0.070)	-0.006 (0.057)	**0.248 (0.045)**
W	**0.211 (0.083)**	**0.145 (0.067)**	**0.192 (0.058)**	**0.161 (0.041)**
I	**0.156 (0.078)**	**0.145 (0.062)**	**0.246 (0.050)**	**0.167 (0.043)**
G	**-0.139 (0.042)**	-0.104 (0.055)	**-0.128 (0.043)**	0.028 (0.040)
Dimension composites:				
BC	-0.033 (0.056)	-0.053 (0.039)	-0.031 (0.032)	**0.149 (0.032)**
E	**0.124 (0.052)**	**0.140 (0.040)**	**0.298 (0.035)**	**0.387 (0.036)**
Global composite	**0.077 (0.039)**	0.074 (0.038)	**0.224 (0.032)**	**0.433 (0.029)**

Note. Standardized regression coefficients are presented. Standard errors are in parentheses. * Controlled for the amount of time spent on the four types of instruction. ** Controlled for Perceived Potential Learning and Perceived Prior Knowledge. LBC = Behavioral/Cognitive Engagement in Lecture; LE = Emotional Engagement in Lecture; WBC (passive) = Behavioral (passive)/Cognitive Engagement in Whole-Class Interaction; WB (active) = Active Behavioral Engagement in Whole-Class Interaction; WE = Emotional Engagement in Whole-Class Interaction; IBC = Behavioral/Cognitive Engagement in Individual Work; IE = Emotional Engagement in Individual Work; GBC = Behavioral/Cognitive Engagement in Group Work; GE = Emotional Engagement in Group Work; L = Engagement in Lecture; W = Engagement in Whole-Class Interaction; I = Engagement in Individual Work; G = Engagement in Group Work; BC = Behavioral/Cognitive Engagement; E = Emotional Engagement. Statistically significant correlations (p < 0.05) are in bold.

In sum, I found that, overall, student engagement was a significant predictor of achievement, as indicated by most regressions with global engagement as a predictor. However, breaking engagement down by dimensions, instruction types, or subscales showed that not all aspects of engagement predict achievement in the same way. In particular, regression analyses with dimension, instruction type, or subscale engagement demonstrated which aspects may be beneficial for achievement, which aspects do not appear to affect achievement, and which aspects may be even detrimental for achievement. The analysis also showed that the effects of engagement on achievement differed between measures of achievement. Thus, examining different aspects of engagement with different achievement outcomes provided an opportunity to identify specific engagement aspects that have the potential to be the most beneficial. For example, for actual grades in percent, I found that among dimensions, the emotional dimension seemed to be the most beneficial, whereas among instruction types, engagement in Whole-Class Interaction seemed to be the most beneficial. Breaking engagement down even more and examining engagement subscales that incorporate dimensions and instruction types simultaneously, I found that specific aspects of engagement, particularly beneficial for actual grades in percent, are Active Engagement in Whole-Class Interaction and Emotional Engagement in Individual Work. Thus, dimension engagement and instruction type engagement provide more information than global engagement, and subscale engagement provides more information than dimension engagement and instruction type engagement.

Conclusion. Overall, with correlational and regression analyses, I examined the assumption within the external aspect of Messick's model. The assumption states that expected relationships with relevant constructs need to be demonstrated. In particular, I examined discriminant and predictive validity. I found evidence for discriminant validity with all external variables under investigation: effort, persistence, metacognitive strategies, feeling and value components of interest, intellect, social efficacy with peers, preference for group work, and public speaking anxiety. Further, I found that, overall, student engagement positively predicts student achievement. However, there was variability in sizes and directions of effects within dimensions, instruction types, and subscales. Thus, dimension engagement and instruction type engagement may provide more information about effects of engagement on achievement than global engagement, and subscale engagement may provide more information than dimension engagement and instruction type engagement. These results support multidimensional, instruction specific measurement of student engagement.

Internal Consistency

One of the assumptions within the generalizability aspect of Messick's model is the assumption of the internal consistency of subscales. This assumption states that subscales need to exhibit adequate internal consistency. Initially, I planned to examine internal consistency via Cronbach's alpha. However, this assumption was developed under the premise that engagement was a reflective construct. As formative indicators are not expected to have positive correlations, they are not expected to be internally consistent (Bollen & Lennox, 1991). Here, I present results of the internal consistency analysis for completeness of the discussion (see Table 36). Overall, internal consistency of subscales was adequate. Subscales with more items showed higher internal consistency. Such internal consistency is expected since most items within subscales were positively correlated, albeit not strongly. In sum, I provided evidence for internal consistency of subscales. However, internal consistency is not necessary for formative constructs and, therefore, should not be used as evidence for or against subscale quality.

Table 36. Internal consistency of engagement subscales

Engagement Composite	Number of items	Cronbach's alpha
LBC	9	0.81
LE	4	0.74
WBC (passive)	8	0.84
WB (active)	3	0.87
WE	4	0.69
IBC	11	0.84
IE	4	0.69
GBC	11	0.89
GE	4	0.71

Note. LBC = Behavioral/Cognitive Engagement in Lecture; LE = Emotional Engagement in Lecture; WBC (passive) = Behavioral (passive)/Cognitive Engagement in Whole-Class Interaction; WB (active) = Active Behavioral Engagement in Whole-Class Interaction; WE = Emotional Engagement in Whole-Class Interaction; IBC = Behavioral/Cognitive Engagement in Individual Work; IE = Emotional Engagement in Individual Work; GBC = Behavioral/Cognitive Engagement in Group Work; GE = Emotional Engagement in Group Work.

Chapter Five

The goal of the present study was to develop and provide initial validity evidence for the instrument that measures multidimensional, instruction-specific student engagement in undergraduate mathematics-based classes. I aimed the instrument to measure three engagement dimensions (behavioral, cognitive, and emotional) in four types of instruction (Lecture, Whole-Class Interaction, Individual Work, and Group Work). In developing and validating instrument scores, I employed the argument-based approach to validity (Kane, 2013; 2015). To inform my selection of the assumptions, required for the score inferences, as well as to inform my selection of the means to evaluate these assumptions, I used two general methodological frameworks for instrument development and validation: the model of Gehlbach and Brinkworth (2011) and the unified construct-based model of validity (Messick, 1995). In this chapter, I describe how the interpretation/use and validity arguments for instrument scores changed as a result of this study. Further, I discuss two major contributions of this study. Specifically, first, the study demonstrated the potential for combining multidimensionality and instructional specificity in engagement measurement. Second, the study results suggested that student engagement may be conceptualized as a formative construct rather than as a reflective construct. Finally, I discuss study limitations and directions for future research.

Revised Interpretation/Use and Validity Arguments

Score interpretations and uses. In the interpretation/use argument, I stated that there are 12 subscale composite scores, designed to be interpreted as levels of each engagement dimension (behavioral, cognitive, and emotional) in each type of instruction (lecture, whole-class interaction, individual work, and group work) in a particular undergraduate mathematics-based class. However, the results of the study showed that among items designed to indicate behavioral engagement in whole-class interaction, active behaviors separated out and indicated active behavioral engagement in whole-class interaction. Further, the results showed that, except for active behavioral engagement in whole-class interaction, behavioral and cognitive engagement dimensions within each instruction type were too similar and could not be differentiated empirically. Thus, the number of subscales has changed to nine. Composite scores on the four behavioral/cognitive or the four emotional instruction-specific subscales should be interpreted as levels of behavioral/cognitive or emotional engagement in each type of instruction in a particular undergraduate mathematics-based class. Behaviors in the behavioral/cognitive engagement in whole-class interaction subscale were only passive. Composite scores on the active behavioral engagement in whole-class interaction subscale should be interpreted as the levels of active behavioral engagement in whole-class interaction. Next, instead of three dimension composite scores (behavioral, cognitive, and emotional), there are two: behavioral/cognitive and emotional. They are designed to be interpreted as levels of behavioral/cognitive or emotional engagement in a particular undergraduate mathematics-based class. The interpretation of instruction type composite scores and global engagement composite scores have not changed. Four instruction type composite scores are designed to be interpreted as the levels of student engagement in four types of instruction (lecture, whole-class interaction, individual work, and group work) in a particular undergraduate mathematics-based class. Global engagement composite scores (i.e., composite scores on the overall instrument) are designed to be interpreted as the level of overall student engagement in a particular undergraduate mathematics-based class. For each of these composite scores, the term "level" refers to the frequency of student engagement over the course of the semester.

The intended uses of the scores, produced by the instrument, have not changed. First, educators can use the scores to identify which kind of engagement students lack. In particular, if educators find that students lack a particular dimension of engagement, engagement in a particular instruction type, or a particular dimension of engagement in a particular instruction type, then educators may focus their efforts on an instructional intervention that aims to increase engagement of this kind. Second, the scores can be used in research that aims to inform the development of engagement interventions. In particular, researchers can use the scores to identify instructional facilitators of engagement and develop engagement interventions that focus on these facilitators.

Theory-based inference for item scores. At the stage of working with items, the theory-based inference provides a link between items and constructs, where the constructs are subscale engagement, engagement dimensions, engagement in instruction types, or global engagement. The assumptions, required for the theory-based inference for item scores and stated in the interpretation/use argument, have not changed as a result of the study. Yet, the methods employed to empirically examine the assumptions and stated in the validity argument have been extended for one assumption – the assumption within the structural aspect of Messick's model. Overall, the methods provided evidence for the assumptions; however, the strength of the evidence differed.

The first assumption states that the items need to be relevant to and representative of the construct being measured (the content aspect). This assumption was examined via expert reviews. The second assumption states that observed response processes need to match the intended ones and be aligned with respondents' behaviors (the cognitive modeling component of the substantive aspect). This assumption was examined via cognitive interviews. Expert reviews and cognitive interviews provided some evidence to support these assumptions, as I identified and tried to address multiple problems with the instrument. Yet, I cannot say that all problems were eliminated.

The third assumption stated that item response characteristics also need to be consistent with expected characteristics (the scale functioning component of the substantive aspect). Initial evaluation of this assumption was conducted using responses to the items during cognitive interviews. Further evaluation of this assumption was conducted on the field-testing data. To examine this assumption, I explored item response characteristics via descriptive statistics (using both cognitive interview and field-testing data) and item correlations (using field-testing data). The analysis of descriptive statistics showed that some items were not normally distributed; the items also varied in their means. The analysis of item correlations showed that within-subscale item correlations were typically positive but not as high as desired. Yet, within-subscale item correlations were stronger than between-subscale item correlations. Thus, I provided some evidence for the scale functioning component of Messick's model.

The fourth assumption states that the internal structure of the instrument needs to be determined (the structural aspect). Initially, I planned to identify the internal structure via an ESEM analysis. However, conducting the ESEM analysis, I found that even the most interpretable ESEM model might not represent the internal structure of the instrument well. The most interpretable model was the 7-factor model with the following six substantive factors: four behavioral/cognitive engagement factors in each type of instruction (with behaviors in the behavioral/cognitive factor in whole-class interaction being passive), active behavioral engagement in whole-class interaction, and emotional engagement. One problem with the internal structure, suggested by this model, was the emergence of the "Difficulty" factor. Loadings on this factor came from cross-loadings of items that also loaded on their substantive factors. In particular, positive loadings on the "Difficulty" factor came from cross-loadings of "difficult" items (i.e., items with lower means and often larger variances). Typically, the more "difficult" the item was, the stronger the loading on the "Difficulty" factor was, and the more attenuated the loading on the item's substantive factor was. Smaller negative cross-loadings came from very "easy" items. Thus, the "difficult" items, i.e., the items that cross-loaded substantially (positively) on the "Difficulty" factor, are likely to be the items that have the most information about students' engagement and the most ability to discriminate between engagement levels of moderately or highly engaged students. A potential reason for the

“Difficulty” factor to emerge is an inappropriate application of linear methods. An application of nonlinear methods may be able to resolve the “Difficulty” factor.

Another problem with the internal structure is the lack of fit, particularly for emotional engagement items. These items, regardless of the instruction type, tended to load on a single factor. Yet, there were also a number of cross-loadings of emotional engagement items on behavioral/cognitive factors, mostly those in the corresponding instruction type. Thus, the lack of fit and cross-loadings of emotional engagement items may indicate that these items are not similar enough to indicate a common factor. Yet, different emotions may provide unique information about students’ emotional engagement. Therefore, including different emotions to indicate emotional engagement may provide more information about students’ levels of emotional engagement and may help better discriminate between students with different levels of emotional engagement.

Given that the ESEM model might not represent the internal structure of the instrument due to the problems discussed above, I reconsidered my approach to engagement measurement. A common practice in engagement measurement and more broadly in measurement of psychological constructs is to approach measurement as reflective. In this study, I suggest that engagement measurement may be better conceptualized as formative rather than reflective. In particular, student engagement may be an entity constructed by researchers rather than an internal quality of a student. Further, in terms of the direction of causality between specific ways of engagement and the construct of student engagement, it is more plausible that students are said to be engaged because they engage in particular ways, rather than that students engage in particular ways because they are engaged. Next, it is likely that specific ways of engagement, used to measure the construct of engagement, are not interchangeable. Unique aspects of different ways of engagement are informative for the construct rather than irrelevant to it. Similarly to reflective measurement, formative indicators of the same construct have conceptual unity. Finally, the problems with the ESEM analyses can be explained from the perspective of formative measurement. The problem of poor ESEM model fit and the presence of cross-loadings does not apply to formative measurement because indicators are not expected to fit together. The "Difficulty" factor is also not necessarily a problem in formative measurement. First, data-driven methods to developing the internal structure are not required in formative measurement. In this case, only constructs, specified by researchers, would be created. Second, in contrast to reflect measurement, conceptually similar indicators that are substantially different from each other are desired to achieve a full representation of the construct.

In deciding on the constructs and indicators (i.e., on the internal structure), I used theoretical conceptualizations from the construct conceptualizations and evidence from the ESEM analysis for guidance. The suggestions of the ESEM analysis for the development of formative engagement constructs appeared to be theoretically meaningful. Specifically, based on the ESEM results, one construct that clearly stood out was active behavioral engagement in whole-class interaction. Next, also based on the ESEM results, I decided to merge behavioral and cognitive indicators within each instruction type (with the exception of active behaviors in whole-class interaction), as these indicators tended to be conceptually too similar to indicate different constructs. Yet, the ESEM analysis showed that behavioral/cognitive engagement can be differentiated by instruction type. Thus, I developed separate constructs of behavioral/cognitive engagement in each instruction types. A final decision concerned emotional engagement items. In the ESEM analysis, they tended to load on a single factor; yet, they did not fit well and had substantial cross-loadings. Thus, based on these observations and the initial design of emotional engagement by instruction type, I developed four emotional engagement constructs within each instruction type. In total, nine subscale constructs were created. In sum, the ESEM analysis provided some evidence for the internal structure of the instrument. However, ESEM was not designed to identify internal structures of formative measures. For this reason, the internal structure developed for the instrument in this study should be interpreted as tentative and approached with caution.

Scoring inference. The scoring inference employs scoring rules to develop composite scores from item scores. The scoring rules have not changed, compared to the scoring rules specified in the interpretation/use argument. However, I considered an alternative scoring rule for subscale composites. The two assumptions required for the scoring inference and stated in the interpretation/use argument have not changed as a result of the study. One assumption states that the scoring rules need to be plausible. The methods employed to evaluate this assumption and stated in the validity argument have changed. Another assumption states that all information required for computing composite scores need to be available. This assumption was addressed as part of the validity argument and, therefore, will not be discussed further.

I originally planned and eventually adopted the approach of creating subscale composite scores via averaging items in these subscales (i.e., the not weighted approach). However, after observing that some items were more “difficult” than others, I questioned the appropriateness of having all items contribute to composite scores in the same way. Thus, I considered an alternative approach to creating composite scores, which would take item “difficulty” into account (the weighted approach). This approach gives larger weights to more “difficult” items. Both approaches were implemented and evaluated by comparing descriptive statistics and computing correlations of composite scores, created via each approach. I found that the two scoring approaches produced similar composite scores. As expected, the not weighted approach produced somewhat higher scores than the weighted approach. The weighted approach is also more appropriate from the theoretical point of view. Yet, as the differences between the two approaches were minimal and the not weighted approach is easier to implement, I selected the not weighted approach as a scoring method for subscale composites.

The assumption of the plausibility of this scoring method was initially planned to be evaluated via testing whether items within a subscale were parallel. The planned analysis included extending the ESEM model by constraining item loadings to be the same within subscales and item error variances to be the same within subscales. However, as the ESEM model was not accepted as a model representing the internal structure of the instrument and was not viewed as suitable for formative measurement more generally, testing parallel ESEM models did not seem appropriate. Instead, the evidence for the plausibility of the scoring rule was provided via comparisons with the weighted approach, for which there was a theoretical rationale. In sum, I provided some evidence for the plausibility of the scoring rule for subscale composites.

The scoring rules for dimension composite scores, instruction type composite scores, and global composite scores have not changed as a result of this study, with a small exception. Initially, I planned to use subscale composite scores to create dimension and instruction type composite scores. However, I separated active behavioral engagement from passive behavioral and cognitive engagement in whole-class interaction. Thus, to develop dimension composite scores, I created a single behavioral/cognitive engagement dimension in whole-class interaction. I refer to behavioral/cognitive and emotional dimensions in each instruction type as 2x4 composites. Thus, instead of subscale composites, 2x4 composites were used to develop dimension and instruction type composites. Besides this modification, the process remained as described in the interpretation/use argument. Engagement dimension composite scores were computed via summing the 2x4 composite scores across instruction types, weighted by the amount of instruction, and dividing the sums by the total amount of time spent on the four types of instruction. Instruction type composite scores were computed via averaging the 2x4 composite scores across engagement dimensions. Finally, global engagement composite scores were computed via averaging dimension engagement composite scores. The assumption of the plausibility of the scoring rules for dimension composite scores, instruction type composite scores, and global composite scores were addressed in the validity argument and, therefore, will not be discussed further.

Theory-based inference for composite scores. The theory-based inference for composite scores is an inference from composite scores to construct scores that represent levels of the constructs. The assumptions required for the theory-based inference for composite scores and stated in the interpretation/use argument have not changed as a result of the study. The methods employed to empirically examine the assumptions and stated in the validity argument also have not changed. Overall, the methods provided evidence for the assumptions; however, the strength of the evidence differed.

The first assumption states that characteristics of composite scores need to be as expected (the scale functioning component within the substantive aspect, applied to composite characteristics). This assumption was evaluated via an examination of descriptive statistics and correlations of composite scores. I found that distributions of composite scores were approximately normal. Correlations were typically low-to-moderate in magnitude, supporting discriminant validity between the constructs. However, for some constructs, correlations were somewhat higher than desired.

The second assumption states that expected relationships with relevant external constructs need to be demonstrated (the external aspect). Specifically, discriminant validity needed to be demonstrated with effort, persistence, feeling and value components of interest, metacognitive strategies, intellect, social efficacy with peers, preference for group work, and public speaking anxiety. To examine discriminant validity, I conducted a correlational analysis of these external constructs and engagement composite scores. Predictive validity needed to be demonstrated with course achievement, to indicate which four measures were used: actual grades in percent, actual letter grades, expected grades, and perceived learning. To examine predictive validity, I regressed each measure of course achievement on engagement composite scores (separately for each kind of composite scores: subscale composite scores, dimension composite scores, instruction type composite scores, and global composite scores).

In general, the relationships, found as a result of the correlational analysis, were consistent with the expected relationships. When relationships did not match expectations, the deviations were minor. Overall, correlations between engagement composites and external constructs ranged from statistically non-significant to moderate in magnitude, providing support for discriminant validity. In particular, correlations with effort, persistence, feeling and value components of interest, and metacognitive strategies provided discriminant validity for dimension composites and subscale composites of the same dimension. Correlations with social efficacy with peers, preference for group work, and public speaking anxiety provided discriminant validity for instruction type composites and subscale composites of the same instruction type. Next, based on prior research, I expected engagement composites to positively predict achievement or not to predict it. Most regression results matched this expectation, with the exception of behavioral/cognitive engagement in group work (and also overall engagement in group work), which had small negative effects on some achievement variables. Thus, overall, I found some support for the predictive validity of engagement composites.

It is important to note that there are differences in relationships between achievement with engagement subscale composites, dimension composites, and instruction type composites. Thus, dimension engagement and instruction type engagement may provide more information about relationships with achievement than global engagement, and subscale engagement may provide more information than dimension engagement and instruction type engagement. These results support the usefulness of measuring engagement by dimension and instruction type.

Generalization inference. The generalization inference provides expected composite scores over a universe of possible composite scores. The assumptions of score generalizability and reliability required for the generalization inference and stated in the interpretation/use argument have not changed as a result of this study. The methods employed to empirically examine these assumptions and stated in the validity argument also have not changed. Overall, the results provided initial evidence for the assumption of score generalizability. The assumption of score reliability and the assumption of generalizability of relationships between engagement and external variables were not evaluated in this study. Finally, the assumption of internal consistency was removed from the interpretation/use and validity arguments.

The assumption of score generalizability states that scores need to be generalizable and reliable regardless of the setting, group membership, form, or format. As an initial step in evaluating this assumption, I examined whether the field-testing sample included a variety of classes from different settings, all forms, all formats, and students from a variety of backgrounds. Frequencies of students across settings, forms, formats, and groups showed some support for this assumption. However, some settings, formats, and groups were represented by a low number of students in the field-testing sample. A more rigorous investigation of generalizability of scores (e.g., via testing for measurement invariance) is outside of the scope of the initial validation. Further, I pretested the paper-and-pencil and online formats via cognitive interviews to explore whether format-specific aspects of instrument administration could prevent generalizability across formats. I did not find evidence for such format-specific aspects.

The assumption of internal consistency states that subscales need to exhibit adequate internal consistency. However, this assumption is relevant to reflective constructs where items are expected to be positively correlated. It is not relevant for formative constructs where items are not expected to be positively correlated. Thus, I removed this assumption from the interpretation/use and validity arguments. Nevertheless, for the completeness of discussion, I presented Cronbach's alphas, used to evaluate subscale internal consistency. Cronbach's alphas showed that the internal consistency of the subscales was adequate. Yet, these finding should not be used as evidence for or against the quality of subscales.

Finally, evaluating the assumption of reliability of scores (e.g., via test-retest reliability) across settings, forms, formats, and groups is outside of the scope of initial validation; the assumption of the generalizability of external relationships (e.g., via differential prediction) are also outside of the scope of the initial validation.

Decision inference. The decision inference allows for the use of scores produced by the instrument to identify the kinds of engagement that students lack and implement instructional interventions that aim to increase these kinds of engagement. The assumptions required for the decision inference and stated in the interpretation/use argument have not changed as a result of the study. The first assumption for this inference states that students need to benefit from instructional interventions that aim to increase particular kinds of engagement. The second assumption states that the benefits need to substantially outweigh the negative consequences of the interventions. An evaluation of these assumptions is outside the scope of the initial validation. For this reason, the validity argument for the decision inference have not been developed yet.

Summary. In this study, I developed the interpretation/use and validity arguments for the scores produced by the instrument. As a result of the study, student engagement has been re-conceptualized from a reflective construct to a formative construct. Due to this re-conceptualization, some parts of the arguments have been changed. Overall, I provided initial evidence for the assumptions I aimed to evaluate during this study. Further validation, in terms of both providing stronger evidence for the evaluated assumptions and providing evidence for the assumptions that have not been evaluated yet, is needed. However, from the initial validity evidence, two main contributions of this research occurred. First, the results demonstrated the potential for combining multidimensionality and instructional specificity in engagement measurement. Second, the results suggested that student engagement may be conceptualized as a formative construct rather than as a reflective construct. I discuss these two contributions in more detail below.

Combining Multidimensionality and Instructional Specificity

Providing rationale for developing this instrument, I argued that combining multidimensionality and instructional specificity would allow educators to identify both how (the dimension) and where (the instruction type) students are not engaged. With such information, educators would be able to develop more targeted and, therefore, more efficient interventions, compared to what would have been possible with information only about engagement in instruction types or engagement dimensions. Yet, to my knowledge, multidimensionality and instructional specificity had not been applied together to engagement measurement prior to this study. Overall, in this study, I developed and provided initial validity evidence for an instrument that measures multidimensional, instruction-specific engagement in mathematics-based undergraduate classes. Thus, I demonstrated the potential for combining multidimensionality and instructional specificity in mathematics-based undergraduate classes.

Dimensionality. The instrument was designed to measure behavioral, cognitive, and emotional engagement dimensions. One argument in the rationale for measuring multiple dimensions of student engagement was the existence of qualitative differences between engagement dimensions. In this study, for each dimension, I aimed to design items that were qualitatively different from items that indicate other dimensions. Specifically, I conceptualized the three dimensions in the following way. Behavioral engagement referred to students' expected observable on-task behaviors, including both verbal and non-verbal. Cognitive engagement referred to students' expected cognitive processes of selecting, organizing, and integrating. Emotional engagement referred to students' positive activating, negative activating, positive deactivating, and negative deactivating emotions. While the three dimensions are conceptually distinct, developing items that would indicate only one dimension was particularly challenging for behavioral and cognitive dimensions. In expert reviews, experts often viewed some items, designed to be behavioral, as cognitive, and vice versa. The ESEM analysis showed that the two dimensions could not be differentiated empirically. As a result, I reduced the number of dimensions in the instrument from three (behavioral, cognitive, and emotional) to two (behavioral/cognitive and emotional) qualitatively different dimensions. Another argument in the rationale for measuring multiple dimensions of student engagement was the existence of quantitative differences between engagement dimensions. In this study, I found the correlation between behavioral/cognitive and emotional dimension composites to be low. This finding suggests that a student may be engaged behaviorally and cognitively but not emotionally, or vice versa. Thus, as hypothesized, for some students, there are quantitative differences in engagement levels across dimensions.

Similarly to my findings, emotional engagement has been typically found to be empirically distinct from behavioral and cognitive engagement in prior research. Specifically, in the studies where conceptualizations of engagement dimensions were similar to mine, researchers found that emotional engagement had low-to-moderate correlations with behavioral and cognitive engagement (e.g., Reeve, 2013; Whitney et al., 2019), although in some studies the correlations were moderate-to-high (e.g., Uzzaman & Karim, 2016; Z. Wang et al., 2014). As my finding of differentiation of emotional engagement from behavioral and cognitive is in line with findings from other studies, I will not discuss it further. Instead, I will focus the discussion on the similarity of behavioral and cognitive engagement, as this finding is different from the results of other studies.

Contrary to my study, prior research has typically found behavioral and cognitive engagement to be distinct from each other. Specifically, some researchers found that behavioral engagement had a low-to-moderate correlation with cognitive engagement (Awang Hashim & Murad Sani, 2008; Reeve, 2013; Whitney et al., 2019), although quite high correlations were also observed (Z. Wang et al., 2014). One reason for the empirical differentiation between behavioral and cognitive dimensions may lie in different conceptualizations and item referents. For example, in the study of Awang Hashim and Murad Sani (2008), behavioral engagement was conceptualized as school compliance (e.g., preparedness, tardiness, aggressive behavior, etc.), and cognitive engagement was conceptualized as metacognitive strategies used while studying for the class. In my study, both behaviors and cognitive processes were applied in the context of in-class learning. A more subtle difference in item referents can be seen in the study of Reeve (2013). Here, behavioral items had a referent of "in this class," where some items (e.g., the item about listening) likely referred to the time in a classroom. However, cognitive items had a referent of the time spent studying for a particular class (e.g., "when I study for this class") or similar. This referent may be interpreted as a broader (or different) referent to also (or only) include the study time outside of the classroom. In my study, behavioral and cognitive items referred to the same thing – a particular instruction type. Some other studies seemed to use similar (or mixed) referents across behavioral and cognitive dimensions. For example, in the study of Whitney et al. (2019), items referred to either in-class or out-of-class behaviors and cognitive processes, or potentially to both.

The reason why behavioral engagement was empirically distinguished in the study of Whitney et al. (2019) might not be in the intended distinction between behaviors and cognitive processes. Instead, it may be possible for the distinction to be due to the differences in item “difficulties,” as the ESEM results in my study suggested. Specifically, with the ESEM analysis, I found that “easy” behavioral and cognitive items seemed to be grouped together, while more “difficult” items seemed to cross-load on a different factor with loadings reflecting item “difficulty.” Further, whether the items were behaviors or cognitive processes did not seem to play a role, suggesting that behaviors and cognitive processes in a classroom are conceptually similar entities. In the literature, when referents, similar to mine (or at least mixed), were used, researchers typically found that average cognitive engagement tended to be lower than behavioral (e.g., Fredricks et al., 2003; Lam et al., 2014; Liu et al., 2018). This difference may suggest that it is more “difficult” for students to engage cognitively than behaviorally. This finding seems to be logical because cognitive processes may be considered as an adds-on to behaviors. In fact, one of the ways cognitive engagement has been defined in the literature is investment in learning that is “more than just behavioral engagement” (Fredricks et al., 2004, p. 64). Thus, there may be students who engage behaviorally but not cognitively, as well as students who engage both behaviorally and cognitively. However, it is not likely for students to be engaged cognitively but not behaviorally.

Although Whitney et al. (2019) did not report descriptive statistics, it may be plausible that behaviors, such as listening and/or reading carefully, taking good notes, reviewing notes, or talking to the teacher about own progress in the class, are “easier” to do than cognitive processes, such as trying to connect different topics from course material, summarizing the material learned in class or from other course materials, identifying key information from any reading assignment, or examining the strengths and weaknesses of own views on a topic or issue. Yet, these speculations alone cannot count as an argument against the distinction between the two dimensions based on their behavioral or cognitive nature. However, the cognitive scale of Whitney et al. (2019) seemed to include not only the aforementioned cognitive items but also two items that may be considered behavioral. In particular, these behavioral items within the cognitive dimension are the item about discussing course topics, ideas, or concepts with the teacher outside of class and the item about asking questions or contributing to course discussions in other ways. It may be plausible that these two items loaded on the cognitive dimension because they are more similar in "difficulty" to the cognitive items than to the items in the behavioral dimension. If the difference between dimensions was indeed based on their behavioral or cognitive nature, these two behavioral items would have loaded on the behavioral dimension.

Another example is the study of Z. Wang et al. (2014), where the authors empirically distinguished between a cognitive engagement dimension and two behavioral engagement dimensions (compliance and effortful class participation). Although Z. Wang et al. (2014) did not report descriptive statistics, it may be plausible that the distinction between behavioral and cognitive dimensions occurred due to the differences in item "difficulties." Specifically, listening very carefully (a behavioral item) may be "easier" than going back over things they do not understand (a cognitive item). Getting really involved in class activities (a behavioral item) may be “easier” than judging the quality of own ideas or work during class activities (a cognitive item). Furthermore, in this study, similarly to the study of Whitney et al. (2019), one may see that not all items in the cognitive dimension seem to be cognitive processes. Specifically, the item “If I’m not sure about things, I check my book or use other materials like charts” may be considered behavioral. Another item in the cognitive dimension seems to be double-barreled, asking about both behaviors and cognitive processes. This item is the item about searching for information from different places (a behavior) and thinking about how to put it together (a cognitive process). Further, not all items on the behavioral dimensions seem to be behaviors. Specifically, the item “I form new questions in my mind as I join in class activities” may be considered cognitive. The correlations between behavioral and cognitive engagement dimensions, found by Z. Wang et al. (2014), were higher than those found by Whitney et al. (2019). Thus, it seems that behaviors and cognitive processes, specified by Z. Wang et al. (2014), are more likely to co-occur than those specified by Whitney et al. (2019).

The occurrence of behavioral items among cognitive items and vice versa may challenge the assumption that cognitive engagement is necessarily more than behavioral engagement in the case when behavioral engagement is conceptualized as behaviors, and cognitive engagement is conceptualized as cognitive processes. In my study, I found that some behaviors, such as drawing own pictures and making own remarks, are more “difficult” than some of the cognitive processes (particularly those that are quite commonly used), such as identifying important information or connecting information with prior knowledge. Further, some behaviors in my study are those that are likely to occur conditional on some cognitive processes. For example, a behavior of writing down more than one way of solving or of thinking about the task even if the student already has an answer assumes that the student already thought about different ways of solving or answering the task even if they already have an answer (a cognitive process). Thus, these behaviors and cognitive processes may co-occur, or the cognitive process may occur without the behavior, but the behavior is unlikely to occur without the cognitive process. As a result, this behavior was more “difficult” than this cognitive process.

Notably, in contrast to my study, in neither of the explored studies, a “Difficulty” factor with cross-loading items emerged. In prior research, behavioral engagement may be an “easy” factor, and cognitive engagement may be a “difficult” factor. I can speculate that cross-loadings between the behavioral and cognitive engagement dimensions did not occur in these studies because items within a dimension were approximately of the same “difficulty.” In contrast, in my instrument, there was a range of “difficulties” of items that nevertheless had conceptual unity. This range of item “difficulties” may have caused the “Difficulty” factor with cross-loading items to occur in my study. It should also be noted that not in all prior studies, average cognitive engagement was lower than average behavioral engagement. For example, in the study of Reeve (2013), average behavioral engagement tended to be higher than average cognitive engagement; yet, in the study of Reeve (2013), referents also differed between the two dimensions.

Finally, I will note that in this discussion, I explored only internal structures of engagement measures that were developed via EFA or via CFA that was based on EFA. I did not explore internal structures based on CFA because they typically showed only adequate model fit, as indicated by very high Chi Square values and acceptable approximate global fit indices (e.g., Lam et al., 2014; Wang et al., 2011, 2016). It is fairly common in applied research to discard the Chi-Square test on the grounds of sample size sensitivity. Specifically, it is often stated that with a large enough sample size, the Chi Square test would show a statistically significant result even if model misspecifications are trivial or if data are non-normal. However, while the statistically significant Chi Square test may indeed indicate data non-normality or trivial model misspecifications, it may also indicate the presence of severe model misspecifications (McIntosh, 2007). Thus, ignoring a statistically significant Chi Square may lead to retaining a severely misspecified model, resulting in obtaining distorted parameter estimates and, subsequently, misinterpreting them (McIntosh, 2007). Approximate global fit indices (e.g., RMSEA, SRMR, CFI) also do not provide a better alternative in evaluating model fit because their use constitutes the same challenges as the use of the Chi Square test. Being global, approximate global fit indices are similarly unable to demonstrate whether the “approximate” fit is due to multiple trivial misspecifications or a few severe misspecifications (McDonald & Ho, 2002). To determine the degree of model misspecification, one needs to locate misspecifications by employing diagnostic procedures, such as residual analysis, the Lagrangian multiplier (LM) test, etc. (McIntosh, 2007). However, in the investigations of engagement dimensionality, the extent of examining sources of misfit is unknown. Thus, I did not explore CFA models because I was not confident that their internal structures and parameter estimates could be trusted. It should be noted that EFA models can also be misspecified, e.g., by specifying a wrong number of dimensions or by overlooking items that are not locally independent. Yet, in CFA models, in addition to these sources of misspecification, there is an additional source of misspecification in the form of misspecified cross-loadings (i.e., specifying cross-loadings fixed to zero when they are significantly larger than zero).

In sum, in my study, I found that emotional engagement was empirically distinct from behavioral and cognitive engagement. This finding is consistent with findings of other studies. I also found that in my study, behavioral engagement was not empirically distinct from cognitive engagement. However, this finding is not consistent with prior research. I hypothesized that the differentiation between behavioral and cognitive engagement in prior research may have occurred not because of the conceptual differences between behaviors and cognitive processes but because of the differences in the "difficulty" of behavioral and cognitive items. However, this hypothesis has been developed based on the limited information and, hence, should be approached with caution.

Specificity. The instrument was designed to measure student engagement in four instruction types (lecture, whole-class interaction, individual work, and group work) in mathematics-based classes. Thus, the instrument was designed to produce four instruction-specific composite scores and one class-specific global composite score. One argument in the rationale for measuring student engagement in multiple instruction types was the existence of qualitative differences between engagement in different instruction types. In this study, for engagement in each instruction type, I aimed to design items that were qualitatively different from engagement in other instruction types where appropriate. The four instruction types were conceptualized in the following way. Lecture was conceptualized as the time in class when the instructor explains the material without interacting with students, e.g., when the instructor lectures in a traditional sense, presents the material without asking questions along the way, etc. Whole-class interaction was conceptualized as the time in class when the instructor interacts with students addressing the class as a whole, e.g., when the instructor asks questions to the whole class, does interactive lecture, holds a whole-class discussion, or when other students ask questions in front of the whole class, etc. Individual work was conceptualized as the time in class when a student works on a task without interacting with other students (excluding exams and formal quizzes), e.g., when the student does the task on their own, starts working on the task by themselves before turning to others, etc. Group work was conceptualized as the time in class when a student interacts with other students about a task, e.g., when the student discusses the task with a neighbor, works in a group or pair, checks their answers with people sitting nearby, etc.

Qualitative differences between items that indicate engagement in different instruction types took two main forms. Some qualitatively different items were unique to the instruction types. For example, the item about trying to make sure you know why you use particular strategies or reasoning to solve or answer the task is specific to individual work. As another example, the item about justifying your thinking about the task when speaking with other students is specific to group work. Other qualitatively different items were similar in terms of the behavior, cognitive process, or emotion they measured but different in the instruction type they referred to. For example, there were four items about critical thinking applied to each instruction type: critical thinking about your instructor's explanations (lecture), critical thinking about what is being said between your instructor and other students (whole-class interaction), critical thinking about your task solution, answer, or solution attempts (individual work), and critical thinking about other students' thinking about the task, their solutions, or answers (group work). As another example, there were two items about drawing own pictures: drawing your own pictures of your instructor's explanations (lecture) and drawing your own pictures of what is being said between your instructor and other students (whole-class interaction).

Another argument in the rationale for measuring student engagement in different instruction types was the existence of quantitative differences in engagement across different instruction types. In this study, I found the correlations between engagement in different instruction types to be moderate. The highest correlations on the upper boundary of acceptable were found between engagement in lecture and engagement in whole-class interaction, as well as between engagement in lecture and engagement in individual work. The former correlation may have occurred because both lecture and whole-class interaction are teacher-centered types of instruction. It is possible that teacher-centered types of instruction are more similar than student-centered types of instruction. The latter correlation may have occurred because both lecture and individual work heavily employ students' individual thinking. Thus, in both cases, relatively high correlations are logical, as they reflect particular similarities between instruction types. Nevertheless, moderate-to-high correlations suggest that students may be more engaged in some types of instruction than in others.

Studies that explored student engagement within instruction types were studies that employed ESM (Shernoff et al., 2003; Uekawa et al., 2007). These studies measured student engagement at the level of activities, which were later classified by instruction types. While no correlations were reported in these studies, mean comparisons between engagement levels in different instruction types (Shernoff et al., 2003) or regression coefficients of instruction types with student engagement as an outcome (Uekawa et al., 2007) were available. In my study, I did not perform any significance testing to detect if mean differences between engagement in different instruction types were statistically significant. Thus, I will describe the observed mean differences in terms of tendencies. I found that engagement in group work tended to be higher than engagement in lecture, which in turn tended to be somewhat higher than engagement in individual work, which in turn tended to be higher than engagement in whole-class interaction. The lowest mean for engagement in whole-class interaction seemed to be primarily due the mean of active behavioral engagement being substantially lower than the mean of passive behavioral/cognitive engagement. Yet, both were included in the calculation of scores for the entire engagement in whole-class interaction. In the studies of Shernoff et al. (2003) and Uekawa et al. (2007), whole-class interaction was not included as a type of instruction. Engagement in group work, similarly to my results, was higher than engagement in lecture. Comparing engagement in lecture and engagement in individual work, Shernoff et al. (2003) found that engagement in individual work was higher than engagement in lecture, while Uekawa et al. (2007) did not find a significant difference between the two. Differently from the results of these studies, I found that engagement in individual work tended to be somewhat higher than engagement in lecture, although this difference was minimal. Finally, Shernoff et al. (2003) did not find a significant difference between engagement in individual work and engagement in group work. Differently from the results of Shernoff et al. (2003), I found that engagement in individual work tended to be lower than engagement in group work. Thus, some of my results seem to be consistent with prior research, whereas others seem to differ.

While there may be multiple potential reasons for the inconsistencies, I will note some. First, conceptualizations of engagement in the studies of Shernoff et al. (2003) and Uekawa et al. (2007) were more narrow than in my study. Shernoff et al. (2003) employed three items to measure engagement, and Uekawa et al. (2007) employed eight items. Second, Shernoff et al. (2003) and Uekawa et al. (2007) used an activity as a unit of analysis, whereas I used a student as a level of analysis. Third, engagement measures of Shernoff et al. (2003) and Uekawa et al. (2007) were the same regardless of the activity (and, subsequently, of the instruction type), whereas I employed similarly worded items for emotional engagement but at least partially different items across instruction types for behavioral and cognitive items. Thus, comparisons between engagement levels in different types of instruction should be approached with caution in my study. While it is plausible that emotional engagement scales are comparable across instruction types, the comparability of behavioral/cognitive scales should be examined further in the future validation work.

In addition to the ESM research, engagement in group work should be discussed in relation to the research that conceptualized student engagement from the multidimensional perspective. In particular, engagement in group work and engagement in whole-class interaction partially overlap with some conceptualizations of social engagement. For example, Rimm-Kaufman et al. (2015) conceptualized social engagement as “students’ day-to-day social exchanges with peers that are tethered to the instructional content” (p. 172). M.-T. Wang et al. (2016) conceptualized social engagement as “the quality of social interactions with peers and adults, as well as the willingness to invest in the formation and maintenance of relationships while learning” (p. 17). The conceptualization of Rimm-Kaufman et al. (2015) includes only interactions with students, whereas the conceptualization of M.-T. Wang et al. (2016) includes interactions with both students and adults. Further, the conceptualizations of both Rimm-Kaufman et al. (2015) and M.-T. Wang et al. (2016) do not seem to differentiate between whole-class and small group settings. In contrast, in my study, I differentiated between the two settings. Engagement in whole-class interaction included active and passive involvement during the time when the instructor interacts with the whole class. Engagement in group work included interactions with other students in a small group setting. Separating the two settings allows for obtaining more information about student engagement with others in multiple settings. Low-to-moderate correlations between engagement in these instruction types provided empirical evidence for this statement. A final note needs to be made about applicability of social engagement measures. My instrument explicitly asks a student about applicability of each instruction type to the student's class (by asking about the amount of time spent on each instruction type). However, from the dimensional perspective, social engagement is assumed to be applicable to all classes, regardless of whether opportunities for interaction were actually present in a particular class.

In sum, in my study, I found that the constructs of engagement in different instruction types were empirically distinct, with the most similar engagement levels being between lecture and whole-class interaction, as well as between lecture and individual work. Further, some differences in mean levels of engagement in different types of instruction were similar to those observed in prior research, whereas others were not. Yet, comparisons of my results for instruction-specific engagement with the results from prior research should be approached with caution due to the differences in methodology. Finally, engagement in group work and engagement in whole-class interaction can be compared to social engagement. Yet, measuring student engagement in these two instruction types, as my instrument does, may provide more information about student engagement with others than measuring social engagement as a single dimension.

Dimensionality and specificity. In this study, I aimed to combine multidimensionality and instructional specificity in engagement measurement. Specifically, I measured the following kinds of engagement: emotional engagement in each instruction type and behavioral/cognitive engagement in each instruction type with the exception of Whole-Class Interaction, where I specified passive behavioral/cognitive engagement and active behavioral engagement. I found that correlations between the engagement subscale constructs were typically low-to-moderate, with the highest correlations between behavioral/cognitive engagement in lecture and passive behavioral/cognitive engagement in whole-class interaction, as well as between emotional engagement in lecture and whole-class interaction. As mentioned above, these results may have occurred due to the similarity between these instruction types in terms of being teacher-centered. It should be noted that active behavioral engagement in whole-class interaction correlates with any other subscale engagement construct only weakly. Overall, the observed correlations suggest that students may be more engaged in some kinds of engagement than in others, where "kind" incorporates both dimensionality and instructional specificity. Therefore, the instrument may be able to identify the kinds of engagement, in which students are not engaged. Thus, I provided initial evidence for the first use of my instrument: identifying where and how students are not engaged.

As no other study, to my knowledge, measured multidimensional, instruction-specific engagement, no direct comparisons with prior research can be made. Yet, Shernoff et al. (2003) reported mean differences between instruction types separately for engagement items. Two items (interest and enjoyment) could be considered emotional, and one item (concentration) could be considered behavioral or cognitive. Shernoff et al. (2003) found similar results for each of their items (interest, enjoyment, and concentration). Specifically, similarly to overall engagement, they found that average interest, enjoyment, and concentration in group work and in individual work were higher than in lecture. Yet, they did not find significant differences in average interest, enjoyment, and concentration between group work and individual work. In my study, I did not perform any significance testing to detect if mean differences between subscale engagement in different instruction types were statistically significant. Thus, as in the previous section, I will describe the observed mean differences in terms of tendencies. For emotional engagement, similarly to the results of Shernoff et al. (2003), I found that students tended to be more emotionally engaged in group work than in lecture. Yet, contrary to the results of Shernoff et al. (2003), I found that students tended to be more emotionally engaged in group work or in lecture than in individual work. Additionally, I found that emotional engagement in whole-class interaction tended to be lower than in group work (albeit minimally) but higher than in individual work or lecture. In terms of behavioral/cognitive engagement, my results are consistent with the results of Shernoff et al. (2003). Specifically, I found that emotional engagement in group work and individual work was higher than in lecture. However, the levels of behavioral/cognitive engagement were the same between group work and individual work. Additionally, behavioral/cognitive engagement in group work, individual work, and lecture tended to be higher than passive behavioral/cognitive engagement in whole-class interaction, which in turn tended to be higher than active behavioral engagement in whole-class interaction. Thus, while Shernoff et al. (2003) found similar patterns of mean differences across emotional and behavioral/cognitive items, I found different patterns. This observation suggests that multidimensional, instruction-specific measurement of engagement may be able to better detect differences in students' levels of engagement dimensions across instruction types than ESM measurement.

Further, results from the regression analyses showed that subscale engagement constructs differentially predicted achievement. Let's consider actual grades in percent as an example. I found that global student engagement was very weakly related to actual grades in percent. Considering engagement by dimensions, I found that it was actually emotional engagement that predicted actual grades in percent. Considering engagement by instruction types, I found that it was engagement in whole-class interaction and engagement in individual work that positively predicted actual grades in percent. However, when engagement was considered by both dimensions and instruction types, I found that actual grades in percent were predicted specifically by active behavioral engagement in whole-class interaction and emotional engagement in individual work. Thus, combining multidimensionality and instructional specificity in engagement measurement provides more information than either the multidimensional measurement or the instruction-specific measurement. Thus, multidimensional, instruction-specific measurement of engagement allows educators to identify the specific kinds of engagement that have the largest effects on achievement. This information has the potential to help educators focus their efforts to increase student engagement on the kinds of engagement that are most beneficial for student achievement.

A final note needs to be made about active behavioral engagement in whole-class interaction. The separation of the active behavioral items in whole-class interaction may indicate the need to re-consider the type of engagement that these items represent. For example, this subscale construct may be better conceptualized as agentic engagement (Reeve, 2013; Reeve & Tseng, 2011), measured at the class level, rather than instruction-specific engagement. Reeve and Tseng (2011) defined agentic engagement as “students’ constructive contributions into the flow of the instruction they receive” (p. 258). My item about asking questions to the instructor in front of the whole class is similar to the item of Reeve (2013) about asking questions in class to help oneself learn. However, some other items of Reeve (2013) are focused on expressing opinions and preferences. Such items seem to be conceptually different from my items that focus on contributing to instructor’s interactions with the whole class. Yet, whether my items are empirically distinct from the items that indicate agentic engagement is a question for future research.

In sum, I found low-to-moderate correlations between subscale constructs, suggesting that a student may have different levels of different kinds of engagement. Thus, combining multidimensionality and instructional specificity in engagement measurement provides an opportunity to identify specific kinds of engagement that students may lack. Comparing to the prior research, I found that multidimensional, instruction specific measurement may be able to better detect the differences in students’ levels of engagement dimensions across instruction types. Next, in the regression analyses, I found that combining multidimensionality and instructional specificity in engagement measurement allows for identifying the specific kinds of engagement that affect achievement the most. A final note needs to be made about active behavioral engagement in whole-class interaction, which has some similarities with the concept of agentic engagement.

Student Engagement as a Formative Construct

In educational and psychological research, constructs are routinely treated as reflective. Yet, researchers across different fields in social sciences and beyond have developed and used formative constructs, as well. To illustrate what kinds of constructs have been approached as formative, I provide several examples. One of the most well-known formative constructs is socio-economic status (e.g., Bollen & Lennox, 1991; Borsboom et al., 2003). It should be noted that formative constructs can be first-order constructs or higher-order constructs. An example of first-order formative constructs in educational research are home environment (e.g., number of books at home or frequency of talking about studies at home), school curriculum (e.g., frequency of painting or drawing in an arts class or frequency of class visits to art museum/gallery), personal attributes (e.g., liking to do art work or taking an art course), and art-related not-for-school (extracurricular) activities (e.g., entering an art competition or talking to family/friends about art) (Xu et al., 2018). In psychology, a construct considered as potentially being a first-order formative construct is executive function that consists of children's performance on a number of executive function tasks (Willoughby & Blair, 2016). In recreation research, formative measurement was employed for the constructs of leisure constraints with respect to recreation in parks: intrapersonal constraints (e.g., poor health or fear of crime), interpersonal constraints (e.g., not having a companion to go to park with or preference of friends/family to recreate elsewhere), and structural (e.g., lack of time or inability to get to parks) (Kyle & Jun, 2015). In health, an example of a first-order formative construct is quality of life that consists of multiple life domains, such as health, income, education, social contact, happiness, etc. (Felix & Garcia-Vega, 2012).

When second-order formative constructs are specified, first-order constructs can be specified as either reflective or formative. For example, in education, Thien (2019) conceptualized quality of school life as a second-order formative construct that consists of the following first-order reflective constructs: positive affect, teacher-student relations, status, identity, opportunity, and achievement. In sport psychology, examples of second-order formative constructs are dimensions of competitive anxiety: cognitive, physiological, and regulatory (Jones et al., 2019). For instance, the cognitive dimension is indicated by three first-order reflective constructs: worry, private self-focus, and public self-focus. Another example of a second-order formative construct is employee well-being that is indicated by four first-order reflective constructs: work-life balance, job wellness, physical wellness, and purpose in life (Khatri & Gupta, 2019). Rodrigues et al. (2018) presented hierarchical constructs that are formative at each of the three levels. In particular, Rodrigues et al. (2018) conceptualized psychological empowerment as a third-order formative construct that consists of four components: behavioral, cognitive, emotional, and relational empowerment. For instance, behavioral empowerment consisted of three first-order formative constructs: activism, civic engagement, and online civic engagement.

In this study, I re-conceptualized student engagement as a formative construct. Specifically, I re-conceptualized subscale constructs (i.e., multidimensional, instruction-specific constructs) as first-order formative constructs. The subscale constructs are indicated by behaviors, cognitive processes, or emotions, characteristic of particular instruction types. While these indicators are conceptually similar to represent particular kinds of engagement, they are also sufficiently different to represent a variety of aspects of these engagement kinds.

In existing studies, the construct of student engagement has been typically assumed to be reflective. In my study, I also initially approched measurement of student engagement as a reflective construct. Differently from me, prior engagement research, to my knowledge, did not find a “Difficulty” factor. This factor emerges when conceptually similar items with different item “difficulty” are used to indicate a construct. In my study, the “Difficulty” factor was especially profound for behavioral and cognitive items. As I discussed in the section about dimensionality, when (linear) reflective modeling is applied to behavioral and cognitive engagement, it is possible that emerging factors reflect differences in item “difficulties” rather than conceptual differences between dimensions. Linear factor analytic methods (e.g., EFA, CFA, ESEM) are commonly used to analyze the internal structure of engagement. However, an uncritical application of a linear factor analysis perhaps should be challenged when items are of different “difficulty.” In such situations, a linear factor analysis may lead to (1) the development of artificial factors that reflect differences in “difficulty” rather than differences in the content and/or (2) the shrinkage of content validity when researchers try to improve measurement models by removing cross-loading items. However, researchers may remove items that cross-load not because they are poor indicators of the construct but because they have a different level of “difficulty.”

In my study, if I were to develop reflective engagement constructs based on the ESEM results, I would have needed to remove all “difficult” items that cross-loaded on the “Difficulty” factor to obtain clearer engagement factors. However, doing that would have resulted in having constructs of “easy” engagement and, hence, the problem of construct underrepresentation. In representing the constructs of engagement, it is important to capture a variety of possible ways to engage, not just the “easy” ways. Fredricks et al. (2004) discussed what they referred to as “qualitative differences in the level or degree of engagement” (p. 61) within each dimension. Specifically, Fredricks et al. (2004) suggested that behavioral engagement “can range from simply doing the work and following the rules to participating in the student council,” cognitive engagement “can range from simple memorization to the use of self-regulated learning strategies that promote deep understanding and expertise,” and emotional engagement “can range from simple liking to deep valuing of, or identification with the institution” (p. 61). These “qualitative differences in the level or degree of engagement” (p. 61) may be similar to my notion of “difficulty” in that higher degrees of engagement (e.g., participating in the student council) are more “difficult” than lower degrees of engagement (e.g., doing the work). Thus, researchers, who consider student engagement to be a reflective construct, may want to consider applying nonlinear factor analysis to prevent eliminating items that are good indicators of the construct but are related to it in a nonlinear way.

In my view of student engagement as a formative construct, a range in item “difficulty” is desired, as it leads to a better representation of the construct. In other words, formative constructs of engagement are formed via behaviors, cognitive processes, or emotions that represent a variety of ways to engage in class, ranging from “easy” to “difficult.” Such constructs are able to better differentiate between students with different engagement levels, compared to constructs that are more homogeneous in terms of “difficulty.” However, nonlinearity between items and constructs may also apply to formative measurement. While in this study I developed formative constructs as composites based on theoretical grounds (the correspondence between indicators and construct conceptualizations) and on the information from the ESEM analysis, future validation efforts may include an empirical investigation of the internal structure that accounts for nonlinearity. Specifically, a nonlinear PCA may be explored in the future as a way to empirically investigate the internal structure of formative engagement constructs, indicators of which range in “difficulty.”

In addition to the “Difficulty” factor, another problem that I had in my study in the lack of model fit and the presence of cross-loadings unrelated to the “Difficulty” factor. This problem is not atypical for situations when formative constructs are modeled as reflective because formative indicators are not expected to fit together in a way that reflective indicators are. Thus, it is possible that in prior research, good indicators of student engagement may have been excluded because they did not fit well or cross-loaded in a reflective model. One may argue that in some existing studies, student engagement was modeled as reflective but did not seem to have fit problems. The reason may lie in operationalizing engagement in a more homogeneous way. For example, emotional engagement in my study was operationalized as positive activating emotions, negative activating emotions, positive deactivating emotions, and negative deactivating emotions. In contrast, Reeve and Tseng (2011) operationalized emotional engagement as curiosity, interest, enjoyment, and fun. Z. Wang et al. (2014) operationalized emotional engagement as interest, pride, excitement, happiness, and amusement. These operationalizations produce more narrow constructs than mine.

Besides first-order formative engagement constructs (i.e., subscale constructs), I also developed second- and third-order formative engagement constructs. Dimension engagement constructs and instruction type engagement constructs in my study are second-order constructs. It should be noted that second-order constructs were developed not directly based on first-order constructs but based on 2x4 constructs. The eight 2x4 constructs differed from nine first-order constructs in one aspect. Specifically, two subscale constructs – passive behavioral/cognitive engagement and active engagement in whole-class interaction – were combined into a single construct of behavioral/cognitive enagement in whole-class interaction. Thus, a dimension engagement construct is a combination of engagement in instruction types within this dimension, weighted by the amount of time students spent on each instruction type. An instruction type engagement construct is a combination of dimension engagement constructs within this instruction type. Finally, global engagement is a third-order formative construct, which is a combination of dimension engagement constructs.

A final note needs to be made about the variability of conceptualizations of student engagement in the literature. Engagement conceptualizations differ to a various degree between researchers. Some differences in conceptualizations are driven by the choice of theoretical frameworks. For example, engagement within the three-dimensional framework of Fredricks et al. (2004) includes behavioral, cognitive, and emotional dimensions. However, in another framework, student engagement is characterized by energy, dedication, and absorption (Salmela-Aro & Upadaya, 2012). Even within a single framework, conceptualizations also differ to a various degree. For example, Archambault et al. (2009) conceptualized behavioral engagement as school attendance and discipline, whereas Kong et al. (2003) conceptualized it as attentiveness, diligence, and time spent on homework. M.-T. Wang et al. (2011) conceptualized cognitive engagement as self-regulated learning and cognitive strategy use, whereas Miller et al. (1996) conceptualized it as self-regulation, deep strategy use, shallow processing strategy use, effort, and persistence. Further, M.-T. Wang et al. (2011) conceptualized emotional engagement as school belonging and valuing of school education, whereas Lam et al. (2014) conceptualized it as students' feelings about learning and their school. If student engagement was a reflective construct and, therefore, a real entity, then any conceptualization would indicate the same entity and would be interchangeable with another conceptualization. Yet, it does not seem plausible that the meaning of student engagement or its dimensions are the same in the examples above, regardless of the conceptualization. Christenson et al. (2012) suggested that the consensus among researchers about the definitions of student engagement and its dimensions is not likely to be possible; moreover, the consensus might not be needed. However, in the view of Christenson et al. (2012), researchers need to provide definitions and measures of student engagement. While the variability of definitions presents a problem for conceptualizing student engagement as reflective, it does not present a problem for conceptualizing student engagement as formative. Formative measurement of student engagement views it as an entity constructed by people rather than a real entity. Thus, the emerged variety of conceptualizations may support the view of engagement as a formative construct rather than reflective.

In sum, in this study, I re-conceptualized student engagement as a formative construct. Specifically, I re-conceptualized subscale constructs (i.e., multidimensional, instruction-specific constructs) as first-order formative constructs. These constructs represent ways of engagement that are characteristic of a particular dimension and instruction type and encompass different degrees of "difficulty." Further, I conceptualized dimension engagement constructs and instruction type engagement constructs as second-order formative constructs. Finally, I conceptualized global engagement as a third-order formative construct. Conceptualizing student engagement as formative may explain how the construct of student engagement can have multiple conceptualizations present in the literature.

Limitations and Directions for Future Research

This study entailed development and initial validation of the instrument that is designed to measure multidimensional, instruction-specific engagement in mathematics-based undergraduate classes. Due to the scope of the study, some assumptions in the interpretation/use argument were not evaluated empirically. Among the assumptions I planned to evaluate in this study, not for all assumptions sufficient validity evidence was provided. For example, the analysis of internal structure was initially planned for engagement as a reflective construct. Thus, empirical evidence from this analysis for engagement as a formative construct is limited. Therefore, future research should continue working on providing further validation for the interpretation and use of instrument scores.

Results from this study also provide some suggestions for potential revisions in the levels of specificity. First, I found relatively high correlations between subscale engagement constructs in lecture and subscale engagement constructs in whole-class interaction (with the exception of active behavioral engagement). Currently, engagement in lecture refers to engagement with instructors' explanations when the instructor does not interact with the class. Engagement in whole-class interaction (with the exception of active behavioral engagement) refers to engagement with the interactions between the instructor and other students. Active behavioral engagement refers to students' own interactions with the class. It is possible that more information could be obtained if the focus of engagement changed from the instruction type in the whole-class setting (lecture or whole-class interaction) to the source of talk in the whole-class setting: instructor, other students, or self. Thus, one type of engagement could refer to engagement with the instructor's explanations, which would capture instructor's explanations both during lecture and whole-class interaction. Another type of engagement could refer to engagement with other students' contributions made during whole-class interaction. Finally, the last type of engagement could be active behavioral engagement, emerged in this study. In addition to increasing the amount of information, such revisions may also help to avoid problems with differentiation between referents for lecture and whole-class interaction that occurred during cognitive interviews for some students. Of course, engagement with instructor's explanantions assumes that it does not differ between the settings of lecture and whole-class interaction. However, this assumption is testable.

Second, a critical aspect of behavioral/cognitive engagement in group work is its sole focus on the social aspect of group work. Yet, behavioral/cognitive engagement in group work involves not only engagement with other students but also engagement with the task. At the stage of item writing, I considered including behavioral/cognitive items that indicate both task engagement in the group setting and engagement with group members. However, I decided not to include task engagement because it is not a distinguishing feature of behavioral/cognitive engagement in group work. Including task engagement would also increase the length of an already long instrument. However, I am concerned that with the current conceptualization of group work, we are not able to learn about task engagement of students who did not work on tasks individually. Another concern is the content validity of group work engagement. A revision to consider is to change the focus of behavioral/cognitive engagement from the instruction type (individual work or group work) to the source of engagement (task or other students) when working on a task. Task engagement could refer to individual behaviors and cognitive processes when working on the task, either in the individual or group setting. Engagement with other students could refer to the social aspect of group work. Such revisions may capture all aspects of working on tasks in both individual and group settings. Of course, task engagement in this conceptualization assumes that it does not differ between individual and group settings. However, this assumption is testable.

A problem of the lack of separability of behavioral and cognitive engagement dimensions should be investigated further. One way to improve our ability to differentiate between the two dimensions is to provide more specific conceptualizations of each dimension that would permit clearer classifications of indicators. Improving conceptualizations is particularly important since empirical methods to examine the separability between behavioral and cognitive dimensions in formative measurement are limited. However, empirical methods such as nonlinear PCA may be promising and should be explored. Further, one may consider employing other methods, such as sorting tasks conducted by experts. In such tasks, experts would sort items by dimensions. As a result, one will be able to see whether an item is consistently viewed as behavioral or cognitive, or whether an item receives inconsistent judgements.

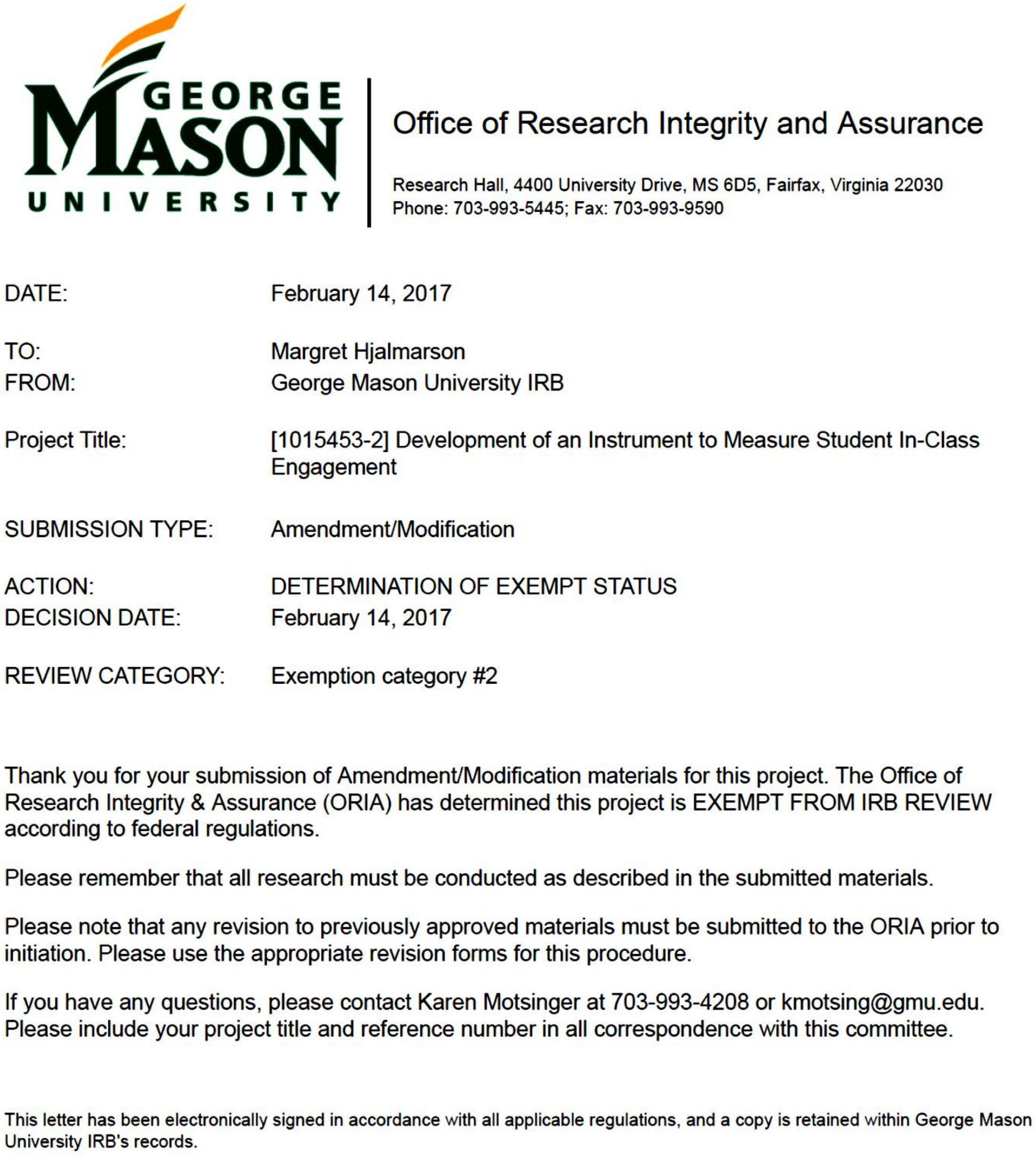
Finally, a note should be made about content validity. The instrument underwent expert reviewing, which provided some evidence for content representativeness and some suggestions to improve it. Yet, expert reviews were conducted with reflective measurement in mind and were not designed to evaluate whether subscales contained a census of indicators. Additionally, I found that distributions of item “difficulty” differed between subscales. Thus, in the future work, I may need to revise items to provide evidence for census of indicators within subscales and comparable distributions of item “difficulty.”

Conclusion

In this study, I developed and provided initial validity evidence for the instrument to measure multidimensional, instruction-specific student engagement in mathematics-based undergraduate classes. The results of this study demonstrated the potential for combining multidimensionality and instructional specificity in engagement measurement. Further, the results suggested that student engagement may be conceptualized as a formative construct rather than as a reflective construct. Scores on formative multidimensional, instruction-specific engagement constructs allow educators to identify how (the dimension) and where (the instruction type) students lack engagement. This information may inform the foci of instructional interventions that aim to improve student engagement.

Appendix A

IRB approval for exploratory interviews, expert reviews, and cognitive interviews



GEORGE MASON UNIVERSITY

Office of Research Integrity and Assurance

Research Hall, 4400 University Drive, MS 6D5, Fairfax, Virginia 22030
Phone: 703-993-5445; Fax: 703-993-9590

DATE:	February 14, 2017
TO:	Margret Hjalmarson
FROM:	George Mason University IRB
Project Title:	[1015453-2] Development of an Instrument to Measure Student In-Class Engagement
SUBMISSION TYPE:	Amendment/Modification
ACTION:	DETERMINATION OF EXEMPT STATUS
DECISION DATE:	February 14, 2017
REVIEW CATEGORY:	Exemption category #2

Thank you for your submission of Amendment/Modification materials for this project. The Office of Research Integrity & Assurance (ORIA) has determined this project is EXEMPT FROM IRB REVIEW according to federal regulations.

Please remember that all research must be conducted as described in the submitted materials.

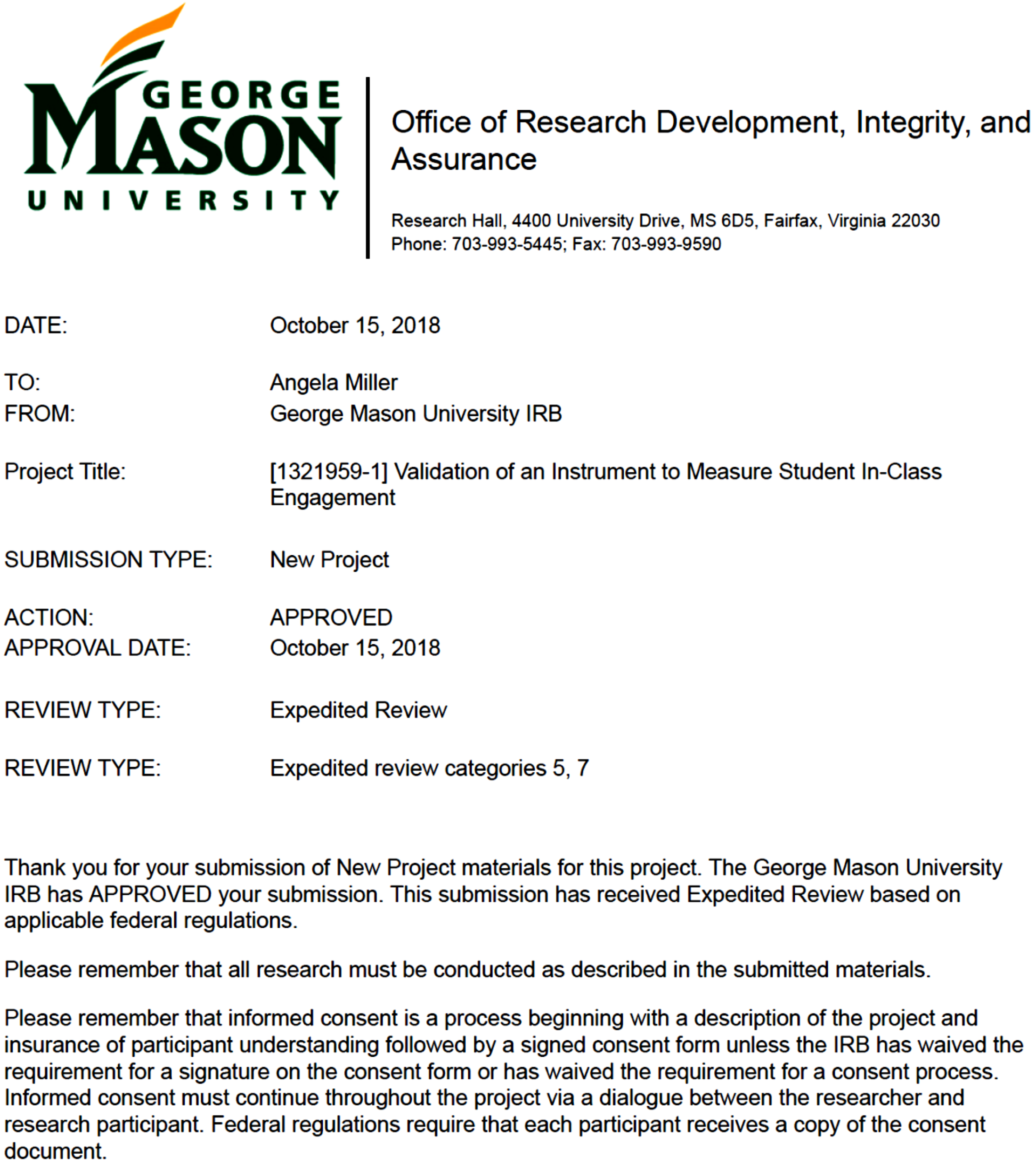
Please note that any revision to previously approved materials must be submitted to the ORIA prior to initiation. Please use the appropriate revision forms for this procedure.

If you have any questions, please contact Karen Motsinger at 703-993-4208 or kmotsing@gmu.edu. Please include your project title and reference number in all correspondence with this committee.

This letter has been electronically signed in accordance with all applicable regulations, and a copy is retained within George Mason University IRB's records.

Appendix B

IRB approval for field-testing



GEORGE MASON UNIVERSITY

Office of Research Development, Integrity, and Assurance

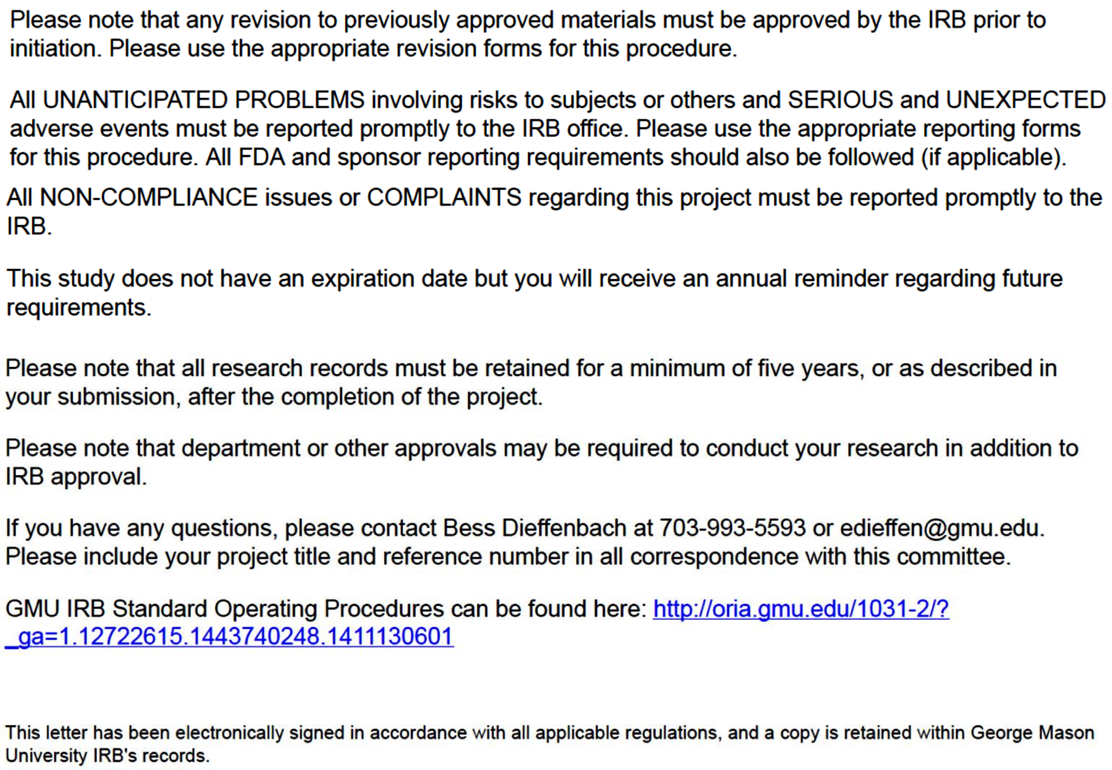
Research Hall, 4400 University Drive, MS 6D5, Fairfax, Virginia 22030
Phone: 703-993-5445; Fax: 703-993-9590

DATE:	October 15, 2018
TO:	Angela Miller
FROM:	George Mason University IRB
Project Title:	[1321959-1] Validation of an Instrument to Measure Student In-Class Engagement
SUBMISSION TYPE:	New Project
ACTION:	APPROVED
APPROVAL DATE:	October 15, 2018
REVIEW TYPE:	Expedited Review
REVIEW TYPE:	Expedited review categories 5, 7

Thank you for your submission of New Project materials for this project. The George Mason University IRB has APPROVED your submission. This submission has received Expedited Review based on applicable federal regulations.

Please remember that all research must be conducted as described in the submitted materials.

Please remember that informed consent is a process beginning with a description of the project and insurance of participant understanding followed by a signed consent form unless the IRB has waived the requirement for a signature on the consent form or has waived the requirement for a consent process. Informed consent must continue throughout the project via a dialogue between the researcher and research participant. Federal regulations require that each participant receives a copy of the consent document.



Please note that any revision to previously approved materials must be approved by the IRB prior to initiation. Please use the appropriate revision forms for this procedure.

All UNANTICIPATED PROBLEMS involving risks to subjects or others and SERIOUS and UNEXPECTED adverse events must be reported promptly to the IRB office. Please use the appropriate reporting forms for this procedure. All FDA and sponsor reporting requirements should also be followed (if applicable).

All NON-COMPLIANCE issues or COMPLAINTS regarding this project must be reported promptly to the IRB.

This study does not have an expiration date but you will receive an annual reminder regarding future requirements.

Please note that all research records must be retained for a minimum of five years, or as described in your submission, after the completion of the project.

Please note that department or other approvals may be required to conduct your research in addition to IRB approval.

If you have any questions, please contact Bess Dieffenbach at 703-993-5593 or edieffen@gmu.edu. Please include your project title and reference number in all correspondence with this committee.

GMU IRB Standard Operating Procedures can be found here: http://oria.gmu.edu/1031-2/?_ga=1.12722615.1443740248.1411130601

This letter has been electronically signed in accordance with all applicable regulations, and a copy is retained within George Mason University IRB's records.

Appendix C

Instructional Time Form items tested during cognitive interviews

Stem:

If you think of all in-class instructional time in the lecture section of this class as 100%, what percentage of time has been spent on:
If you think of all instructional in-class time (excluding exams, formal quizzes, etc.) in this class (lecture section only) as 100%, what percentage of time have been spent on:
If you think of all instructional time (excluding exams, formal quizzes, etc.) in this class (lecture section only) as 100%, what percentage of time have been spent on:
If you think of all instructional time (excluding exams, formal quizzes, etc.) in this class as 100%, what percentage of in-class time have been spent on:
If you think of all instructional time (excluding exams, formal quizzes, etc.) in this class as 100%, what percentage of time have been spent on:

Lecture:

The time in class when your instructor explains the material without interacting with students
The time in class when your instructor explains the material without interacting with students (e.g., when you instructor lectures in a traditional sense, presents the material without asking questions along the way, etc.)
The time in class when your instructor explains the material without interacting with students, e.g., when you instructor lectures in a traditional sense, presents the material without asking questions along the way, etc.
The time in class when your instructor explains the material without interacting with the class

Whole-Class Interaction:

The time in class when your instructor interacts with students
The time in class when your instructor interacts with students addressing the class as a whole

The time in class when your instructor interacts with students addressing the class as a whole (e.g., when your instructor asks questions to the whole class, does interactive lecture, holds a whole-class discussion, etc.)
The time in class when your instructor interacts with students addressing the class as a whole, e.g., when your instructor asks questions to the whole class, does interactive lecture, holds a whole-class discussion, etc.
The time in class when your instructor interacts with students addressing the class as a whole, e.g., when your instructor asks questions to the whole class, does interactive lecture, holds a whole-class discussion, or when other students ask questions in front of the whole class, etc.
The time in class when your instructor interacts with the class
The time in class when your instructor interacts with the class as a whole

Individual Work:

The time in class when YOU interact with other students about a task
The time in class when YOU interact with other students about a task (e.g., when you discuss the task with a neighbor, work in a group or pair, check your answers with people sitting nearby, etc.)
The time in class when you interact with other students about a task
The time in class when you interact with other students about a task (e.g., when you discuss the task with a neighbor, work in a group or pair, check your answers with people sitting nearby, etc.)
The time in class when you interact with other students about a task, e.g., when you discuss the task with a neighbor, work in a group or pair, check your answers with people sitting nearby, etc.

Group Work

The time in class when YOU work on a task without interacting with other students (do not include exams and formal quizzes)
The time in class when YOU work on a task without interacting with other students (excluding exams and formal quizzes)
The time in class when YOU work on a task without interacting with other students (e.g., when you do the task on your own or when you start working on the task by yourself before turning to others; do not include exams and formal quizzes)
The time in class when you work on a task without interacting with other students
The time in class when you work on a task without interacting with other students (e.g., when you work on a task by yourself or when you start working on a task before turning

to other students; excluding exams and formal quizzes)
The time in class when you work on a task without interacting with other students (excluding exams and formal quizzes)
The time in class when you work on a task without interacting with other students (excluding exams and formal quizzes), e.g., when you do the task on your own, start working on the task by yourself before turning to others, etc.
The time in class when you work on a task without interacting with other students (excluding exams and formal quizzes), e.g., when you do the task on your own, when you start working on the task by yourself before turning to others, etc.
The time in class when you work on a task without interacting with other students (e.g., when you do the task on your own or when you start working on the task by yourself before turning to others; do not include exams and formal quizzes)

Not working on a task

The time in class when you are expected to work on a task but you DO NOT work on it
The time in class when you decide not to work on a task
The time in class when you decide not to work on a task (e.g., when you choose to go on social media or have a non-task related conversation with other students while you could and should be working on the task)

Appendix D

Engagement items (and their variations) tested during cognitive interviews

Lecture: Behavioral (LB)

#	During the time when your instructor explains the material without interacting with the class, how often have you…
1	Listened to your instructor? Listened to your instructor's explanations? Listened to your instructor when he/she is explaining the material without interacting with the class?
2	Taken notes? Taken notes on what your instructor is explaining? Taken notes on what your instructor is saying/showing? Taken notes when your instructor is explaining the material without interacting with the class?
3	Read the instructor's notes, PowerPoint slides, etc.? Read what the instructor is writing or showing (e.g., instructor's notes, PowerPoint slides, etc.)? Read what your instructor is writing or showing in class (e.g., instructor's notes, PowerPoint slides, etc.)? Read the instructor's notes, PowerPoint slides, etc. when he/she is explaining the material without interacting with the class?
4	Followed the instructor along with your head and eyes?
5	Listened to all of your instructor's explanations?
6	Read all of what the instructor is writing or showing (e.g., instructor's notes, PowerPoint slides, etc.)?
7	Made your own representations of your instructor's explanations? Made your own representations (e.g., diagrams) of your instructor's explanations? Made your own pictorial or graphical representation (e.g., a diagram) of your instructor's explanations? Drawn your own pictorial representations of your instructor's explanations? Drawn your own pictures of your instructor's explanations? Drawn your own representations of your instructor's explanations?
8	Written your own remarks or comments on your instructor's explanations?

Lecture: Cognitive (LC)

#	During the time when your instructor explains the material without interacting with the class, how often have you…
1	Tried to select relevant information to write down or remember? Tried to select relevant information to write down or remember from what your instructor is saying? Tried to select relevant information to write down or remember from what your instructor is saying/showing? Tried to select relevant information to write down or remember from what your instructor is explaining? Tried to select information to write down or remember from what your instructor is explaining? Tried to select relevant information to write down or remember from what your professor is explaining without interacting with the class? Tried to identify important information from what your instructor is explaining? Tried to identify important information to write down or remember from what your instructor is explaining?
2	Tried to follow your instructor's explanations? Tried to follow what your instructor is saying/showing? Tried to follow your instructor when he/she is explaining the material without interacting with the class?
3	Tried to connect what your instructor is explaining with what he/she explained previously? Tried to connect what your instructor is saying with what he/she said previously? Tried to connect what you instructor is saying, while explaining the material without interacting with the class, with what he/she explained previously?
4	Tried to put what your instructor said/wrote into your own words? Tried to put what your instructor is explaining in your own words? Tried to put what your instructor is explaining into your own words? Tried to put what your instructor is saying/showing into your own words? Tried to put what your instructor is saying, while explaining the material without interacting with the class, into your own words?
5	Tried to summarize what your instructor said/showed?
6	Tried to connect what your instructor is explaining with what you know? Tried to connect what your instructor is saying/showing with what you know? Tried to make connections between what your instructor is explaining with what you know? Tried to make connections between what your instructor is saying with what you know? Tried to make connections between what your instructor is saying, while explaining the material without interacting with the class, with what you know?
7	Tried to make up your own examples or applications of the material? Tried to make up your own examples or applications of the material, when appropriate? Tried to make up your own examples or applications of the material, when appropriate, while your instructor is explaining the material without interacting with the class?
8	Critically thought of or evaluated what your instructor said/showed? Critically thought about or evaluated what your instructor is saying/showing? Critically thought about or evaluated what your instructor is saying? Critically thought of or evaluated what your instructor is saying while explaining the

	material without interacting with the class? Critically thought about or evaluated what your instructor is saying, while explaining the material without interacting without interacting with the class?
9	Tried to understand what your instructor is saying/showing? Tried to understand what your instructor is explaining? Tried to understand what your instructor is saying, while is explaining the material without interacting with the class?
10	Tried to determine if what your instructor is saying is worth paying attention to? Tried to determine if what your instructor is saying, while explaining the material without interacting with the class, is worth paying attention to?
11	Paid attention to what your instructor is explaining?
12	Critically thought about what your instructor is explaining? Critically thought about your instructor's explanations?
13	Evaluated what your instructor is explaining?
14	Given your full attention to what your instructor is explaining?
15	Tried to follow all of your instructor's explanations?

Lecture: Emotional (LE)

#	During the time when your instructor explains the material without interacting with the class, how often have you…
1	Enjoyed listening to your instructor? Enjoyed listening to your instructor when he/she is explaining the material without interacting with the class? Enjoyed the time when your instructor is explaining the material?
2	Felt interested? Felt interested when your instructor is explaining the material without interacting with the class?
3	Felt frustrated? Felt frustrated during your instructor's explanations? Felt frustrated when listening to your instructor?
4	Felt annoyed? Felt annoyed when your instructor is explaining the material without interacting with the class?
5	Felt calm? Felt calm during your instructor's explanations?
6	Felt comfortable?
7	Felt bored? Felt bored when your instructor is explaining the material without interacting with students? Felt bored when your instructor is explaining the material without interacting with the class? Felt bored when your instructor is explaining the material? Felt bored when your instructor was explaining the material?
8	Felt sluggish?
9	Felt stressed?

	Felt stressed when your instructor is explaining the material without interacting with the class?
10	Felt content? Felt content when your instructor is explaining the material without interacting with the class?
11	Felt relaxed? Felt relaxed when your instructor is explaining the material without interacting with the class?
12	Felt tired? Felt tired when your instructor is explaining the material without interacting with the class?
13	Felt apathetic?
14	Felt excited? Felt excited during your instructor’s explanations?
15	Felt anxious? Felt anxious when your instructor is explaining the material? Felt anxious during the time when your instructor was explaining the material?

Whole-class interaction: Behavioral (WB)

#	During the time when your instructor interacts with the class as a whole, how often have you…
1	Posed questions to your instructor? Posed class-related questions to your instructor? Posed class-related questions to your instructor in front of the whole class? Posed class-related questions to your instructor or made a comment in front of the whole class? Posed class-related questions to your instructor or made comments? Posed questions to your instructor in front of the whole class? Asked questions to your instructor in front of the whole class?
2	Volunteered to answer questions your instructor posed to the class? Volunteered to answer your instructor’s questions in front of the whole class? Answered or been willing to answer your instructor’s questions?
3	Listened to your instructor (e.g., to your instructor’s questions posed to the class or answers to other students’ questions)?
4	Listened to other students (e.g., to other students’ questions posed to the instructor or answers to instructor’s questions)?
5	Taken notes on what your instructor is saying (e.g., questions posed to the class, answers to other students’ questions)?
6	Taken notes on what other students are saying (e.g., questions to the instructor, answers to instructor’s questions)?
7	Turned to a student when he/she was speaking?
8	Followed your instructor along with your head and eyes?
9	Been willing to answer questions your instructor posed to the class? Been willing to answer your instructor’s questions in front of the whole class? Been willing to answer your instructor’s questions?

10	Listened to what is being said? Listened to what is being said while your instructor is interacting with the class?
11	Taken notes on what is being said? Taken notes on what is being said between your instructor and other students? Taken notes on what is being said while your instructor is interacting with the class?
12	Listened to everything that is being said?
13	Listened to everything that is being said, including what other students said to the instructor?
14	Made your own pictorial or graphical representation (e.g., a diagram) of what is being said? Made your own representations (e.g., diagrams) of what is being said between your instructor and other students? Made your own representations of what is being said between your instructor and other students? Drawn your own representations of what is being said between your instructor and other students? Drawn your own pictorial representations of what is being said between your instructor and other students? Drawn your own pictures of what is being said between your instructor and other students?
15	Made comments in front of the whole class on what your instructor or other students are saying?
16	Written your own remarks or comments on what is being said between your instructor and other students?
17	Shared your ideas or thoughts with the whole class?

Whole-class interaction: Cognitive (WC)

#	During the time when your instructor interacts with the class as a whole, how often have you…
1	Tried to determine if what your instructor is saying (e.g., questions posed to the class, answers to other students' questions) is worth paying attention to?
2	Tried to determine if what other students are saying (e.g., questions to the instructor, answers to instructor's questions) is worth paying attention to?
3	Tried to connect what your instructor is saying (e.g., questions posed to the class, answers to other students' questions) with what you know?
4	Tried to connect what other students are saying (e.g., questions to the instructor, answers to instructor's questions) with what you know?
5	Thought of a point to make or a question to ask? Thought of a question to ask your instructor either in class or outside of class? Thought of a point to make or a question to ask while your instructor is interacting with the class?
6	Answered in your head or thought about questions posed in the class (by your instructor or other students)? Answered in your head or thought about questions that your instructor or other students posed in front of the whole class? Answered in your head or thought about questions asked? Answered in your head or thought about questions asked by your instructor or other students?

	Answered in your head or thought about questions your instructor asks the class? Answered in your head or thought about questions your instructor asks?
7	Tried to understand what your instructor is saying (e.g., questions posed to the class, answers to other students' questions)?
8	Tried to understand what other students are saying (e.g., questions to the instructor, answers to instructor's questions)?
9	Tried to determine if what is being said is worth paying attention to? Tried to determine if what is being said, while your instructor is interacting with the class, is worth paying attention to?
10	Tried to select relevant information to write down or remember from what is being said? Tried to select relevant information to write down or remember from what is being said, while your instructor is interacting with the class? Tried to select information to write down or remember from what is being said? Tried to identify important information from what is being said between your instructor and other students? Tried to identify important information to write down or remember from what is being said?
11	Tried to follow what is being said? Tried to follow what is being said, while your instructor is interacting with the class?
12	Tried to connect what is being said with what was said in this class previously? Tried to connect what is being said, while your instructor is interacting with the class, with what was said in this class previously?
13	Tried to put what is being said in your own words? Tried to put what is being said between you instructor and other students in your own words? Tried to put what is being said, while your instructor is interacting with the class, in your own words?
14	Tried to connect what is being said with what you know? Tried to connect what is being said between your instructor and other students with what you know? Tried to connect what is being said, while your instructor is interacting with the class, with what you know?
15	Critically thought about or evaluated what is being said? Critically thought about or evaluated what is being said, while your instructor is interacting with the class?
16	Tried to understand what is being said? Tried to understand what is being said, while your instructor is interacting with the class?
17	Paid attention to what is being said?
18	Critically thought about what is being said? Critically thought about what is being said between your instructor and other students?
19	Evaluated what is being said?
20	Given your full attention to what is being said? Given your full attention to what is being said between your instructor and other students?
21	Tried to follow everything that is being said?

Whole-class interaction: Emotional (WE)

#	During the time when your instructor interacts with the class as a whole, how often have you…
1	Enjoyed listening to the interaction between your instructor and the class? Enjoyed the time in class when your instructor interacts with the class? Enjoyed the time when your instructor interacts with the class? Enjoyed the time when your instructor interacts with students?
2	Felt interested? Felt interested when your instructor is interacting with the class?
3	Felt frustrated? Felt frustrated during the time your instructor interacts with the class? Felt frustrated during your instructor's interactions with the class?
4	Felt annoyed? Felt annoyed when your instructor is interacting with the class?
5	Felt calm? Felt calm while your instructor interacts with the class?
6	Felt comfortable?
7	Felt bored? Felt bored when your instructor interacts with the class? Felt bored when your instructor is interacting with the class?
8	Felt sluggish?
9	Felt stressed? Felt stressed when your instructor is interacting with the class?
10	Felt content? Felt content when your instructor is interacting with the class?
11	Felt relaxed? Felt relaxed when your instructor is interacting with the class?
12	Felt tired? Felt tired when your instructor is interacting with the class?
13	Felt apathetic?
14	Felt excited? Felt excited during the interaction between your instructor and the class? Felt excited when your instructor is interacting with student addressing the whole class?
15	Felt anxious? Felt anxious during your instructor's interactions with the class?

Individual work: Behavioral (IB)

#	During the time when you work on a task without interacting with other students, how often have you...
1	Read the task?
2	Written down your thinking about the task (e.g., task solution, answer, solution attempts)? Written down your task solution, answer, or solution attempts? Written down your task solution, answer, or thinking about task?
3	Looked at your notes or other resources?

	Looked at your notes or other resources (e.g., lecture notes, Internet)? Looked at your notes or other resources (e.g., Internet)? Looked at your notes or other resources (e.g., Internet, textbook)?
4	Worked on the task?
5	Re-read the task?
6	Reviewed your task solution, answer, or thinking about the task?
7	Referred back to the task?
8	Written down in detail your task solution or thinking about the task?
9	Tried different ways of solving or thinking about the task even if you already have an answer?
10	Re-read the task before trying to solve it?

Individual work: Cognitive (IC)

#	During the time when you work on a task without interacting with other students, how often have you...
1	Tried to understand what the task asks?
2	Tried to recall from memory the content needed to solve/answer the task?
3	Tried to select relevant task information from other sources (e.g., lecture notes)? Tried to select relevant task information from other resources (e.g., lecture notes, Internet, textbook)? Tried to select relevant task information from your notes or other resources (e.g., Internet)?
4	Critically thought about or evaluated your solution/answer?
5	Thought about how to solve/answer the task?
6	Verified your task solution, answer, or thinking about the task with what the task says? Checked that your work or answer on the task fits with the task instructions/question?
7	Tried to select key information from the task? Tried to identify the most important information from the task?
8	Tried to relate the task to what you know?
9	Critically thought about your solution/answer? Critically thought about your task solution, answer, or solution attempts?
10	Evaluated your solution/answer? Evaluated your thinking about the task, your solution, or answer?
11	Checked your work on the task or task answer?
12	Tried to put the task instructions/question in your own words?
13	Tried to make sure you know why you use particular strategies or reasoning to solve or answer the task?
14	Tried to keep the task instructions/question in mind while solving or answering the task?

Individual work: Emotional (IE)

#	During the time when you work on a task without interacting with other students, how often have you...
1	Enjoyed working on your own? Enjoyed working on the task in class on your own? Enjoyed working on the task on your own?
2	Felt interested? Felt interested when working on the task in class on your own?
3	Felt frustrated? Felt frustrated working on your own in class?
4	Felt annoyed? Felt annoyed when working on the task in class on your own?
5	Felt calm? Felt calm doing the task in class by yourself?
6	Felt comfortable?
7	Felt bored? Felt bored when working on the task in class on your own? Felt bored when you work on the task in class by yourself? Felt bored working on the task by yourself?
8	Felt sluggish?
9	Felt stressed? Felt stressed when working on the task in class on your own?
10	Felt content? Felt content when working on the task in class on your own?
11	Felt relaxed? Felt relaxed when working on the task in class on your own?
12	Felt tired? Felt tired when working on the task in class on your own?
13	Felt apathetic?
14	Felt excited? Felt excited while working on the task in class by yourself? Felt excited while working on the task in class without interacting with other students?
15	Felt anxious? Felt anxious doing the task in class on your own? Felt anxious when working on the task in class without interacting with other students? Felt anxious when you work on a task without interacting with other students?

Group work: Behavioral (GB)

#	During the time when you interact with other students about a task, how often have you...
1	Discussed with other students how to solve/answer the task?
2	-
3	Listened to other students’ questions? Listened to other students’ questions, while interacting with them about the task?

4	Listened to other students’ explanations, ideas, etc.? Listened to other students’ explanations, ideas, etc., while interacting with them about the task?
5	Written down your own task solution/answer based on the other students’ input? Written down or revise the task solution/answer, while or after interacting with other students? Written down your own task solution/answer while or after interacting with them about the task?
6	Taken notes on other students’ thinking about the task or on their solution/answer? Taken notes on other students’ thinking about the task or on their solution/answer, while interacting with them about the task?
7	Looked together with other students at your or their notes or other resources? Looked together with other students at your or their notes or other resources (e.g., Internet)? Looked together with other students at your or their notes or other resources (e.g., Internet, textbook)? Looked together with other students at your or their notes or other resources, while interacting with them about the task? Looked together with other students at your or their notes or other resources (e.g., Internet, textbook), while interacting with them about the task?
8	Turned to a student when he/she was speaking?
9	Listened to other students’ questions, explanations, ideas, etc.? Listened to other students?
10	Worked on the task together with other students?
11	Shared your ideas about the task? Shared your thinking about the task with other students?
12	Asked other students a question about the task? Asked other students about their solutions, answers, or thinking about the task?
13	Bounced your ideas about the task off other students?
14	Helped other students with the task?
15	Asked other students to help you with the task?
16	Been willing to listen to other students?
17	Checked with other students to see if your answers, solutions, or approaches match theirs?
18	Looked at what other students wrote about the task?
19	Justified your thinking about the task when speaking with other students?

Group work: Cognitive (GC)

#	During the time when you interact with other students about a task, how often have you...
1	Tried to select relevant information from what other students are saying? Tried to select relevant information from what other students are saying about the task? Tried to select relevant information from what other students are saying while you are interacting with them about the task? Tried to identify relevant information from what other students are saying about the task?

2	Tried to connect other students’ thinking about the task, their solutions, or answers to your own? Tried to connect other students’ thinking about the task, their solutions, or answers to your own while interacting with them about the task?
3	Tried to use other students’ ideas, solutions, or answers in your thinking about the task? Tried to use other students’ ideas, solutions, or answers in your thinking about the task while interacting with them about the task?
4	Tried to follow what other students are saying about the task? Tried to follow what other students are saying about the task (e.g., their explanations or reasoning)? Tried to follow what other students are saying about the task (e.g., their explanations or reasoning) while interacting with them about the task?
5	Critically thought about or evaluated your solution/answer based on other students’ input?
6	Critically thought about or evaluated other students’ solutions/answers?
7	Thought about how to solve/answer the task together with other students?
8	Tried to understand other students’ thinking about the task, their solutions, or answers? Tried to understand other students’ thinking about the task, their solutions, or answers, while interacting with them about the task?
9	Critically thought about or evaluated the task solution/answer? Critically thought about or evaluated the task solution/answer, while interacting with them about the task? Critically thought about or evaluated your solution/answer, while interacting with other students about the task?
10	Tried to determine if what other students are saying is worth paying attention to? Tried to determine if what other students are saying, while interacting with them about the task, is worth paying attention to? Tried to determine if what other students are saying, while you are interacting with them about the task, is worth paying attention to?
11	Paid attention to what other students are saying? Paid attention to what other students are saying about the task?
12	Critically thought about the task solution/answer? Critically thought about other students’ thinking about the task, their solutions, or answers? Critically thought about other students’ thinking about the task, solution, or answer?
13	Evaluated the task solution/answer? Evaluated other students’ thinking about the task, their solutions, or answers?
14	Compared your and other students’ solutions/answers? Compared your and other students’ solutions /answers or ways of thinking about the task? Compared your and other students’ ways of thinking about the task?
15	Given your full attention to what other students are saying about the task?
16	Tried to follow everything that other students are saying about the task?
17	Considered what other students are saying about the task?
18	Tried to put what other students are saying about the task in your own words?

Group work: Emotional (GE)

#	During the time when you interact with other students about a task, how often have you...
1	Enjoyed interacting with other students about the task?
2	Felt interested? Felt interested when interacting with other students about the task?
3	Felt frustrated? Felt frustrated when you are interacting with other students about the task? Felt frustrated when you interact with other students about the task?
4	Felt annoyed? Felt annoyed when interacting with other students about the task?
5	Felt calm? Felt calm while talking the other students about the task?
6	Felt comfortable?
7	Felt bored? Felt bored when interacting with other students about the task? Felt bored when you interact with other students about the task?
8	Felt sluggish?
9	Felt stressed? Felt stressed when interacting with other students about the task?
10	Felt content? Felt content when interacting with other students about the task?
11	Felt relaxed? Felt relaxed when interacting with other students about the task?
12	Felt tired? Felt tired when interacting with other students about the task?
13	Felt apathetic?
14	Felt excited? Felt excited during the time you talk with other students about the task?
15	Felt anxious? Felt anxious interacting with other students? Felt anxious when talking to other students about the task? Felt anxious when you interact with other students about the task?

Appendix E

Frequencies and descriptive statistics for engagement items tested during cognitive interviews (item numbers refer to items from Appendix D)

Behavioral Engagement in Lecture: 5-point scale

	LB1	LB2	LB3	LB4	LB5	LB6
1	0	6	0	0	0	0
2	0	3	0	0	0	1
3	2	10	3	1	0	3
4	11	7	10	2	4	1
5	24	19	24	3	5	4
Mean	4.59	3.67	4.57	4.33	4.56	3.89
SD	0.60	1.43	0.65	0.82	0.53	1.17
N	37	46	37	6	9	9

Behavioral Engagement in Lecture: 7-point scale

	LB1	LB2	LB3	LB5	LB6	LB7	LB8
1	0	2	0	0	0	0	2
2	0	0	0	0	0	0	3
3	0	2	0	0	0	2	1
4	1	0	1	0	0	5	3
5	1	3	3	0	1	6	2
6	8	4	2	1	0	1	1
7	2	3	6	1	1	0	1
Mean	5.92	4.86	6.08	6.50	6.00	4.43	3.54
SD	0.79	2.07	1.08	0.71	1.41	0.85	1.90
N	12	14	12	2	2	14	13

Cognitive Engagement in Lecture: 5-point scale

	LC1	LC2	LC3	LC4	LC5	LC6	LC7	LC8	LC9
1	2	0	0	3	1	0	16	0	0
2	4	0	0	8	1	0	15	0	0
3	3	2	6	13	1	3	6	8	0
4	12	14	4	15	2	18	7	6	9
5	25	21	7	6	1	25	2	7	22
Mean	4.17	4.51	4.06	3.29	3.17	4.48	2.22	3.95	4.71
SD	1.16	0.61	0.90	1.12	1.47	0.62	1.21	0.86	0.46
N	46	37	17	46	6	46	46	21	31

Cognitive Engagement in Lecture: 5-point scale (continued)

	LC10	LC11	LC12	LC13	LC14	LC15
1	4	0	3	0	0	0
2	2	0	3	2	1	0
3	1	1	7	1	0	1
4	0	5	8	2	4	4
5	3	12	4	5	6	4
Mean	2.60	4.61	3.28	4.00	4.36	4.33
SD	1.78	0.61	1.24	1.25	0.92	0.71
N	11	18	25	10	11	9

Cognitive Engagement in Lecture: 7-point scale

	LC1	LC4	LC6	LC7	LC12	LC14	LC15
1	0	0	0	0	0	0	0
2	1	0	0	0	0	0	0
3	0	2	0	1	0	1	0
4	1	2	1	0	3	1	0
5	4	5	3	0	7	2	0
6	4	2	5	0	3	8	0
7	4	3	5	0	1	2	1
Mean	5.57	5.14	6.00	3.00	5.14	5.64	7.00
SD	1.40	1.35	0.96		0.86	1.08	
N	14	14	14	1	14	14	1

Emotional Engagement in Lecture: 5-point scale

	LE1	LE2	LE3	LE4	LE5	LE6	LE7	LE8	LE9	LE10	LE11	LE12	LE13	LE14	LE15
1	1	1	9	17	0	0	7	3	5	1	0	1	1	4	4
2	4	2	8	6	2	0	17	0	3	1	3	2	1	4	4
3	11	7	12	5	7	0	18	3	9	5	7	6	6	4	5
4	14	14	6	2	12	4	3	0	0	2	5	4	1	3	2
5	15	7	0	0	13	2	1	0	2	5	9	0	1	0	0
Mean	3.84	3.77	2.43	1.73	4.06	4.33	2.43	2.00	2.53	3.64	3.83	3.00	3.00	2.40	2.33
SD	1.07	0.99	1.07	0.98	0.92	0.52	0.91	1.10	1.22	1.28	1.09	0.91	1.05	1.12	1.05
N	46	31	35	31	35	6	46	6	19	15	25	13	12	15	15

Emotional Engagement in Lecture: 7-point scale

	LE1	LE3	LE5	LE7	LE14	LE15
1	0	3	0	1	0	5
2	0	2	0	2	1	2
3	2	4	0	2	3	2
4	7	5	4	7	6	5
5	3	0	2	1	2	0
6	2	0	6	1	2	0
7	0	0	2	0	0	0
Mean	4.36	2.79	5.43	3.57	4.07	2.50
SD	0.93	1.19	1.09	1.28	1.14	1.34
N	14	14	14	14	14	14

Behavioral Engagement in Whole-Class Interaction: 5-point scale

	WB1	WB2	WB3	WB4	WB5	WB6	WB7	WB8	WB9	WB10	WB11	WB12	WB13
1	15	5	0	0	2	3	0	0	5	0	6	0	0
2	6	5	0	0	1	0	1	0	4	0	6	1	0
3	18	1	0	1	1	2	4	2	10	2	5	1	1
4	8	1	2	3	0	1	1	2	10	13	10	1	2
5	2	2	5	3	3	1	1	3	6	19	15	1	2
Mean	2.51	2.29	4.71	4.29	3.14	2.57	3.29	4.14	3.23	4.50	3.52	3.50	4.20
SD	1.21	1.44	0.49	0.76	1.86	1.62	0.95	0.90	1.29	0.62	1.47	1.29	0.84
N	50	14	7	7	7	7	7	7	36	34	43	4	5

Behavioral Engagement in Whole-Class Interaction: 7-point scale

	WB1	WB2	WB10	WB11	WB13	WB14	WB15	WB16	WB17
1	2	3	0	2	0	2	2	2	2
2	3	3	0	3	0	2	0	4	2
3	3	1	0	3	0	2	0	3	2
4	2	2	2	1	0	3	1	2	3
5	4	5	5	3	1	2	0	0	1
6	0	0	3	1	0	3	0	1	0
7	0	0	2	1	1	0	0	1	0
Mean	3.21	3.21	5.42	3.50	6.00	3.71	2.00	3.08	2.90
SD	1.48	1.67	1.00	1.87	1.41	1.77	1.73	1.80	1.37
N	14	14	12	14	2	14	3	13	10

Cognitive Engagement in Whole-Class Interaction: 5-point scale

	WC1	WC2	WC3	WC4	WC5	WC6	WC7	WC8	WC9	WC10	WC11
1	2	0	0	0	2	1	0	0	1	3	0
2	2	1	0	0	2	1	0	0	2	3	0
3	0	3	0	3	10	8	0	3	4	6	4
4	2	2	4	1	5	17	2	1	3	13	9
5	1	1	3	3	0	23	5	3	1	18	21
Mean	2.71	3.43	4.43	4.00	2.95	4.20	4.71	4.00	3.09	3.93	4.50
SD	1.60	0.98	0.53	1.00	0.91	0.93	0.49	1.00	1.14	1.22	0.71
N	7	7	7	7	19	50	7	7	12	43	34

Cognitive Engagement in Whole-Class Interaction: 5-point scale (continued)

	WC12	WC13	WC14	WC15	WC16	WC17	WC18	WC19	WC20	WC21
1	0	5	1	0	1	0	1	1	0	0
2	1	7	1	2	0	0	3	0	0	0
3	3	11	6	2	3	1	5	2	2	2
4	4	13	8	6	5	7	11	2	5	2
5	4	7	27	7	19	13	6	6	3	5
Mean	3.92	3.23	4.37	4.06	4.46	4.57	3.69	4.09	4.10	4.33
SD	1.00	1.25	0.98	1.03	0.96	0.60	1.09	1.30	0.74	0.87
N	12	43	43	17	28	21	26	11	10	9

Cognitive Engagement in Whole-Class Interaction: 7-point scale

	WC5	WC6	WC10	WC13	WC14	WC18	WC20	WC21
1	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0
3	1	0	0	3	1	1	0	0
4	1	1	2	2	1	5	4	0
5	0	3	3	4	4	4	4	1
6	0	5	5	5	5	2	5	0
7	0	5	4	0	3	2	1	0
Mean	3.50	6.00	5.79	4.79	5.57	4.93	5.21	5.00
SD	0.71	0.96	1.05	1.19	1.16	1.21	0.97	
N	2	14	14	14	14	14	14	1

Emotional Engagement in Whole-Class Interaction: 5-point scale

	WE1	WE2	WE3	WE4	WE5	WE6	WE7	WE8	WE9	WE10	WE11	WE12	WE13	WE14	WE15
1	2	1	18	11	1	0	12	1	7	1	1	5	3	2	6
2	4	0	9	14	6	1	21	4	7	0	5	3	4	4	3
3	14	11	8	8	4	1	12	1	5	7	6	4	5	4	3
4	18	13	2	2	8	3	4	0	1	4	4	2	0	3	1
5	11	10	1	1	17	2	1	1	2	3	11	0	1	2	0
Mean	3.65	3.89	1.92	2.11	3.94	3.86	2.22	2.43	2.27	3.53	3.70	2.21	2.38	2.93	1.92
SD	1.05	0.93	1.08	1.01	1.24	1.07	0.97	1.27	1.24	1.06	1.30	1.12	1.12	1.28	1.04
N	50	35	38	36	38	7	50	7	22	17	28	14	14	15	14

Emotional Engagement in Whole-Class Interaction: 7-point scale

	WE1	WE3	WE5	WE7	WE14	WE15
1	0	3	0	0	0	5
2	0	3	0	2	0	4
3	1	4	0	5	3	2
4	5	3	3	5	8	3
5	4	1	1	2	2	0
6	1	0	5	0	1	0
7	3	0	5	0	0	0
Mean	5.00	2.71	5.86	3.50	4.07	2.21
SD	1.30	1.27	1.17	0.94	0.83	1.19
N	14	14	14	14	14	14

Behavioral Engagement in Individual Work: 5-point scale

	IB1	IB2	IB3	IB4	IB5	IB6	IB7	IB8	IB9	IB10
1	0	0	3	0	1	0	0	0	1	0
2	0	2	3	0	1	0	1	1	4	0
3	0	6	9	1	1	0	0	3	1	1
4	1	6	14	0	5	3	2	3	2	2
5	15	19	12	1	11	0	4	1	0	3
Mean	4.94	4.27	3.71	4.00	4.26	4.00	4.29	3.50	2.50	4.33
SD	0.25	0.98	1.19	1.41	1.15	0.00	1.11	0.93	1.07	0.82
N	16	33	41	2	19	3	7	8	8	6

Behavioral Engagement in Individual Work: 7-point scale

	IB3	IB8	IB9	IB10
1	0	0	0	0
2	2	1	3	0
3	0	1	1	0
4	4	4	2	1
5	0	2	4	4
6	3	2	2	3
7	3	2	0	4
Mean	4.92	4.75	4.08	5.83
SD	1.83	1.54	1.51	1.03
N	12	12	12	12

Cognitive Engagement in Individual Work: 5-point scale

	IC1	IC2	IC3	IC4	IC5	IC6	IC7	IC8	IC9	IC10	IC11	IC12	IC13	IC14
1	0	2	2	0	0	0	1	0	0	0	1	1	0	0
2	1	2	3	3	0	1	1	0	5	0	0	2	0	0
3	1	2	1	1	1	1	4	0	4	3	0	3	1	0
4	4	8	8	6	3	4	7	4	5	2	1	5	4	2
5	20	27	2	8	12	12	11	13	9	3	5	3	3	2
Mean	4.65	4.37	3.31	4.06	4.69	4.50	4.08	4.76	3.78	4.00	4.29	3.50	4.25	4.50
SD	0.75	1.11	1.30	1.11	0.60	0.86	1.10	0.44	1.20	0.93	1.50	1.22	0.71	0.58
N	26	41	16	18	16	18	25	17	23	8	7	15	8	4

Cognitive Engagement in Individual Work: 7-point scale

	IC2	IC6	IC7	IC9	IC12	IC13	IC14
1	0	0	0	0	0	0	0
2	0	0	1	0	0	1	0
3	1	0	1	1	2	0	0
4	4	1	0	0	3	2	1
5	1	2	1	5	1	2	2
6	2	4	1	1	3	5	3
7	4	5	8	5	3	2	6
Mean	5.33	6.08	6.00	5.75	5.17	5.33	6.17
SD	1.50	1.00	1.76	1.29	1.53	1.44	1.03
N	12	12	12	12	12	12	12

Emotional Engagement in Individual Work: 5-point scale

	IE1	IE2	IE3	IE4	IE5	IE6	IE7	IE8	IE9	IE10	IE11	IE12	IE13	IE14	IE15
1	3	1	3	11	1	0	9	2	7	0	1	5	1	4	3
2	9	3	8	7	3	0	13	3	7	1	2	2	1	4	2
3	12	6	13	7	7	0	14	0	4	1	2	4	4	4	3
4	9	8	5	1	11	2	4	0	2	8	6	0	2	3	6
5	7	5	1	1	8	2	1	0	1	2	9	0	1	0	0
Mean	3.20	3.57	2.77	2.04	3.73	4.50	2.39	1.60	2.19	3.92	4.00	1.91	3.11	2.40	2.86
SD	1.20	1.12	0.97	1.09	1.08	0.58	1.02	0.55	1.17	0.79	1.21	0.94	1.17	1.12	1.23
N	41	23	30	27	30	5	41	5	21	13	21	11	10	15	14

Emotional Engagement in Individual Work: 7-point scale

	IE1	IE3	IE5	IE7	IE14	IE15
1	0	1	0	1	0	2
2	0	1	0	3	2	2
3	2	3	0	2	3	2
4	3	7	5	4	3	5
5	4	0	4	1	3	1
6	3	0	0	1	1	0
7	0	0	3	0	0	0
Mean	4.67	3.33	5.08	3.33	3.83	3.08
SD	1.07	0.98	1.24	1.44	1.27	1.31
N	12	12	12	12	12	12

Behavioral Engagement in Group Work: 5-point scale

	GB1	GB3	GB4	GB5	GB6	GB7	GB8	GB9	GB10
1	0	0	0	1	18	4	0	1	1
2	1	0	0	2	9	7	0	1	1
3	5	3	2	8	11	6	1	4	3
4	7	5	4	7	8	7	1	2	1
5	7	5	7	6	1	6	5	16	6
Mean	4.00	4.15	4.38	3.63	2.26	3.13	4.57	4.29	3.83
SD	0.92	0.80	0.77	1.10	1.21	1.36	0.79	1.16	1.40
N	20	13	13	24	47	30	7	24	12

Behavioral Engagement in Group Work: 5-point scale (continued)

	GB11	GB12	GB13	GB14	GB15	GB16	GB17	GB18
1	0	0	0	0	0	0	0	0
2	0	3	1	1	0	0	2	0
3	5	5	0	0	2	0	0	6
4	7	11	7	1	0	0	1	2
5	8	9	2	1	1	1	6	1
Mean	4.15	3.93	4.00	3.67	3.67	5.00	4.22	3.44
SD	0.81	0.98	0.82	1.53	1.15		1.30	0.73
N	20	28	10	3	3	1	9	9

Behavioral Engagement in Group Work: 7-point scale

	GB6	GB11	GB12	GB17	GB18	GB19
1	6	0	0	0	0	0
2	0	0	0	0	0	0
3	2	1	1	0	0	0
4	2	2	2	1	1	1
5	3	4	3	1	1	2
6	1	3	2	5	0	2
7	0	4	6	7	2	0
Mean	2.93	5.50	5.71	6.29	5.75	5.20
SD	1.90	1.29	1.38	0.91	1.50	0.84
N	14	14	14	14	4	5

Cognitive Engagement in Group Work: 5-point scale

	GC1	GC2	GC3	GC4	GC5	GC6	GC7	GC8	GC9
1	0	0	0	0	0	1	0	0	0
2	4	5	2	2	0	0	0	1	1
3	11	12	15	9	5	2	5	9	4
4	20	17	21	8	1	3	3	8	5
5	12	13	9	15	1	1	11	12	6
Mean	3.85	3.81	3.79	4.06	3.43	3.43	4.32	4.03	4.00
SD	0.91	0.97	0.81	0.98	0.79	1.27	0.89	0.93	0.97
N	47	47	47	34	7	7	19	30	16

Cognitive Engagement in Group Work: 5-point scale (continued)

	GC10	GC11	GC12	GC13	GC14	GC15	GC16	GC17
1	4	0	2	0	0	0	0	0
2	0	0	3	0	2	2	2	0
3	2	4	4	2	6	2	1	3
4	4	4	9	1	16	5	5	3
5	1	8	6	4	23	3	1	3
Mean	2.82	4.25	3.58	4.29	4.28	3.75	3.56	4.00
SD	1.54	0.86	1.25	0.95	0.85	1.06	1.01	0.87
N	12	16	24	7	47	12	9	9

Cognitive Engagement in Group Work: 7-point scale

	GC1	GC2	GC3	GC12	GC14	GC15	GC16	GC17	GC18
1	0	0	0	0	0	0	0	0	1
2	0	0	1	0	1	0	0	0	0
3	1	1	0	1	1	1	0	0	0
4	1	3	4	1	3	2	1	0	0
5	3	4	4	4	2	4	0	6	3
6	3	2	2	6	4	4	1	2	1
7	6	4	3	2	3	3	0	5	1
Mean	5.86	5.36	5.07	5.50	5.14	5.43	5.00	5.92	4.83
SD	1.29	1.34	1.44	1.09	1.56	1.22	1.41	0.95	2.04
N	14	14	14	14	14	14	2	13	6

Emotional Engagement in Group Work: 5-point scale

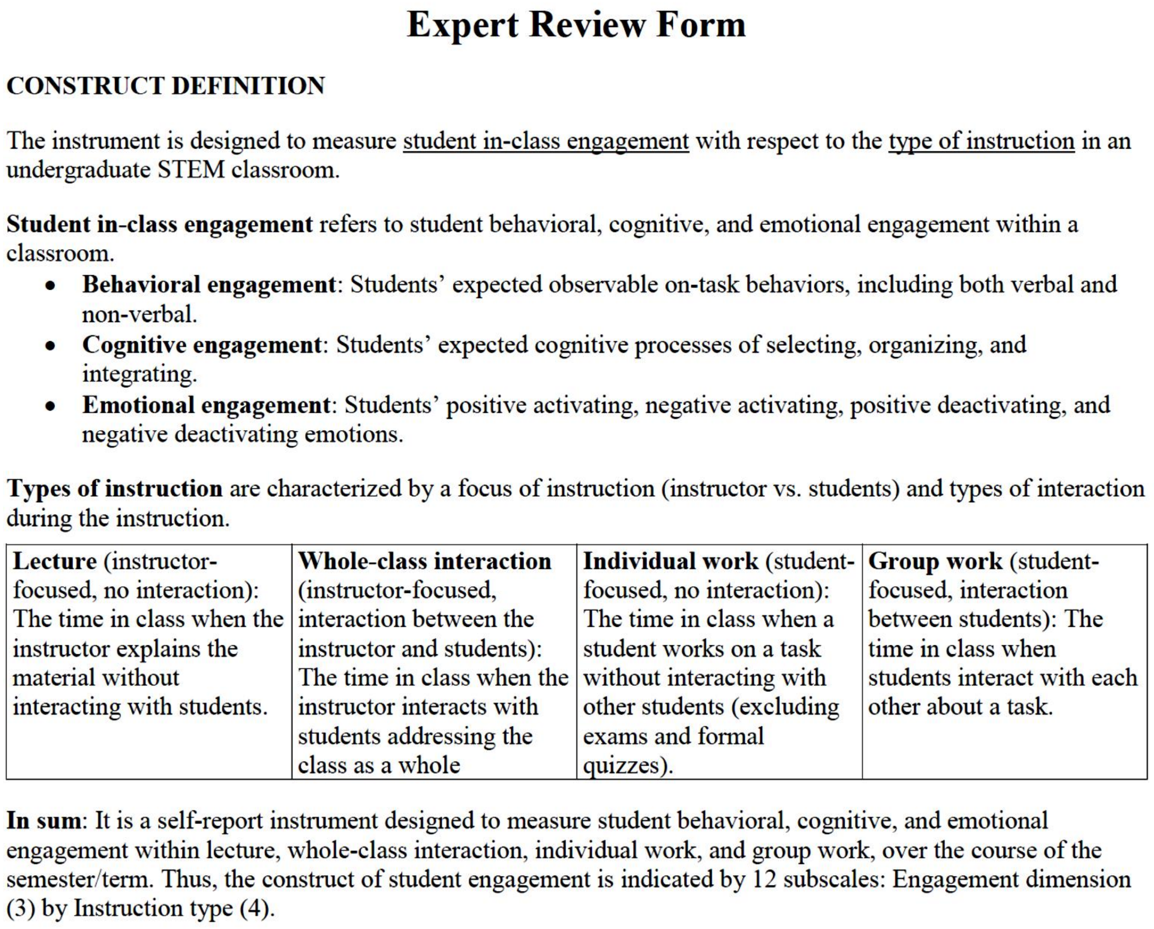
	GE1	GE2	GE3	GE4	GE5	GE6	GE7	GE8	GE9	GE10	GE11	GE12	GE13	GE14	GE15
1	3	1	8	11	2	1	11	1	7	0	2	3	2	1	8
2	5	1	10	9	5	0	21	4	5	1	3	4	4	5	2
3	11	7	11	8	4	1	12	1	5	4	3	1	1	7	4
4	8	15	4	1	12	3	1	0	1	5	7	3	4	3	2
5	20	6	2	2	12	2	2	1	2	5	7	0	0	1	0
Mean	3.79	3.80	2.49	2.16	3.77	3.71	2.19	2.43	2.30	3.93	3.64	2.36	2.64	2.88	2.00
SD	1.28	0.92	1.15	1.16	1.24	1.38	0.97	1.27	1.30	0.96	1.33	1.21	1.21	0.99	1.15
N	47	30	35	31	35	7	47	7	20	16	23	11	12	17	16

Emotional Engagement in Group Work: 7-point scale

	GE1	GE3	GE5	GE7	GE14	GE15
1	0	2	0	3	1	5
2	0	4	0	3	0	5
3	2	3	0	6	2	0
4	1	5	3	0	3	2
5	3	0	2	2	6	2
6	5	0	5	0	1	0
7	3	0	4	0	1	0
Mean	5.43	2.79	5.71	2.64	4.43	2.36
SD	1.34	1.12	1.14	1.28	1.45	1.50
N	14	14	14	14	14	14

Appendix F

Expert Review Form (Round 1)



Expert Review Form

CONSTRUCT DEFINITION

The instrument is designed to measure student in-class engagement with respect to the type of instruction in an undergraduate STEM classroom.

Student in-class engagement refers to student behavioral, cognitive, and emotional engagement within a classroom.

- **Behavioral engagement**: Students' expected observable on-task behaviors, including both verbal and non-verbal.
- **Cognitive engagement**: Students' expected cognitive processes of selecting, organizing, and integrating.
- **Emotional engagement**: Students' positive activating, negative activating, positive deactivating, and negative deactivating emotions.

Types of instruction are characterized by a focus of instruction (instructor vs. students) and types of interaction during the instruction.

Lecture (instructor-focused, no interaction): The time in class when the instructor explains the material without interacting with students.	**Whole-class interaction** (instructor-focused, interaction between the instructor and students): The time in class when the instructor interacts with students addressing the class as a whole	**Individual work** (student-focused, no interaction): The time in class when a student works on a task without interacting with other students (excluding exams and formal quizzes).	**Group work** (student-focused, interaction between students): The time in class when students interact with each other about a task.

In sum: It is a self-report instrument designed to measure student behavioral, cognitive, and emotional engagement within lecture, whole-class interaction, individual work, and group work, over the course of the semester/term. Thus, the construct of student engagement is indicated by 12 subscales: Engagement dimension (3) by Instruction type (4).

USE OF THE INSTRUMENT

To be valid and, therefore, useful, scores on an educational or psychological measure need to have clearly stated interpretations and uses (AERA, APA, & NCME, 2014; Kane, 2006). This instrument is designed to be used by researchers and practitioners (instructors teaching undergraduate STEM courses).

- Researchers may use the instrument to advance theories of teaching and learning through foundational and/or discipline-based educational research (DBER). Connecting engagement to instruction, the instrument will allow researchers to develop a more comprehensive picture of student engagement in a classroom as well as to investigate the impact of particular teaching practices on student engagement.
- Practitioners may use the instrument for the purposes of teaching improvement: learning about the levels of student engagement in a classroom and changing instruction to increase engagement.

EXPERT REVIEW INSTRUCTIONS

In the form below, you will be asked to review the questions one by one and rate each question on the following:

- **Relevance:** The question reflects, samples, and measures the construct.
- **Clarity:** The question is well written and easy to understand; the wording is clear; the length is appropriate.

After each subscale, you will be asked to rate the subscale on the following:

- **Representativeness:** All facets of the construct are included.

We ask you to rate question relevance and clarity, as well as subscale representativeness, on the following scale:

1 = Not Acceptable (major modifications needed)	2 = Below Expectations (some modifications needed)	3 = Meets Expectations (no modifications needed but could be improved with minor changes)	4 = Exceeds Expectations (no modifications needed)

Additionally, you will be asked to note potential problems, revisions, or other comments about a question or subscale.

STRUCTURE OF THE INSTRUMENT

Part 1: In-class time

Students are asked to report the percentage of instructional time spent on each instruction type: lecture, whole-class interaction, individual work, and group work, as well as the percentage of time when they decide not to work on a task. The last category is added to account for the time when the students do not work on the task either individually or in a group.

This information serves two purposes:

- Exclusion criteria for engagement subscales (i.e., if no time was spent on a particular instruction type, then a student will skip the section with engagement questions for this instruction type).
- Weights to compute scores of global engagement (i.e., engagement aggregated across subscales).

Note: The time spent on the student-focused instruction types is not expected to be necessarily similar across students in the same class, as how and whether to work on a task is ultimately at students' discretion.

In-Class Time: Lecture Section

If you think of all in-class instructional time in the <u>lecture section</u> of this class as 100%, what percentage of time has been spent on:

	The time in class when your instructor explains the material without interacting with students
	The time in class when your instructor interacts with students addressing the class as a whole
	The time in class when YOU work on a task without interacting with other students (do <u>not</u> include exams and formal quizzes)
	The time in class when YOU interact with other students about a task
	The time in class when YOU decide not to work on a task
Total = 100%	

Please verify that the percentages you specified add up to 100%.

Clarity (1 - 4):	
Suggested revisions or other comments:	

Part 2: Student Engagement Scale

The Student Engagement Scale consists of questions about student engagement. These questions are split into four sections that correspond to the instruction types. Each section has one stem and multiple questions covering behavioral, cognitive, and emotional engagement dimensions.

Response options are as follows:
1 - Never or almost never, 2 - Rarely, 3 - Sometimes, 4 - Often, and 5 - Always or almost always.

Subscale construct: BEHAVIORAL engagement in LECTURE				
#	In the LECTURE SECTION of this class, during the time in class when your instructor explains the material without interacting with students, how often have you…	Relevance (1 - 4)	Clarity (1 - 4)	Suggested revisions or other comments
1	Listened to your instructor's explanations?			
2	Taken notes on what your instructor is explaining?			
3	Read what the instructor is writing or showing (e.g., instructor's notes, PowerPoint slides, etc.)?			
Subscale representativeness (1 - 4):				
Additional comments about the subscale:				

Subscale construct: COGNITIVE engagement in LECTURE				
#	In the LECTURE SECTION of this class, during the time in class when your instructor explains the material without interacting with students, how often have you…	Relevance (1 - 4)	Clarity (1 - 4)	Suggested revisions or other comments
1	Paid attention to what your instructor is explaining?			
2	Tried to select information to write down or remember from what your instructor is explaining?			
3	Tried to follow your instructor's explanations?			
4	Tried to put what your instructor is explaining into your own words?			
5	Tried to connect what your instructor is explaining with what you know?			
6	Tried to make up your own examples or applications of the material?			

7	Critically thought about your instructor's explanations?			
Subscale representativeness (1 - 4):				
Additional comments about the subscale:				

Subscale construct: EMOTIONAL engagement in LECTURE

#	In the LECTURE SECTION of this class, during the time in class when your instructor explains the material without interacting with students, how often have you…	Relevance (1 - 4)	Clarity (1 - 4)	Suggested revisions or other comments
1	Enjoyed listening to your instructor?			
2	Felt interested?			
3	Felt annoyed?			
4	Felt frustrated?			
5	Felt calm?			
6	Felt bored?			
Subscale representativeness (1 - 4):				
Additional comments about the subscale:				

Subscale construct: BEHAVIORAL engagement in WHOLE-CLASS INTERACTION

#	In the LECTURE SECTION of this class, during the time in class when your instructor interacts with students addressing the class as a whole, how often have you…	Relevance (1 - 4)	Clarity (1 - 4)	Suggested revisions or other comments
1	Posed questions to your instructor?			
2	Been willing to answer your instructor's questions?			
3	Listened to what is being said?			
4	Taken notes on what is being said?			
Subscale representativeness (1 - 4):				
Additional comments about the subscale:				

Subscale construct: COGNITIVE engagement in WHOLE-CLASS INTERACTION

#	In the LECTURE SECTION of this class, during the time in class when your instructor interacts with students addressing the class as a whole, how often have you…	Relevance (1 - 4)	Clarity (1 - 4)	Suggested revisions or other comments

1	Paid attention to what is being said?			
2	Answered in your head or thought about questions your instructor asks the class?			
3	Tried to select information to write down or remember from what is being said?			
4	Tried to follow what is being said?			
5	Tried to put what is being said in your own words?			
6	Tried to connect what is being said with what you know?			
7	Critically thought about what is being said?			
Subscale representativeness (1 - 4):				
Additional comments about the subscale:				

Subscale construct: EMOTIONAL engagement in <u>WHOLE-CLASS INTERACTION</u>				
#	In the LECTURE SECTION of this class, during the time in class when your instructor interacts with students addressing the class as a whole, how often have you…	Relevance (1 - 4)	Clarity (1 - 4)	Suggested revisions or other comments
1	Enjoyed the time in class when your instructor interacts with the students?			
2	Felt interested?			
3	Felt annoyed?			
4	Felt frustrated?			
5	Felt calm?			
6	Felt bored?			
Subscale representativeness (1 - 4):				
Additional comments about the subscale:				

Subscale construct: BEHAVIORAL engagement in <u>INDIVIDUAL WORK</u>				
#	In the LECTURE SECTION of this class, during the time in class when you work on a task without interacting with other students, how often have you...	Relevance (1 - 4)	Clarity (1 - 4)	Suggested revisions or other comments
1	Written down your task solution, answer, or thinking about the task?			
2	Looked at your notes or other resources (e.g., Internet)?			

3	Re-read the task?			
4	Checked your work or answer on the task?			
Subscale representativeness (1 - 4):				
Additional comments about the subscale:				

Subscale construct: COGNITIVE engagement in <u>INDIVIDUAL WORK</u>				
#	In the LECTURE SECTION of this class, during the time in class when you work on a task without interacting with other students, how often have you...	Relevance (1 - 4)	Clarity (1 - 4)	Suggested revisions or other comments
1	Tried to recall from memory the content needed to solve/answer the task?			
2	Critically thought about your task solution, answer, or solution attempts?			
3	Tried to relate the task to what you know?			
4	Verified your work or answer on the task with the task instructions/question?			
5	Tried to select key information from the task?			
Subscale representativeness (1 - 4):				
Additional comments about the subscale:				

Subscale construct: EMOTIONAL engagement in <u>INDIVIDUAL WORK</u>				
#	In the LECTURE SECTION of this class, during the time in class when you work on a task without interacting with other students, how often have you...	Relevance (1 - 4)	Clarity (1 - 4)	Suggested revisions or other comments
1	Enjoyed working on the task on your own?			
2	Felt interested?			
3	Felt annoyed?			
4	Felt frustrated?			
5	Felt calm?			
6	Felt bored?			
Subscale representativeness (1 - 4):				
Additional comments about the subscale:				

Subscale construct: BEHAVIORAL engagement in GROUP WORK				
#	In the LECTURE SECTION of this class, during the time in class when you interact with other students about a task, how often have you...	Relevance (1 - 4)	Clarity (1 - 4)	Suggested revisions or other comments
1	Bounced your ideas about the task off other students?			
2	Compared your and other students' solutions/answers or ways of thinking about the task?			
3	Listened to other students?			
4	Taken notes on other students' thinking about the task or on their solution/answer?			
5	Asked other students a question about the task?			
6	Shared your thinking about the task with other students?			
Subscale representativeness (1 - 4):				
Additional comments about the subscale:				

Subscale construct: COGNITIVE engagement in GROUP WORK				
#	In the LECTURE SECTION of this class, during the time in class when you interact with other students about a task, how often have you...	Relevance (1 - 4)	Clarity (1 - 4)	Suggested revisions or other comments
1	Paid attention to what other students are saying?			
2	Tried to select relevant information from what other students are saying?			
3	Tried to connect other students' thinking about the task, their solutions, or answers to your own?			
4	Tried to use other students' ideas, solutions, or answers in your thinking about the task?			
5	Tried to follow what other students are saying about the task?			
6	Critically thought about other students' thinking about the task, their solutions, or answers?			
Subscale representativeness (1 - 4):				
Additional comments about the subscale:				

Subscale construct: EMOTIONAL engagement in GROUP WORK

#	In the LECTURE SECTION of this class, during the time in class when you interact with other students about a task, how often have you...	Relevance (1 - 4)	Clarity (1 - 4)	Suggested revisions or other comments
1	Enjoyed interacting with other students about the task?			
2	Felt interested?			
3	Felt annoyed?			
4	Felt frustrated?			
5	Felt calm?			
6	Felt bored?			
Subscale representativeness (1 - 4):				
Additional comments about the subscale:				

RESPONSE OPTIONS to engagement questions

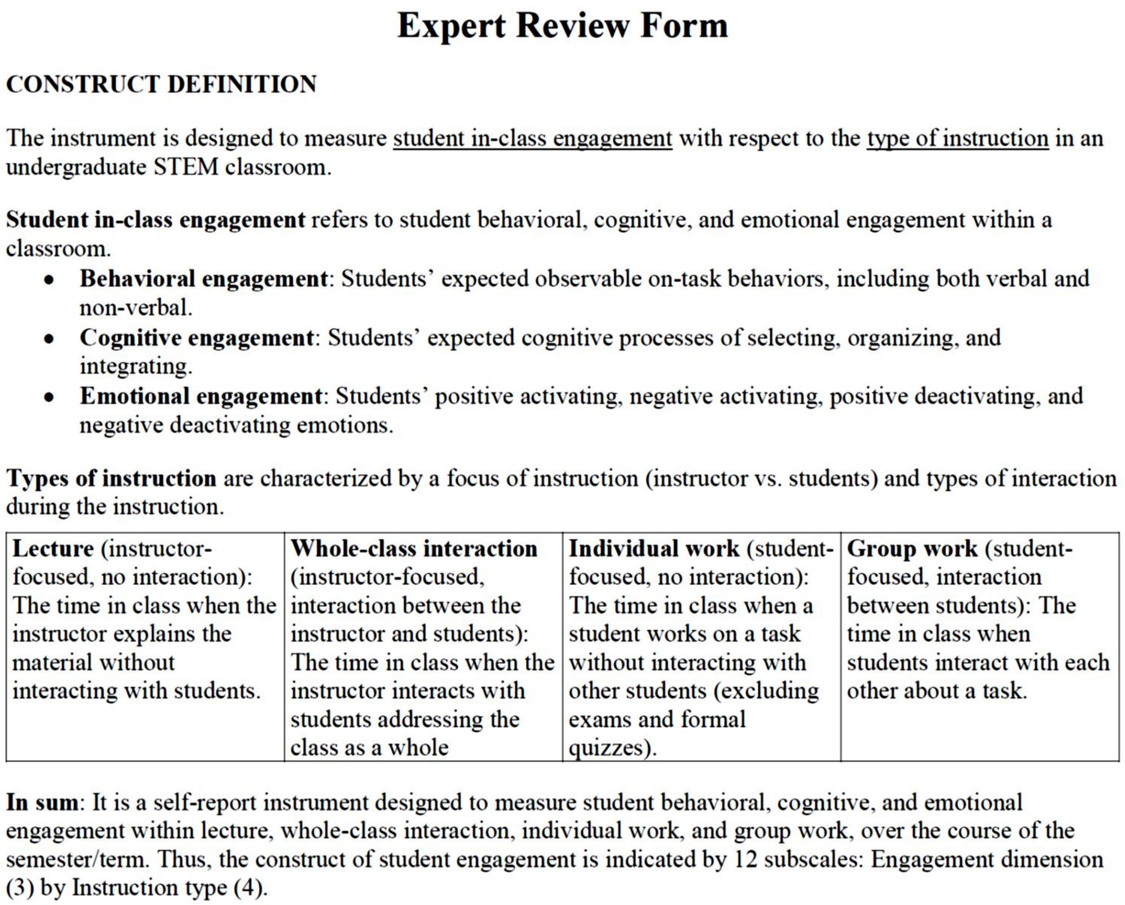
Never or almost never	Rarely	Sometimes	Often	Always or almost always
1	2	3	4	5

Comments about the response options:

Additional comments about the overall scale:

Appendix G

Expert Review Form (Round 2)



Expert Review Form

CONSTRUCT DEFINITION

The instrument is designed to measure student in-class engagement with respect to the type of instruction in an undergraduate STEM classroom.

Student in-class engagement refers to student behavioral, cognitive, and emotional engagement within a classroom.

- **Behavioral engagement**: Students' expected observable on-task behaviors, including both verbal and non-verbal.
- **Cognitive engagement**: Students' expected cognitive processes of selecting, organizing, and integrating.
- **Emotional engagement**: Students' positive activating, negative activating, positive deactivating, and negative deactivating emotions.

Types of instruction are characterized by a focus of instruction (instructor vs. students) and types of interaction during the instruction.

Lecture (instructor-focused, no interaction): The time in class when the instructor explains the material without interacting with students.	**Whole-class interaction** (instructor-focused, interaction between the instructor and students): The time in class when the instructor interacts with students addressing the class as a whole	**Individual work** (student-focused, no interaction): The time in class when a student works on a task without interacting with other students (excluding exams and formal quizzes).	**Group work** (student-focused, interaction between students): The time in class when students interact with each other about a task.

In sum: It is a self-report instrument designed to measure student behavioral, cognitive, and emotional engagement within lecture, whole-class interaction, individual work, and group work, over the course of the semester/term. Thus, the construct of student engagement is indicated by 12 subscales: Engagement dimension (3) by Instruction type (4).

USE OF THE INSTRUMENT

To be valid and, therefore, useful, scores on an educational or psychological measure need to have clearly stated interpretations and uses (AERA, APA, & NCME, 2014; Kane, 2006). This instrument is designed to be used by researchers and practitioners (instructors teaching undergraduate STEM courses).

- Researchers may use the instrument to advance theories of teaching and learning through foundational and/or discipline-based educational research (DBER). Connecting engagement to instruction, the instrument will allow researchers to develop a more comprehensive picture of student engagement in a classroom as well as to investigate the impact of particular teaching practices on student engagement.
- Practitioners may use the instrument for the purposes of teaching improvement: learning about the levels of student engagement in a classroom and changing instruction to increase engagement.

EXPERT REVIEW INSTRUCTIONS

In the form below, you will be asked to review the questions one by one and rate each question on the following:

- **Relevance:** The question reflects, samples, and measures the construct.
- **Clarity:** The question is well written and easy to understand; the wording is clear; the length is appropriate.

After each subscale, you will be asked to rate the subscale on the following:

- **Representativeness:** All facets of the construct are included.

We ask you to rate question relevance and clarity, as well as subscale representativeness, on the following scale:

1 = Not Acceptable (major modifications needed)	2 = Below Expectations (some modifications needed)	3 = Meets Expectations (no modifications needed but could be improved with minor changes)	4 = Exceeds Expectations (no modifications needed)

Additionally, you will be asked to note potential problems, revisions, or other comments about a question or subscale.

STRUCTURE OF THE INSTRUMENT

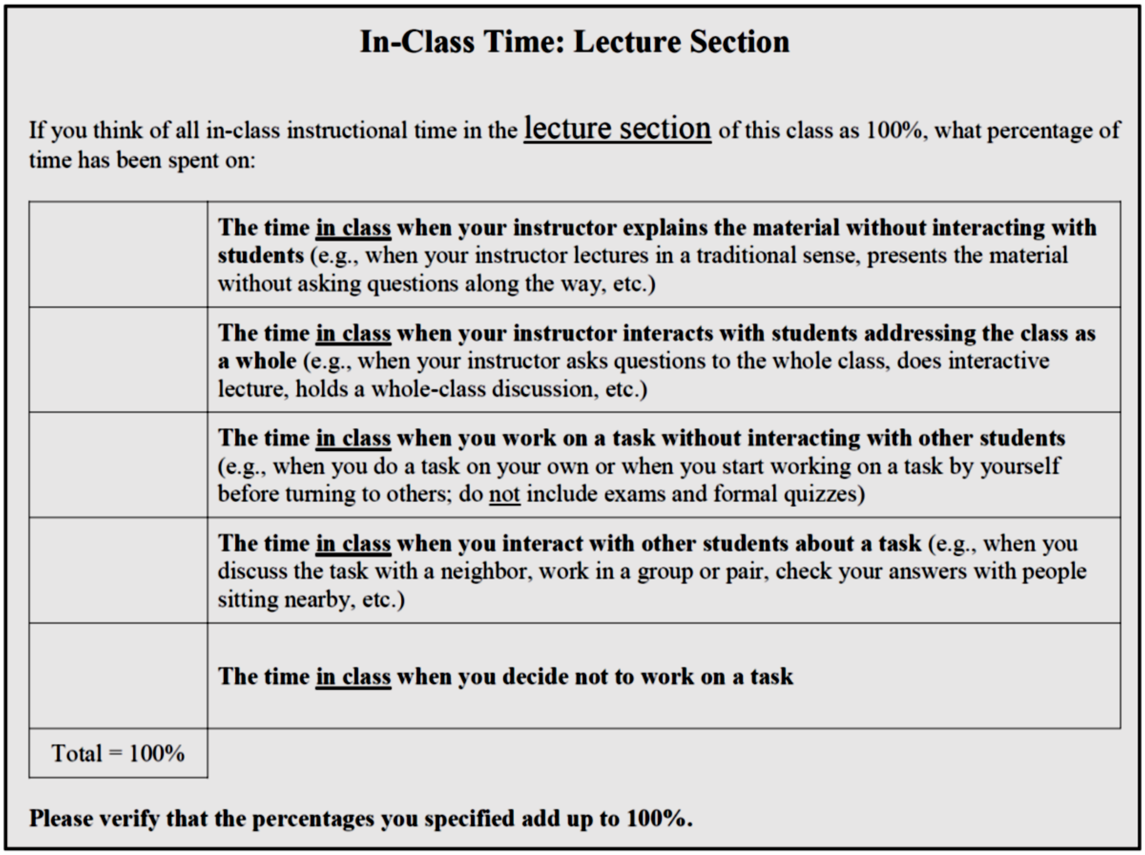
Part 1: In-class time

Students are asked to report the percentage of instructional time spent on each instruction type: lecture, whole-class interaction, individual work, and group work, as well as the percentage of time when they decide not to work on a task. The last category is added to account for the time when the students do not work on the task either individually or in a group.

This information serves two purposes:

- Exclusion criteria for engagement subscales (i.e., if no time was spent on a particular instruction type, then a student will skip the section with engagement questions for this instruction type).
- Weights to compute scores of global engagement (i.e., engagement aggregated across subscales).

Note: The time spent on the student-focused instruction types is not expected to be necessarily similar across students in the same class, as how and whether to work on a task is ultimately at students' discretion.



In-Class Time: Lecture Section

If you think of all in-class instructional time in the lecture section of this class as 100%, what percentage of time has been spent on:

	The time in class when your instructor explains the material without interacting with students (e.g., when your instructor lectures in a traditional sense, presents the material without asking questions along the way, etc.)
	The time in class when your instructor interacts with students addressing the class as a whole (e.g., when your instructor asks questions to the whole class, does interactive lecture, holds a whole-class discussion, etc.)
	The time in class when you work on a task without interacting with other students (e.g., when you do a task on your own or when you start working on a task by yourself before turning to others; do not include exams and formal quizzes)
	The time in class when you interact with other students about a task (e.g., when you discuss the task with a neighbor, work in a group or pair, check your answers with people sitting nearby, etc.)
	The time in class when you decide not to work on a task
Total = 100%	

Please verify that the percentages you specified add up to 100%.

Clarity (1 - 4):	
Suggested revisions or other comments:	

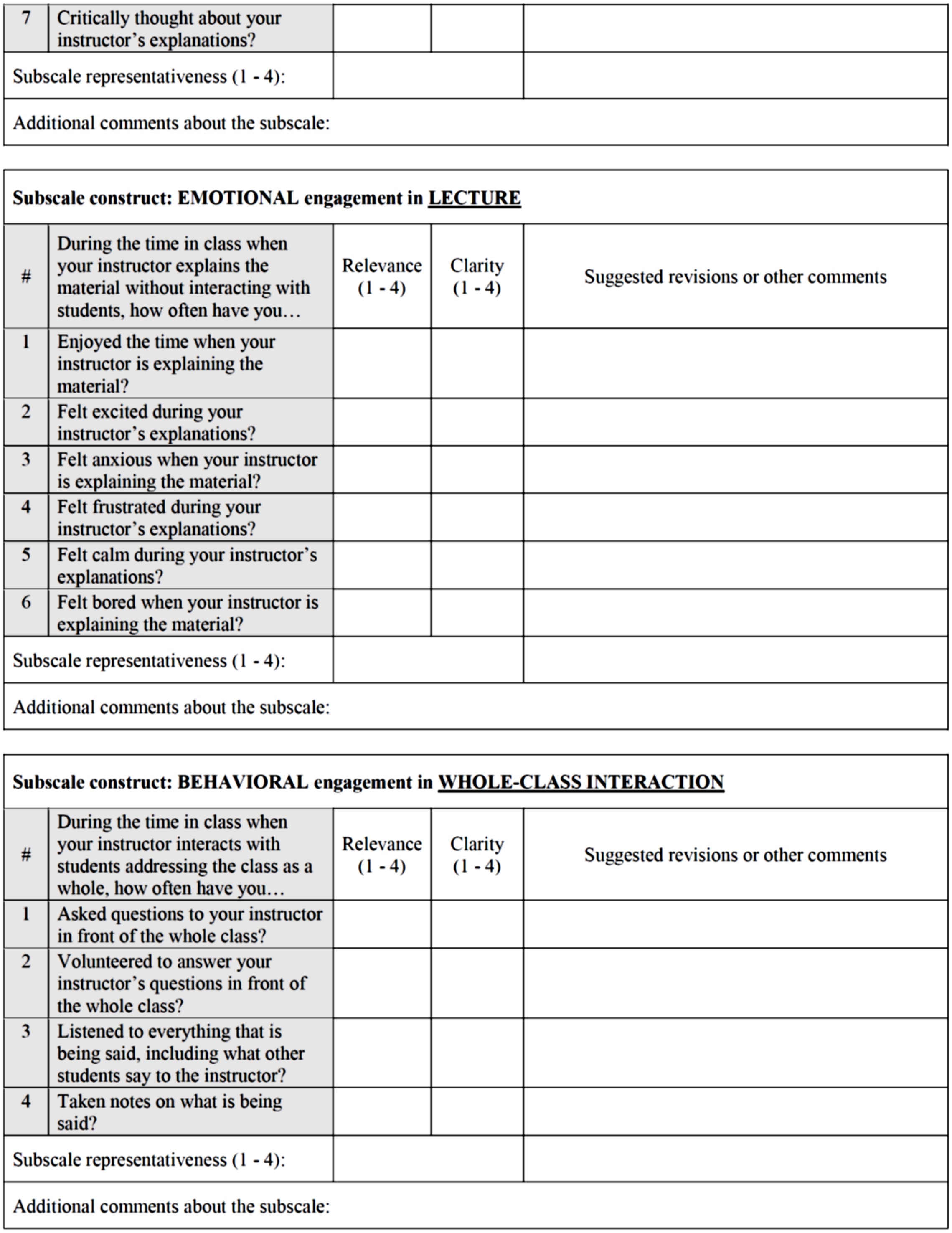
Part 2: Student Engagement Scale

The Student Engagement Scale consists of questions about student engagement. These questions are split into four sections that correspond to the instruction types. Each section has one stem and multiple questions covering behavioral, cognitive, and emotional engagement dimensions.

Response options are as follows:
1 - Never or almost never, 2 - Rarely, 3 - Sometimes, 4 - Often, and 5 - Always or almost always.

Subscale construct: BEHAVIORAL engagement in LECTURE				
#	During the time in class when your instructor explains the material without interacting with students, how often have you…	Relevance (1 - 4)	Clarity (1 - 4)	Suggested revisions or other comments
1	Listened to all of your instructor's explanations?			
2	Taken notes on what your instructor is explaining?			
3	Read all of what the instructor is writing or showing (e.g., instructor's notes, PowerPoint slides, etc.)?			
Subscale representativeness (1 - 4):				
Additional comments about the subscale:				

Subscale construct: COGNITIVE engagement in LECTURE				
#	During the time in class when your instructor explains the material without interacting with students, how often have you…	Relevance (1 - 4)	Clarity (1 - 4)	Suggested revisions or other comments
1	Given your full attention to what your instructor is explaining?			
2	Tried to identify important information to write down or remember from what your instructor is explaining?			
3	Tried to follow all of your instructor's explanations?			
4	Tried to put what your instructor is explaining into your own words?			
5	Tried to connect what your instructor is explaining with what you know?			
6	Tried to make up your own examples or applications of the material?			



7	Critically thought about your instructor’s explanations?			
Subscale representativeness (1 - 4):				
Additional comments about the subscale:				

Subscale construct: EMOTIONAL engagement in <u>LECTURE</u>				
#	During the time in class when your instructor explains the material without interacting with students, how often have you…	Relevance (1 - 4)	Clarity (1 - 4)	Suggested revisions or other comments
1	Enjoyed the time when your instructor is explaining the material?			
2	Felt excited during your instructor’s explanations?			
3	Felt anxious when your instructor is explaining the material?			
4	Felt frustrated during your instructor’s explanations?			
5	Felt calm during your instructor’s explanations?			
6	Felt bored when your instructor is explaining the material?			
Subscale representativeness (1 - 4):				
Additional comments about the subscale:				

Subscale construct: BEHAVIORAL engagement in <u>WHOLE-CLASS INTERACTION</u>				
#	During the time in class when your instructor interacts with students addressing the class as a whole, how often have you…	Relevance (1 - 4)	Clarity (1 - 4)	Suggested revisions or other comments
1	Asked questions to your instructor in front of the whole class?			
2	Volunteered to answer your instructor’s questions in front of the whole class?			
3	Listened to everything that is being said, including what other students say to the instructor?			
4	Taken notes on what is being said?			
Subscale representativeness (1 - 4):				
Additional comments about the subscale:				

Subscale construct: COGNITIVE engagement in WHOLE-CLASS INTERACTION				
#	During the time in class when your instructor interacts with students addressing the class as a whole, how often have you…	Relevance (1 - 4)	Clarity (1 - 4)	Suggested revisions or other comments
1	Given your full attention to what is being said?			
2	Answered in your head or thought about questions your instructor asks the class?			
3	Tried to identify important information to write down or remember from what is being said?			
4	Tried to follow everything that is being said?			
5	Tried to put what is being said in your own words?			
6	Tried to connect what is being said with what you know?			
7	Critically thought about what is being said?			
Subscale representativeness (1 - 4):				
Additional comments about the subscale:				

Subscale construct: EMOTIONAL engagement in WHOLE-CLASS INTERACTION				
#	During the time in class when your instructor interacts with students addressing the class as a whole, how often have you…	Relevance (1 - 4)	Clarity (1 - 4)	Suggested revisions or other comments
1	Enjoyed the time when your instructor interacts with the class?			
2	Felt excited during the interaction between your instructor and the class?			
3	Felt anxious during your instructor's interactions with the class?			
4	Felt frustrated during the time your instructor interacts with the class?			
5	Felt calm while your instructor interacts with the class?			
6	Felt bored when your instructor interacts with the class?			
Subscale representativeness (1 - 4):				
Additional comments about the subscale:				

Subscale construct: BEHAVIORAL engagement in INDIVIDUAL WORK				
#	During the time in class when you work on a task without interacting with other students (excluding exams and formal quizzes), how often have you...	Relevance (1 - 4)	Clarity (1 - 4)	Suggested revisions or other comments
1	Written down in detail your task solution or thinking about the task?			
2	Looked at your notes or other resources (e.g., Internet)?			
3	Re-read the task before trying to solve or answer it?			
4	Tried different ways of solving or thinking about the task even if you already have an answer?			
Subscale representativeness (1 - 4):				
Additional comments about the subscale:				

Subscale construct: COGNITIVE engagement in INDIVIDUAL WORK				
#	During the time in class when you work on a task without interacting with other students (excluding exams and formal quizzes), how often have you...	Relevance (1 - 4)	Clarity (1 - 4)	Suggested revisions or other comments
1	Tried to recall from memory the content needed to solve/answer the task?			
2	Critically thought about your task solution, answer, or solution attempts?			
3	Checked that your work or answer on the task fits with the task instructions/question?			
4	Tried to keep the task instructions/question in mind while solving or answering the task?			
5	Tried to identify the most important information from the task?			
6	Tried to put the task instructions/question in your own words?			
7	Tried to make sure you know why you use particular strategies or reasoning to solve or answer the task?			
Subscale representativeness (1 - 4):				

Additional comments about the subscale:

Subscale construct: EMOTIONAL engagement in INDIVIDUAL WORK

#	During the time in class when you work on a task without interacting with other students (excluding exams and formal quizzes), how often have you...	Relevance (1 - 4)	Clarity (1 - 4)	Suggested revisions or other comments
1	Enjoyed working on the task in class on your own?			
2	Felt excited while working on the task in class by yourself?			
3	Felt anxious doing the task in class on your own?			
4	Felt frustrated working on your own in class?			
5	Felt calm doing the task in class by yourself?			
6	Felt bored when you work on the task in class by yourself?			
Subscale representativeness (1 - 4):				

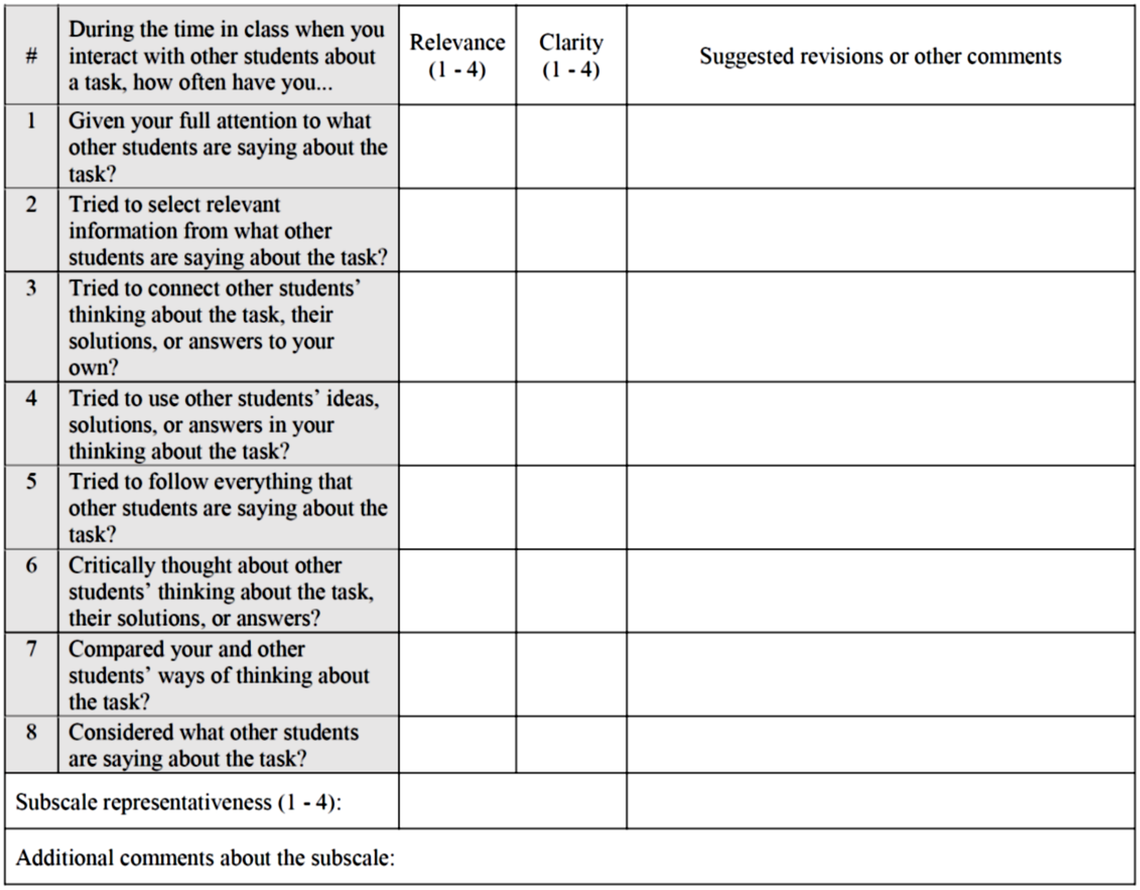
Additional comments about the subscale:

Subscale construct: BEHAVIORAL engagement in GROUP WORK

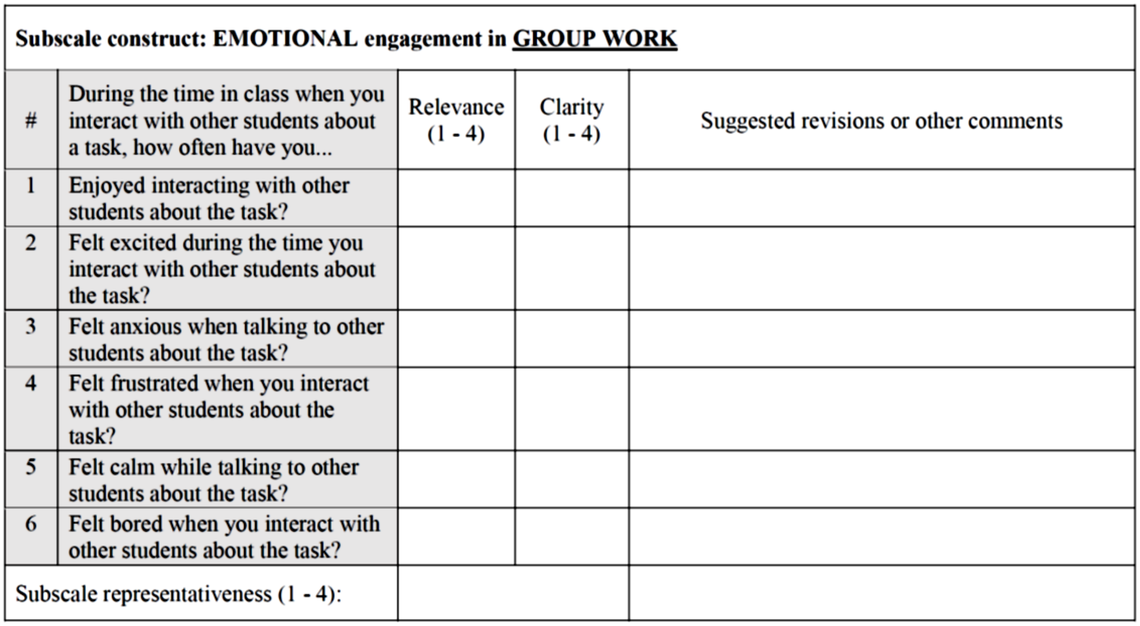
#	During the time in class when you interact with other students about a task, how often have you...	Relevance (1 - 4)	Clarity (1 - 4)	Suggested revisions or other comments
1	Shared your thinking about the task with other students?			
2	Taken notes on other students' thinking about the task or on their solution/answer?			
3	Asked other students about their solutions, answers, or thinking about the task?			
4	Looked at what other students wrote about the task?			
5	Checked with other students to see if your answers, solutions, or approaches match theirs?			
Subscale representativeness (1 - 4):				

Additional comments about the subscale:

Subscale construct: COGNITIVE engagement in GROUP WORK



#	During the time in class when you interact with other students about a task, how often have you...	Relevance (1 - 4)	Clarity (1 - 4)	Suggested revisions or other comments
1	Given your full attention to what other students are saying about the task?			
2	Tried to select relevant information from what other students are saying about the task?			
3	Tried to connect other students' thinking about the task, their solutions, or answers to your own?			
4	Tried to use other students' ideas, solutions, or answers in your thinking about the task?			
5	Tried to follow everything that other students are saying about the task?			
6	Critically thought about other students' thinking about the task, their solutions, or answers?			
7	Compared your and other students' ways of thinking about the task?			
8	Considered what other students are saying about the task?			
Subscale representativeness (1 - 4):				
Additional comments about the subscale:				



Subscale construct: EMOTIONAL engagement in GROUP WORK

#	During the time in class when you interact with other students about a task, how often have you...	Relevance (1 - 4)	Clarity (1 - 4)	Suggested revisions or other comments
1	Enjoyed interacting with other students about the task?			
2	Felt excited during the time you interact with other students about the task?			
3	Felt anxious when talking to other students about the task?			
4	Felt frustrated when you interact with other students about the task?			
5	Felt calm while talking to other students about the task?			
6	Felt bored when you interact with other students about the task?			
Subscale representativeness (1 - 4):				

Additional comments about the subscale:

RESPONSE OPTIONS to engagement questions

Never or almost never	Rarely	Sometimes	Often	Always or almost always
1	2	3	4	5

Comments about the response options:

Additional comments about the overall scale:

STUDENT VERSION of the instrument (please find attached as a separate document)

Comments about the presentation of the instrument to students:

Appendix H

Expert Ratings (Round 1)

Subscale representativeness:

	Expert #1	Expert #2	Expert #3
Behavioral Engagement in Lecture (LB)	3	3	
Cognitive Engagement in Lecture (LC)	3	3	
Emotional Engagement in Lecture (LE)		3	
Behavioral Engagement in Whole-Class Interaction (WB)	2	3	
Cognitive Engagement in Whole-Class Interaction (WC)	3	3	
Emotional Engagement in Whole-Class Interaction (WE)	3	3	
Behavioral Engagement in Individual Work (IB)	2	2	
Cognitive Engagement in Individual Work (IC)	3	3	
Emotional Engagement in Individual Work (IE)		3	
Behavioral Engagement in Group Work (GB)	2	2	
Cognitive Engagement in Group Work (GC)	3	3	
Emotional Engagement in Group Work (GE)	3	3-4	

Item Relevance:

	Expert #1	Expert #2	Expert #3
Behavioral Engagement in Lecture (LB)			
Listened to your instructor’s explanations?	3	3	4
Taken notes on what your instructor is explaining?	3	3	4
Read what the instructor is writing or showing (e.g., instructor’s notes, PowerPoint slides, etc.)?	3	3	4
Cognitive Engagement in Lecture (LC)			
Paid attention to what your instructor is explaining?	2	2	2
Tried to select information to write down or remember from what your instructor is explaining?	2	3	2

	Expert #1	Expert #2	Expert #3
Tried to follow your instructor’s explanations?	3	2	2
Tried to put what your instructor is explaining into your own words?	3	4	3
Tried to connect what your instructor is explaining with what you know?	4	4	4
Tried to make up your own examples or applications of the material?	4	4	4
Critically thought about your instructor’s explanations?	4	4	3
Emotional Engagement in Lecture (LE)			
Enjoyed listening to your instructor?	4	3	4
Felt interested?	4	4	4
Felt annoyed?	4	2	4
Felt frustrated?	4	4	4
Felt calm?	4	4	3
Felt bored?	4	4	4
Behavioral Engagement in Whole-Class Interaction (WB)			
Posed questions to your instructor?	3	2	4
Been willing to answer your instructor’s questions?	4	3	3
Listened to what is being said?	2	4	4
Taken notes on what is being said?	2	3	4
Cognitive Engagement in Whole-Class Interaction (WC)			
Paid attention to what is being said?	3	2	
Answered in your head or thought about questions your instructor asks the class?	3	3	
Tried to select information to write down or remember from what is being said?	3	3	
Tried to follow what is being said?	3	2	
Tried to put what is being said in your own words?	3	4	
Tried to connect what is being said with what you know?	3	4	
Critically thought about what is being said?	3	4	
Emotional Engagement in Whole-Class Interaction (WE)			
Enjoyed the time in class when your instructor interacts with the students?	4	3	4
Felt interested?	4	4	4
Felt annoyed?	4	3	4
Felt frustrated?	4	4	4
Felt calm?	4	4	3

	Expert #1	Expert #2	Expert #3
Felt bored?	4	4	4
Behavioral Engagement in Individual Work (IB)			
Written down your task solution, answer, or thinking about the task?	4	2	4
Looked at your notes or other resources (e.g., Internet)?	4	3	4
Re-read the task?	4	2	4
Checked your work or answer on the task?	2	2	4
Cognitive Engagement in Individual Work (IC)			
Tried to recall from memory the content needed to solve/answer the task?	3	3	4
Critically thought about your task solution, answer, or solution attempts?	3	3	2
Tried to relate the task to what you know?	4	4	4
Verified your work or answer on the task with the task instructions/question?	2	3	2
Tried to select key information from the task?	3	3	3
Emotional Engagement in Individual Work (IE)			
Enjoyed working on the task on your own?	4	3	4
Felt interested?	4	4	4
Felt annoyed?	4	2	4
Felt frustrated?	4	3	4
Felt calm?	4	3	3
Felt bored?	4	4	4
Behavioral Engagement in Group Work (GB)			
Bounced your ideas about the task off other students?	2	2	3
Compared your and other students' solutions/answers or ways of thinking about the task?	2	2	3
Listened to other students?	3	4	4
Taken notes on other students' thinking about the task or on their solution/answer?	4	3	4
Asked other students a question about the task?	4	2	4
Shared your thinking about the task with other students?	3	2	4
Cognitive Engagement in Group Work (GC)			
Paid attention to what other students are saying?	3	2	2
Tried to select relevant information from what other students are saying?	3	3	4
Tried to connect other students' thinking about the task, their solutions, or answers to your own?	3	4	4

	Expert #1	Expert #2	Expert #3
Tried to use other students' ideas, solutions, or answers in your thinking about the task?	4	4	4
Tried to follow what other students are saying about the task?	4	2	3
Critically thought about other students' thinking about the task, their solutions, or answers?	3	4	3
Emotional Engagement in Group Work (GE)			
Enjoyed interacting with other students about the task?	4	3	4
Felt interested?	4	4	4
Felt annoyed?	4	4	4
Felt frustrated?	4	4	4
Felt calm?	4	3	3
Felt bored?	4	4	4

Item Clarity:

	Expert #1	Expert #2	Expert #3
Instructional Time Form	2	3	
Behavioral Engagement in Lecture (LB)			
Listened to your instructor's explanations?	4	3	4
Taken notes on what your instructor is explaining?	4	3	4
Read what the instructor is writing or showing (e.g., instructor's notes, PowerPoint slides, etc.)?	4	3-4	4
Cognitive Engagement in Lecture (LC)			
Paid attention to what your instructor is explaining?	3	3	4
Tried to select information to write down or remember from what your instructor is explaining?	3	3	3
Tried to follow your instructor's explanations?	3	3	2
Tried to put what your instructor is explaining into your own words?	4	3	4
Tried to connect what your instructor is explaining with what you know?	4	3	4
Tried to make up your own examples or applications of the material?	4	3	4
Critically thought about your instructor's explanations?	4	3	2
Emotional Engagement in Lecture (LE)			
Enjoyed listening to your instructor?	4	2	4
Felt interested?	4	4	3
Felt annoyed?	4	3	5
Felt frustrated?	4	3	3

	Expert #1	Expert #2	Expert #3
Felt calm?	4	4	3
Felt bored?	4	4	3
Behavioral Engagement in Whole-Class Interaction (WB)			
Posed questions to your instructor?	3	3	3
Been willing to answer your instructor's questions?	2	3	3
Listened to what is being said?	3	3	4
Taken notes on what is being said?	3	3	4
Cognitive Engagement in Whole-Class Interaction (WC)			
Paid attention to what is being said?	2	3	
Answered in your head or thought about questions your instructor asks the class?	2	2	
Tried to select information to write down or remember from what is being said?	3	3	
Tried to follow what is being said?	3	3	
Tried to put what is being said in your own words?	3	3	
Tried to connect what is being said with what you know?	3	3	
Critically thought about what is being said?	3	3	
Emotional Engagement in Whole-Class Interaction (WE)			
Enjoyed the time in class when your instructor interacts with the students?	4	2	4
Felt interested?	4	4	4
Felt annoyed?	4	3	4
Felt frustrated?	4	3	4
Felt calm?	4	4	4
Felt bored?	4	4	4
Behavioral Engagement in Individual Work (IB)			
Written down your task solution, answer, or thinking about the task?	2	3	4
Looked at your notes or other resources (e.g., Internet)?	4	3	4
Re-read the task?	4	2	4
Checked your work or answer on the task?	2	3	4
Cognitive Engagement in Individual Work (IC)			
Tried to recall from memory the content needed to solve/answer the task?	3	3	4
Critically thought about your task solution, answer, or solution attempts?	2	3	3
Tried to relate the task to what you know?	4	4	4

	Expert #1	Expert #2	Expert #3
Verified your work or answer on the task with the task instructions/question?	2	3	2
Tried to select key information from the task?	2	1-2	3
Emotional Engagement in Individual Work (IE)			
Enjoyed working on the task on your own?	4	3	4
Felt interested?	3	4	4
Felt annoyed?	3	3	4
Felt frustrated?	3	3	4
Felt calm?	3	3	4
Felt bored?	3	4	4
Behavioral Engagement in Group Work (GB)			
Bounced your ideas about the task off other students?	2	3	2
Compared your and other students' solutions/answers or ways of thinking about the task?	4	3	4
Listened to other students?	3	4	4
Taken notes on other students' thinking about the task or on their solution/answer?	4	3	4
Asked other students a question about the task?	4	4	4
Shared your thinking about the task with other students?	3	3	4
Cognitive Engagement in Group Work (GC)			
Paid attention to what other students are saying?	3	3	4
Tried to select relevant information from what other students are saying?	3	3	4
Tried to connect other students' thinking about the task, their solutions, or answers to your own?	3	4	4
Tried to use other students' ideas, solutions, or answers in your thinking about the task?	4	4	4
Tried to follow what other students are saying about the task?	4	3	2
Critically thought about other students' thinking about the task, their solutions, or answers?	3	4	2
Emotional Engagement in Group Work (GE)			
Enjoyed interacting with other students about the task?	4	3	4
Felt interested?	3	4	4
Felt annoyed?	3	4	4
Felt frustrated?	3	4	4
Felt calm?	3	3	4
Felt bored?	3	4	4

Comments:

	Expert #1	Expert #2	Expert #3
Instructional Time Form	The instructions are a little unclear. Are there different "sections" of the course? This is the part that I am unsure of "all in-class instructional time in the lecture section of this class as 100%". Why the in-class instructional time designation and then the lecture section? It seems redundant. Why not just say include a section in the percentage that is non-instructional time (e.g., passing back exams) and then just refer them to the course in total or the total time spent in class? I also think students might get hung up on the distinction between lecture vs. whole-class instruction. You may wish to provide examples of what you mean by these. Does posing a question to the class count as whole-class instruction because it's instructor initiated? If a student raised a hand and asked a question during lecture, because it's student focused/initiated, would that be lecture or whole-class instruction? I find this distinction to be somewhat artificial and likely difficult for students to distinguish among. However, a bigger question might be, what is the utility of the distinction for instructors?	• Inclusion of investigation, labs or design (STEM focused) • Examples might support clarity • Where would students classify note taking? Is note taking part of individual work and potentially concurrent with lecture?	In last question, not sure about the word decide. This is making value judgement that engagement is choice. I would just say The time in class when you are not working on a task
Behavioral Engagement in Lecture (LB)		Inclusion of attention from cognitive engagement subscale; consider focus at	

	Expert #1	Expert #2	Expert #3
		front of the room as an indicator	
Listened to your instructor's explanations?		? use presentation	
Taken notes on what your instructor is explaining?			
Read what the instructor is writing or showing (e.g., instructor's notes, PowerPoint slides, etc.)?	Does showing video count here?	Helpful examples	
Cognitive Engagement in Lecture (LC)	The use of different wording for one item in this scale introduces bias that could influence measurement properties. If tried is the wording for most items, you may wish to use that for all items or vary more consistently across items.	Following along and attention sound more like behavioral engagement	Not clear why some of these items are indicators of cognitive rather than behavioral engagement
Paid attention to what your instructor is explaining?	This doesn't seem that different from the 3 items in behavioral engagement. Maybe something more like "worked to understand" or "given your full attention".	Behavioral engagement	In many scales, paying attention is a measure of behavioral engagement. Why is this considered cognitive?
Tried to select information to write down or remember from what your instructor is explaining?	Again – what's the real difference between taking notes and trying to select information to write down? How does one take notes without selecting information to write down? If the intent is to see to what extent students are filtering for critical information, then I suggest a more direct wording to that effect, such as "tried to figure out the most important pieces	Nice!	This seems more like a function of type of lecture than cognitive engagement

	Expert #1	Expert #2	Expert #3
	of information to write down" or something to that effect.		
Tried to follow your instructor's explanations?		Seems more behavioral	Why is this cognitive?
Tried to put what your instructor is explaining into your own words?		Potentially difficult to do *during* lecture	
Tried to connect what your instructor is explaining with what you know?			
Tried to make up your own examples or applications of the material?		"come up with"	
Critically thought about your instructor's explanations?			I think this would be hard for a student to understand. What does critically mean?
Emotional Engagement in Lecture (LE)	The use of different wording for the first item in this scale introduces bias that could influence measurement properties. Either use consistently or provide more variation within the scale but in a consistent manner. E.g., felt…. Or alternatively….experienced….or…been. Consider adding anxious or nervous.	Consider anxiety (difficulty of material; pace). Item 1 could be about instructor characteristics (humor) or about course material – does that matter? Format difference. Annoyed might elicit social	Need to make sure participant is answering about lecture
Enjoyed listening to your instructor?	"Felt enjoyment"….	Enjoyment of material/content or	

	Expert #1	Expert #2	Expert #3
		instructor?; format is different from remaining items ("experienced enjoyment?"	
Felt interested?			Add in the lecture
Felt annoyed?		Do we care if related to other students vs. teacher?	Add in the lecture
Felt frustrated?		Frustration can be a positive	Add in the lecture
Felt calm?			Add in the lecture
Felt bored?			Add in the lecture
Behavioral Engagement in Whole-Class Interaction (WB)	I think this scale needs work to be distinct from the lecture items. But, this is a point I bring up in my overall comments above. Could there also be ways that students provide feedback to the instructor in other ways, such as use of clickers, etc. that are not captured here?	Consider focused on speaker/other students' contributions	
Posed questions to your instructor?	Posed might be unfamiliar language. Might want to go with "asked".	Similar to item 2 from cognitive engagement; could be cognitive or agentic	
Been willing to answer your instructor's questions?	Willingness to answer questions and demonstrating that behaviorally are not the same. I can sit there and be willing if the instructor points me out, but still not raise my hand or give any indication that I'm "willing". I'd split this into two questions if you're interesting in (a) would I? and (b) did I take		This is predicated on the instructor asking questions

	Expert #1	Expert #2	Expert #3
	steps to indicate that willingness, such as raise my hand, etc.?		
Listened to what is being said?	This is too non-specific to be distinct from the lecture items. See comment below.		
Taken notes on what is being said?	For this to be distinct from the lecture series, perhaps the focus should be on the attention to the interaction. For example, “taken notes on the exchange between the instructor and students”. A similar critique could be made to item 3 in this series.		
Cognitive Engagement in Whole-Class Interaction (WC)		Posed questions	Similar numbers and comments to the lecturer section
Paid attention to what is being said?	I’d change the what is being said to “what is being said during the interaction” or “exchange”.	Behavioral engagement	
Answered in your head or thought about questions your instructor asks the class?	How is this distinct from the behavioral engagement items? Is this the mental effort piece? I think the second part of the question is more cognitive engagement as compared to the behavioral aspect of answering. Need some clarity between this item with BE to be clear on what is the observable behavior and what is the mental effort.	Check grammar; seems like 2 items in 1	
Tried to select information to write down or remember from what is being said?	See the notes elsewhere on “what is being said”, as a little vague.		
Tried to follow what is being	See the notes elsewhere on “what is being	Behavioral engagement	

	Expert #1	Expert #2	Expert #3
said?	said", as a little vague		
Tried to put what is being said in your own words?	See the notes elsewhere on "what is being said", as a little vague	Works better in this instructional format	
Tried to connect what is being said with what you know?	See the notes elsewhere on "what is being said", as a little vague		
Critically thought about what is being said?	See the notes elsewhere on "what is being said", as a little vague		
Emotional Engagement in Whole-Class Interaction (WE)	Consider adding worried, anxious, or nervous. See other notes about use of phrasing.	Consider anxious	Similar numbers and comments to the lecturer section.
Enjoyed the time in class when your instructor interacts with the students?	I like this wording about "instructor interacts with the students"	"other students"; format difference	
Felt interested?			
Felt annoyed?		Better fit in this scale for social interaction	
Felt frustrated?			
Felt calm?			This has not been included in other scales. Calm could be both indicative of engagement/disengagement
Felt bored?			
Behavioral Engagement in Individual Work (IB)	Seeking help from the instructor or other students is a reasonable activity during individual work. Also consider asking about staying on task during the entire period given	Several items seem more cognitive	

	Expert #1	Expert #2	Expert #3
	for the activity. Could also be related to using diagramming or notetaking tools.		
Written down your task solution, answer, or thinking about the task?	There is too much going on in this question. Consider splitting out into writing down the solution vs. writing out the process.	Cognitive; consider, “did the task you were assigned”	
Looked at your notes or other resources (e.g., Internet)?			
Re-read the task?		Task requirements or Ss response draft?	
Checked your work or answer on the task?	I don’t care for this question because it implies that an answer is available. It’s not clear that is the case.	Cognitive/regulation	
Cognitive Engagement in Individual Work (IC)	Could consider elaboration a bit more, such as the use of concept mapping or other tools to organize information.	Include revision; consider checking your understanding of task information; item 4 from behavioral engagement	
Tried to recall from memory the content needed to solve/answer the task?		Without using notes? Consider “restate”	
Critically thought about your task solution, answer, or solution attempts?	Is this review of your work? There is a qualitative distinction between reviewing the work for accuracy, etc., which requires some level of critique and searching for alternative explanations or critiques to the solution that the student has arrived at.	Dependent on type of instruction/task	What does critically mean? I am not sure students would interpret the same way
Tried to relate the task to what you know?			

	Expert #1	Expert #2	Expert #3
Verified your work or answer on the task with the task instructions/question?	How is this different than item 4 in BE?	Rephrase, "checked your work….fit with the task instructions?" responded to?	How is this different than item, checked your work? (Above)
Tried to select key information from the task?	Is this metacognition in that the student is trying to determine why the instructor assigned this task? This question isn't clear.	Unclear, toward what ends? For future learning?	
Emotional Engagement in Individual Work (IE)	Carry consistent language. Consider adding anxious, etc.		
Enjoyed working on the task on your own?		Distinct format; exclude "on your own"	
Felt interested?	Add "in the task". If you want to make these emotions specific to the action in the classroom, then you may wish to finish the sentence with the focal unit.		
Felt annoyed?		At what?	
Felt frustrated?			
Felt calm?			
Felt bored?			
Behavioral Engagement in Group Work (GB)	Many of these items seem like cognitive engagement. The behavioral part of group work, to me, would be more like active participation in discussion, helped to set group work rules, worked on group documents, located resources to help with the group task, and the like.	More behavioral might be, "worked with other students on the task"	
Bounced your ideas about the	This seems like cognitive engagement. This is	Cognitive engagement	Not sure what mean by

	Expert #1	Expert #2	Expert #3
task off other students?	basically brainstorming, correct? If so, that suggests a mental effort. This is kind of a fine line. I think of it as quality of behavioral participation (e.g., how does your group know that you are an active participant) as compared to quality of mental participation. What does this mean behaviorally?		bounce. Share is better word, but then not sure how different than #5
Compared your and other students' solutions/answers or ways of thinking about the task?	This seems like cognitive engagement.	Cognitive engagement	
Listened to other students?		Key item	
Taken notes on other students' thinking about the task or on their solution/answer?		Second phrase could be low quality – such as copying	
Asked other students a question about the task?		Cognitive engagement	
Shared your thinking about the task with other students?		Cognitive. Consider rephrasing as "participated"	
Cognitive Engagement in Group Work (GC)	This is all a bit unidirectional – the student is processing the input from the other students rather than generating or initiating to help elevate the group solution.	Consider items currently under behavioral	
Paid attention to what other students are saying?		Behavioral	This could also be indicator of behavioral engagement
Tried to select relevant information from what other students are saying?			

	Expert #1	Expert #2	Expert #3
Tried to connect other students' thinking about the task, their solutions, or answers to your own?		Nice item	
Tried to use other students' ideas, solutions, or answers in your thinking about the task?			
Tried to follow what other students are saying about the task?	I prefer this to the more passive "paid attention".	Behavioral	How is this different than listening? Not sure what mean by follow
Critically thought about other students' thinking about the task, their solutions, or answers?		Yes!	Not clear what you mean by critically
Emotional Engagement in Group Work (GE)	Consider adding anxious, worried, or nervous.	Anxious	
Enjoyed interacting with other students about the task?		Seems to fit best in this format	
Felt interested?	Add the focal unit to the response.		
Felt annoyed?		Social emotion works well here.	
Felt frustrated?			
Felt calm?			
Felt bored?			
Response options to engagement items	I'm wondering how you will interpret these response options as they are about frequency.		I would suggest deleting almost never and almost

	Expert #1	Expert #2	Expert #3
	To me, engagement is about the quality of participation, mental effort, or emotion. Are you expecting that the quality portion will be contained within the items themselves, even though one could argue that the subscales contain items that reflect different quality. For instance, is always paying attention reflecting an equivalent level of cognitive engagement as always critically thinking about other students' responses. Something to consider….		always, not clear how they are different than 2 and 4
The overall scale	I noted very few negatively worded items to potentially address positive response bias. I would suggest considering the nature of the consistency of the wording within scales to avoid problems with psychometric properties that are related to the use of similar or different wording within scales and also consider revising to either add a couple of negatively worded responses per subscale or reword a sample of existing items.	• Consider providing examples of instructional format o Whole class interactions (e.g., hotseat; discussion; recitation) o Groupwork (includes pair work through groups; excludes all students in a recitation section) • Consider inclusion of inquiry, lab sections or design formats as an instructional format • On the student survey, consider having the N/A option only for the whole scale level, but not the individual item level. There are several instances	

	Expert #1	Expert #2	Expert #3
		where behavioral engagement includes cognitive, and vice versa. This is especially challenging for the independent work subscale	

Appendix I

Expert Ratings (Round 2)

Subscale representativeness:

	Expert #4	Expert #5
Behavioral Engagement in Lecture (LB)	2	4
Cognitive Engagement in Lecture (LC)	3	4
Emotional Engagement in Lecture (LE)	3	
Behavioral Engagement in Whole-Class Interaction (WB)	3	4
Cognitive Engagement in Whole-Class Interaction (WC)	3	4
Emotional Engagement in Whole-Class Interaction (WE)	3	4
Behavioral Engagement in Individual Work (IB)	4	4
Cognitive Engagement in Individual Work (IC)	4	4
Emotional Engagement in Individual Work (IE)	4	4
Behavioral Engagement in Group Work (GB)	4	4
Cognitive Engagement in Group Work (GC)	4	4
Emotional Engagement in Group Work (GE)	4	4

Item relevance:

	Expert #4	Expert #5
Behavioral Engagement in Lecture (LB)		
Listened to all of your instructor’s explanations?	2	4
Taken notes on what your instructor is explaining?	3	4
Read all of what the instructor is writing or showing (e.g., instructor’s notes, PowerPoint slides, etc.)?	3	4
Cognitive Engagement in Lecture (LC)		
Given your full attention to what your instructor is explaining?	4	4
Tried to identify important information to write down or remember from what your instructor is explaining?	4	4
Tried to follow all of your instructor’s explanations?	4	4
Tried to put what your instructor is explaining into your own words?	4	4
Tried to connect what your instructor is explaining with what you know?	4	4
Tried to make up your own examples or applications of the material?	3	4

	Expert #4	Expert #5
Critically thought about your instructor’s explanations?	4	4
Emotional Engagement in Lecture (LE)		
Enjoyed the time when your instructor is explaining the material?	4	4
Felt excited during your instructor’s explanations?	4	4
Felt anxious when your instructor is explaining the material?	4	4
Felt frustrated during your instructor’s explanations?	4	4
Felt calm during your instructor’s explanations?	3	4
Felt bored when your instructor is explaining the material?	4	4
Behavioral Engagement in Whole-Class Interaction (WB)		
Asked questions to your instructor in front of the whole class?	3	4
Volunteered to answer your instructor’s questions in front of the whole class?	3	4
Listened to everything that is being said, including what other students say to the instructor?	4	4
Taken notes on what is being said?	4	4
Cognitive Engagement in Whole-Class Interaction (WC)		
Given your full attention to what is being said?	4	4
Answered in your head or thought about questions your instructor asks the class?	4	4
Tried to identify important information to write down or remember from what is being said?	4	4
Tried to follow everything that is being said?	2	4
Tried to put what is being said in your own words?	3	4
Tried to connect what is being said with what you know?	4	4
Critically thought about what is being said?	4	4
Emotional Engagement in Whole-Class Interaction (WE)		
Enjoyed the time when your instructor interacts with the class?	4	4
Felt excited during the interaction between your instructor and the class?	4	4
Felt anxious during your instructor’s interactions with the class?	4	4
Felt frustrated during the time your instructor interacts with the class?	4	4
Felt calm while your instructor interacts with the class?	3	4
Felt bored when your instructor interacts with the class?	4	4
Behavioral Engagement in Individual Work (IB)		
Written down in detail your task solution or thinking about the task?	4	4
Looked at your notes or other resources (e.g., Internet)?	4	4
Re-read the task before trying to solve or answer it?	4	4
Tried different ways of solving or thinking about the task even if you already have an answer?	3	3

	Expert #4	Expert #5
Cognitive Engagement in Individual Work (IC)		
Tried to recall from memory the content needed to solve/answer the task?	3	4
Critically thought about your task solution, answer, or solution attempts?	4	4
Checked that your work or answer on the task fits with the task instructions/question?	4	4
Tried to keep the task instructions/question in mind while solving or answering the task?	4	4
Tried to identify the most important information from the task?	4	4
Tried to put the task instructions/question in your own words?	4	4
Tried to make sure you know why you use particular strategies or reasoning to solve or answer the task?	4	4
Emotional Engagement in Individual Work (IE)		
Enjoyed working on the task in class on your own?	4	4
Felt excited while working on the task in class by yourself?	4	4
Felt anxious doing the task in class on your own?	4	4
Felt frustrated working on your own in class?	4	4
Felt calm doing the task in class by yourself?	4	4
Felt bored when you work on the task in class by yourself?	4	4
Behavioral Engagement in Group Work (GB)		
Shared your thinking about the task with other students?	4	4
Taken notes on other students’ thinking about the task or on their solution/answer?	4	4
Asked other students about their solutions, answers, or thinking about the task?	4	4
Looked at what other students wrote about the task?	4	4
Checked with other students to see if your answers, solutions, or approaches match theirs?	4	4
Cognitive Engagement in Group Work (GC)		
Given your full attention to what other students are saying about the task?	4	4
Tried to select relevant information from what other students are saying about the task?	4	4
Tried to connect other students’ thinking about the task, their solutions, or answers to your own?	4	4
Tried to use other students’ ideas, solutions, or answers in your thinking about the task?	4	4
Tried to follow everything that other students are saying about the task?	4	4
Critically thought about other students’ thinking about the task, their solutions, or answers?	4	4

	Expert #4	Expert #5
Compared your and other students' ways of thinking about the task?	4	4
Considered what other students are saying about the task?	4	4
Emotional Engagement in Group Work (GE)		
Enjoyed interacting with other students about the task?	4	4
Felt excited during the time you interact with other students about the task?	4	4
Felt anxious when talking to other students about the task?	4	4
Felt frustrated when you interact with other students about the task?	4	4
Felt calm while talking to other students about the task?	4	4
Felt bored when you interact with other students about the task?	4	4

Item clarity:

	Expert #4	Expert #5
Instructional Time Form	3	3.9
Behavioral Engagement in Lecture (LB)		
Listened to all of your instructor's explanations?	3	4
Taken notes on what your instructor is explaining?	3	4
Read all of what the instructor is writing or showing (e.g., instructor's notes, PowerPoint slides, etc.)?	3	4
Cognitive Engagement in Lecture (LC)		
Given your full attention to what your instructor is explaining?	4	4
Tried to identify important information to write down or remember from what your instructor is explaining?	4	4
Tried to follow all of your instructor's explanations?	4	4
Tried to put what your instructor is explaining into your own words?	4	4
Tried to connect what your instructor is explaining with what you know?	4	3
Tried to make up your own examples or applications of the material?	3	4
Critically thought about your instructor's explanations?	4	3
Emotional Engagement in Lecture (LE)		
Enjoyed the time when your instructor is explaining the material?	4	4
Felt excited during your instructor's explanations?	4	4
Felt anxious when your instructor is explaining the material?	4	4
Felt frustrated during your instructor's explanations?	4	4
Felt calm during your instructor's explanations?	3	4
Felt bored when your instructor is explaining the material?	4	4

	Expert #4	Expert #5
Behavioral Engagement in Whole-Class Interaction (WB)		
Asked questions to your instructor in front of the whole class?	3	4
Volunteered to answer your instructor's questions in front of the whole class?	3	4
Listened to everything that is being said, including what other students say to the instructor?	4	4
Taken notes on what is being said?	4	4
Cognitive Engagement in Whole-Class Interaction (WC)		
Given your full attention to what is being said?	4	4
Answered in your head or thought about questions your instructor asks the class?	4	4
Tried to identify important information to write down or remember from what is being said?	4	4
Tried to follow everything that is being said?	2	4
Tried to put what is being said in your own words?	3	4
Tried to connect what is being said with what you know?	4	3
Critically thought about what is being said?	4	3
Emotional Engagement in Whole-Class Interaction (WE)		
Enjoyed the time when your instructor interacts with the class?	4	4
Felt excited during the interaction between your instructor and the class?	4	4
Felt anxious during your instructor's interactions with the class?	4	4
Felt frustrated during the time your instructor interacts with the class?	4	4
Felt calm while your instructor interacts with the class?	3	4
Felt bored when your instructor interacts with the class?	4	4
Behavioral Engagement in Individual Work (IB)		
Written down in detail your task solution or thinking about the task?	4	4
Looked at your notes or other resources (e.g., Internet)?	4	4
Re-read the task before trying to solve or answer it?	4	4
Tried different ways of solving or thinking about the task even if you already have an answer?	3	3
Cognitive Engagement in Individual Work (IC)		
Tried to recall from memory the content needed to solve/answer the task?	3	4
Critically thought about your task solution, answer, or solution attempts?	4	3
Checked that your work or answer on the task fits with the task instructions/question?	4	4
Tried to keep the task instructions/question in mind while solving or answering the task?	3	4

	Expert #4	Expert #5
Tried to identify the most important information from the task?	4	4
Tried to put the task instructions/question in your own words?	4	4
Tried to make sure you know why you use particular strategies or reasoning to solve or answer the task?	4	4
Emotional Engagement in Individual Work (IE)		
Enjoyed working on the task in class on your own?	4	4
Felt excited while working on the task in class by yourself?	4	4
Felt anxious doing the task in class on your own?	4	4
Felt frustrated working on your own in class?	4	4
Felt calm doing the task in class by yourself?	4	4
Felt bored when you work on the task in class by yourself?	4	4
Behavioral Engagement in Group Work (GB)		
Shared your thinking about the task with other students?	4	4
Taken notes on other students' thinking about the task or on their solution/answer?	4	4
Asked other students about their solutions, answers, or thinking about the task?	4	4
Looked at what other students wrote about the task?	4	4
Checked with other students to see if your answers, solutions, or approaches match theirs?	4	4
Cognitive Engagement in Group Work (GC)		
Given your full attention to what other students are saying about the task?	4	4
Tried to select relevant information from what other students are saying about the task?	4	4
Tried to connect other students' thinking about the task, their solutions, or answers to your own?	4	4
Tried to use other students' ideas, solutions, or answers in your thinking about the task?	4	4
Tried to follow everything that other students are saying about the task?	4	4
Critically thought about other students' thinking about the task, their solutions, or answers?	4	4
Compared your and other students' ways of thinking about the task?	4	4
Considered what other students are saying about the task?	4	4
Emotional Engagement in Group Work (GE)		
Enjoyed interacting with other students about the task?	4	4
Felt excited during the time you interact with other students about the task?	4	4
Felt anxious when talking to other students about the task?	4	4
Felt frustrated when you interact with other students about the	4	4

	Expert #4	Expert #5
task?		
Felt calm while talking to other students about the task?	4	4
Felt bored when you interact with other students about the task?	4	4

Comments:

	Expert #4	Expert #5
Instructional Time Form	Some students may have trouble adding percentages. Consider giving them a set number of points (e.g., 10, 20, etc.) to assign to the options instead.	The time in class when you interact with other students "on" a task.
Behavioral Engagement in Lecture (LB)	It seems like negative behaviors should be measured, as well, such as talking to a neighbor, surfing the Web, etc..	
Listened to all of your instructor's explanations?	I'm concerned that social desirability may bias responses. Consider rephrasing the question to make it more difficult to endorse.	Go lower [unclear] on all of these [unclear] they are a continuation of a sentence
Taken notes on what your instructor is explaining?	Frequency is one aspect of this, but what about quality or intensity of notetaking?	
Read all of what the instructor is writing or showing (e.g., instructor's notes, PowerPoint slides, etc.)?	I imagine that responses to this item will be mitigated by instructor behavior, such as rushing through slides or speaking too fast.	
Cognitive Engagement in Lecture (LC)	It may behoove you to explain to respondents what some of these cognitions can look like. For example, I a student is putting the instructor's words into her own when she writes her notes in her own words. Again, there may be some negative indicators that should be included (e.g., daydreaming, spacing out, etc.).	
Given your full attention to what your instructor is explaining?		

Tried to identify important information to write down or remember from what your instructor is explaining?		
Tried to follow all of your instructor's explanations?		
Tried to put what your instructor is explaining into your own words?	I really like this item!	
Tried to connect what your instructor is explaining with what you know?	I like this item, too, but a difficulty I see with both is that some students engage in these cognitions after class, because it is hard to keep up with notes during class.	"what you know" is not clear. Maybe re-write.
Tried to make up your own examples or applications of the material?	Ditto.	
Critically thought about your instructor's explanations?	Again, it depends on instructor pace, or perceived pace.	Instead of "critically" maybe "thought deeply about"
Emotional Engagement in Lecture (LE)	Consider adding an item for apathy.	You have two positive and two negative activating and one positive deactivating and one negative deactivating. Uneven. Might make analysis challenging.
Enjoyed the time when your instructor is explaining the material?		
Felt excited during your instructor's explanations?		
Felt anxious when your instructor is explaining the material?		
Felt frustrated during your instructor's explanations?		
Felt calm during your instructor's	I'm not sure about this one, because "calm"	

explanations?	is a neutral state; one could even say it is the baseline. (But then again, that could be a good thing, because it would provide needed variability.)	
Felt bored when your instructor is explaining the material?		
Behavioral Engagement in Whole-Class Interaction (WB)	Are there behaviors that would indicate lack of engagement?	Offering #'s for each subscale might make analysis challenging.
Asked questions to your instructor in front of the whole class?	This indicator could be confounded with shyness or other personality variables.	
Volunteered to answer your instructor's questions in front of the whole class?	Ditto.	
Listened to everything that is being said, including what other students say to the instructor?		
Taken notes on what is being said?	This is good! Taking notes on not only the lecture but also on students' comments/questions is a great upper-level indicator.	
Cognitive Engagement in Whole-Class Interaction (WC)		
Given your full attention to what is being said?		
Answered in your head or thought about questions your instructor asks the class?	I like this one!	
Tried to identify important information to write down or remember from what is being said?		

Tried to follow everything that is being said?	Hard to distinguish this from listening.	
Tried to put what is being said in your own words?	Point of clarification: put into words by speaking to the whole class, or just in your own mind?	
Tried to connect what is being said with what you know?		Again don't [unclear] "what you know"
Critically thought about what is being said?		Same with critical
Emotional Engagement in Whole-Class Interaction (WE)	Again, consider apathy.	
Enjoyed the time when your instructor interacts with the class?		
Felt excited during the interaction between your instructor and the class?		
Felt anxious during your instructor's interactions with the class?		
Felt frustrated during the time your instructor interacts with the class?		
Felt calm while your instructor interacts with the class?	Same concern about "Calm" as a baseline.	
Felt bored when your instructor interacts with the class?		
Behavioral Engagement in Individual Work (IB)		
Written down in detail your task solution or thinking about the task?		

Looked at your notes or other resources (e.g., Internet)?		
Re-read the task before trying to solve or answer it?		
Tried different ways of solving or thinking about the task even if you already have an answer?	I’m not sure if this is reflective of engagement per se, but of a different construct, such as curiosity. (Of course, it could be partially reflective of engagement.)	Seems more cognitive
Cognitive Engagement in Individual Work (IC)		
Tried to recall from memory the content needed to solve/answer the task?	I’m not sure that recalling from memory is necessarily reflective of engagement. One could be just as cognitively engaged while looking up the content. Maybe ask if some of the content was recalled?	
Critically thought about your task solution, answer, or solution attempts?		Critically
Checked that your work or answer on the task fits with the task instructions/question?		
Tried to keep the task instructions/question in mind while solving or answering the task?	I’m not sure how you could rephrase this, but as is, it could be interpreted as needing to literally have it repeating in your mind.	
Tried to identify the most important information from the task?		
Tried to put the task instructions/question in your own words?		

Tried to make sure you know why you use particular strategies or reasoning to solve or answer the task?	
Emotional Engagement in Individual Work (IE)	Good job on this scale!
Enjoyed working on the task in class on your own?	
Felt excited while working on the task in class by yourself?	
Felt anxious doing the task in class on your own?	I like this as a reverse-scored item!
Felt frustrated working on your own in class?	This, too!
Felt calm doing the task in class by yourself?	
Felt bored when you work on the task in class by yourself?	
Behavioral Engagement in Group Work (GB)	
Shared your thinking about the task with other students?	
Taken notes on other students' thinking about the task or on their solution/answer?	
Asked other students about their solutions, answers, or thinking about the task?	Perhaps add "process" to the list.

Looked at what other students wrote about the task?	
Checked with other students to see if your answers, solutions, or approaches match theirs?	
Cognitive Engagement in Group Work (GC)	
Given your full attention to what other students are saying about the task?	
Tried to select relevant information from what other students are saying about the task?	
Tried to connect other students' thinking about the task, their solutions, or answers to your own?	Consider adding "process."
Tried to use other students' ideas, solutions, or answers in your thinking about the task?	
Tried to follow everything that other students are saying about the task?	
Critically thought about other students' thinking about the task, their solutions, or answers?	
Compared your and other students' ways of thinking about the task?	
Considered what other students are saying about the task?	
Emotional Engagement in Group Work (GE)	

Enjoyed interacting with other students about the task?		
Felt excited during the time you interact with other students about the task?		
Felt anxious when talking to other students about the task?		
Felt frustrated when you interact with other students about the task?		
Felt calm while talking to other students about the task?		
Felt bored when you interact with other students about the task?	I also like this as a reverse-scored item.	
Response options to engagement items	This is a conventional scale, and should work well. However, if you find you are getting skewed response distributions, consider rebalancing the scale with different response options.	I might drop the or almost never
The overall scale	Very thorough, great work!	
Student version of the instrument	It looks good! I like the alternating bands of color, and the use of the response option labels rather than numbers.	

Appendix J

Measure of Student Engagement

Item (abbreviated)	Item
BEHAVIORAL engagement in LECTURE (LB): In class, when your instructor explains the material without interacting with students, how often have you…	
LB7_read	Read what your instructor is writing or showing in class (e.g., instructor's notes, PowerPoint slides, etc.)?
LB10_listen	*Listened to your instructor's explanations?*
LB2_notes	Taken notes on what your instructor is explaining?
LB5_pictures	Drawn your own pictures of your instructor's explanations?
LB13_remarks	Written your own remarks or comments on your instructor's explanations?
COGNITIVE engagement in LECTURE (LC): In class, when your instructor explains the material without interacting with students, how often have you…	
LC3_attention	Given your full attention to what your instructor is explaining?
LC6_identify	Tried to identify important information from what your instructor is explaining?
LC15_connect	Tried to connect what your instructor is explaining with what you know?
LC12_critical	Critically thought about your instructor's explanations?
LC9_ownwords	Tried to put what your instructor is explaining in your own words?
EMOTIONAL engagement in LECTURE (LE): In class, when your instructor explains the material without interacting with students, how often have you…	
LE1_enjoyed	*Enjoyed the time when your instructor is explaining the material?*
LE11_excited	Felt excited during your instructor's explanations?
LE14_calm	Felt calm during your instructor's explanations?
LE4_frustrated_rec	Felt frustrated during your instructor's explanations? (recoded)
LE16_anxious_rec	*Felt anxious when your instructor is explaining the material? (recoded)*
LE8_bored_rec	Felt bored when your instructor is explaining the material? (recoded)
BEHAVIORAL engagement in WHOLE-CLASS INTERACTION (WB): In class, when your instructor interacts with students addressing the class as a whole, how often have you…	
WB6_volunteer	Volunteered to answer your instructor's questions in front of the whole class?
WB14_shared	Shared your ideas or thoughts with the whole class?
WB18_asked	Asked questions to your instructor in front of the whole class?
WB9_listen	*Listened to what is being said between your instructor and other students?*
WB3_notes	Taken notes on what is being said between your instructor and other

Item (abbreviated)	Item
	students?
WB11_pictures	Drawn your own pictures of what is being said between your instructor and other students?
WB16_remarks	Written your own remarks or comments on what is being said between your instructor and other students?
COGNITIVE engagement in WHOLE-CLASS INTERACTION (WC): In class, when your instructor interacts with students addressing the class as a whole, how often have you…	
WC13_answeredhead	*Answered in your head or thought about questions your instructor asks the class?*
WC19_attention	Given your full attention to what is being said between your instructor and other students?
WC7_identify	Tried to identify important information from what is being said between your instructor and other students?
WC10_connect	Tried to connect what is being said between your instructor and other students with what you know?
WC4_critical	Critically thought about what is being said between your instructor and other students?
WC2_ownwords	Tried to put what is being said between your instructor and other students in your own words?
EMOTIONAL engagement in WHOLE-CLASS INTERACTION (WE): In class, when your instructor interacts with students addressing the class as a whole, how often have you…	
WE17_enjoyed	*Enjoyed the time when your instructor interacts with the class?*
WE5_excited	Felt excited during the interaction between your instructor and the class?
WE8_calm	Felt calm while your instructor interacts with the class?
WE15_frustrated_rec	Felt frustrated during the time your instructor interacts with the class? (recoded)
WE12_anxious_rec	*Felt anxious during your instructor's interactions with the class? (recoded)*
WE1_bored_rec	Felt bored when your instructor interacts with the class? (recoded)
BEHAVIORAL engagement in INDIVIDUAL WORK (IB): In class, when you work on a task without interacting with other students (excluding exams and formal quizzes), how often have you...	
IB15_reread	Re-read the task before trying to solve or answer it?
IB11_looked	*Looked at your notes or other resources (e.g., Internet)?*
IB7_checked	Checked that your work or answer on the task fits with the task instructions/question?
IB2_write	Written down in detail your task solution or thinking about the task?
IB17_wrotedifways	Written down more than one way of solving or of thinking about the task even if you already have an answer?
COGNITIVE engagement in INDIVIDUAL WORK (IC): In class, when you work on a task without interacting with other students (excluding exams and formal quizzes), how often have you...	
IC3_recall	Tried to recall from memory the content needed to solve/answer the task?
IC13_keepinmind	Tried to keep the task instructions/question in mind while solving or answering the task?
IC9_why	Tried to make sure you know why you use particular strategies or

Item (abbreviated)	Item
	reasoning to solve or answer the task?
IC6_thoughtdifways	Thought about different ways of solving or answering the task even if you already have an answer?
IC5_identify	Tried to identify the most important information from the task?
IC12_critical	Critically thought about your task solution, answer, or solution attempts?
IC16_ownwords	Tried to put the task instructions/question in your own words?
EMOTIONAL engagement in INDIVIDUAL WORK (IE): In class, when you work on a task without interacting with other students (excluding exams and formal quizzes), how often have you...	
IE8_enjoyed	*Enjoyed working on the task in class on your own?*
IE14_excited	Felt excited while working on the task in class by yourself?
IE10_calm	Felt calm doing the task in class by yourself?
IE18_frustrated_rec	Felt frustrated working on your own in class? (recoded)
IE1_anxious_rec	*Felt anxious doing the task in class on your own? (recoded)*
IE4_bored_rec	Felt bored when you work on the task in class by yourself? (recoded)
BEHAVIORAL engagement in GROUP WORK (GB): In class, when you interact with other students about a task, how often have you...	
GB3_asked	Asked other students about their solutions, answers, or thinking about the task?
GB7_justified	Justified your thinking about the task when speaking with other students?
GB10_checked	Checked with other students to see if your answers, solutions, or approaches match theirs?
GB16_shared	*Shared your thinking about the task with other students?*
GB13_notes	Taken notes on other students' thinking about the task or on their solution/answer?
COGNITIVE engagement in GROUP WORK (GC): In class, when you interact with other students about a task, how often have you...	
GC9_attention	Given your full attention to what other students are saying about the task?
GC17_compared	Compared your and other students' ways of thinking about the task?
GC12_use	Tried to use other students' ideas, solutions, or answers in your thinking about the task?
GC2_identify	Tried to identify relevant information from what other students are saying about the task?
GC15_connect	Tried to connect other students' thinking, solutions, or answers on the task to your own?
GC4_critical	Critically thought about other students' thinking about the task, their solutions, or answers?
GC6_ownwords	Tried to put what other students are saying about the task in your own words?
EMOTIONAL engagement in GROUP WORK (GE): In class, when you interact with other students about a task, how often have you...	
GE14_enjoyed	*Enjoyed interacting with other students about the task?*
GE1_excited	Felt excited during the time you interact with other students about the task?
GE11_calm	Felt calm while talking to other students about the task?

Item (abbreviated)	Item
GE18_frustrated_rec	Felt frustrated when you interact with other students about the task? (recoded)
GE5_anxious_rec	*Felt anxious when talking to other students about the task? (recoded)*
GE8_bored_rec	Felt bored when you interact with other students about the task? (recoded)

Note. Numbers in the abbreviated items are the order, in which the items were administered in an item block. The item block includes all items for a particular instruction type (4 item blocks in total). In cursive are items that were administered but were not used in the computation of composite scores.

Appendix K

Measures of Multi-Item Constructs Needed for Validation

[In cursive are items that were adimistered but were not included in the computation of composite scores; (r) = item was reverse-scored]

Effort:

1. I put a lot of effort into this class.
2. I work very hard in this class.

Persistence:

1. When I become confused about something I'm studying for this class, I go back and try to figure it out.
2. When something that I am studying for this class gets difficult, I spend extra time and effort trying to understand it.
3. In this class, I try to learn all of the testable material "inside and out," even if it is boring.
4. *In this class, regardless of whether or not I like the material, I work my hardest to learn it.*

Interest - Feeling:

1. What we are learning in this class is fascinating to me.
2. I am excited about what we are learning in this class.
3. I like what we are learning in this class.
4. I find the things we study in this class interesting.

Interest - Value:

1. What we are studying in this class is useful for me to know.
2. The things we are studying in this class are important to me.
3. I see how I can apply what we are learning in this class to real life.
4. We are learning valuable things in this class.
5. *What we are learning in this class is important for my future goals.*
6. *I find the content of this class personally meaningful.*

Metacognitive strategies:

1. Before starting an assignment for this class, I try to figure out the best way to do it.
2. Before I begin to study for this class, I think about what I want to get done.
3. For assignments in this class, I double check my work to make sure I am doing it right.
4. When I'm working on assignments for this class, I stop once in a while and go over what I have been doing.
5. In this class, I keep track of how much I understand the work, not just if I am getting the right answers.
6. I try to adapt how I do assignments for this class to fit with what the teacher wants or expects.
7. If what I am working on for this class is difficult to understand, I change the way I learn the material.
8. *In this class, I start my assignments without really planning out what I want to get done. (r)*
9. *I try to change the way I study for this class to fit the type of material I am trying to learn.*

Social Efficacy with Peers:

1. I find it easy to start a conversation with other students in this class.
2. I can explain my point of view to other students in this class.
3. I can get along with most students in this class.
4. I can work well with other students in this class.

Preference for group work:

1. When I have a choice, I try to work in a group instead of by myself.
2. I prefer to work on a team rather than individual tasks.
3. Working in a group is better than working alone.
4. Given the choice, I would rather do a job where I can work alone rather than do a job where I have to work with others in a group. (r)
5. I prefer to do my own work and let others do theirs. (r)
6. I personally enjoy working with others.
7. *I like to interact with others when working on projects.*

Intellect:

1. I am quick to understand things.
2. I have difficulty understanding abstract ideas. (r)
3. I can handle a lot of information.
4. I like to solve complex problems.
5. I avoid difficult reading material. (r)
6. I have a rich vocabulary.
7. I think quickly.
8. I learn things slowly. (r)

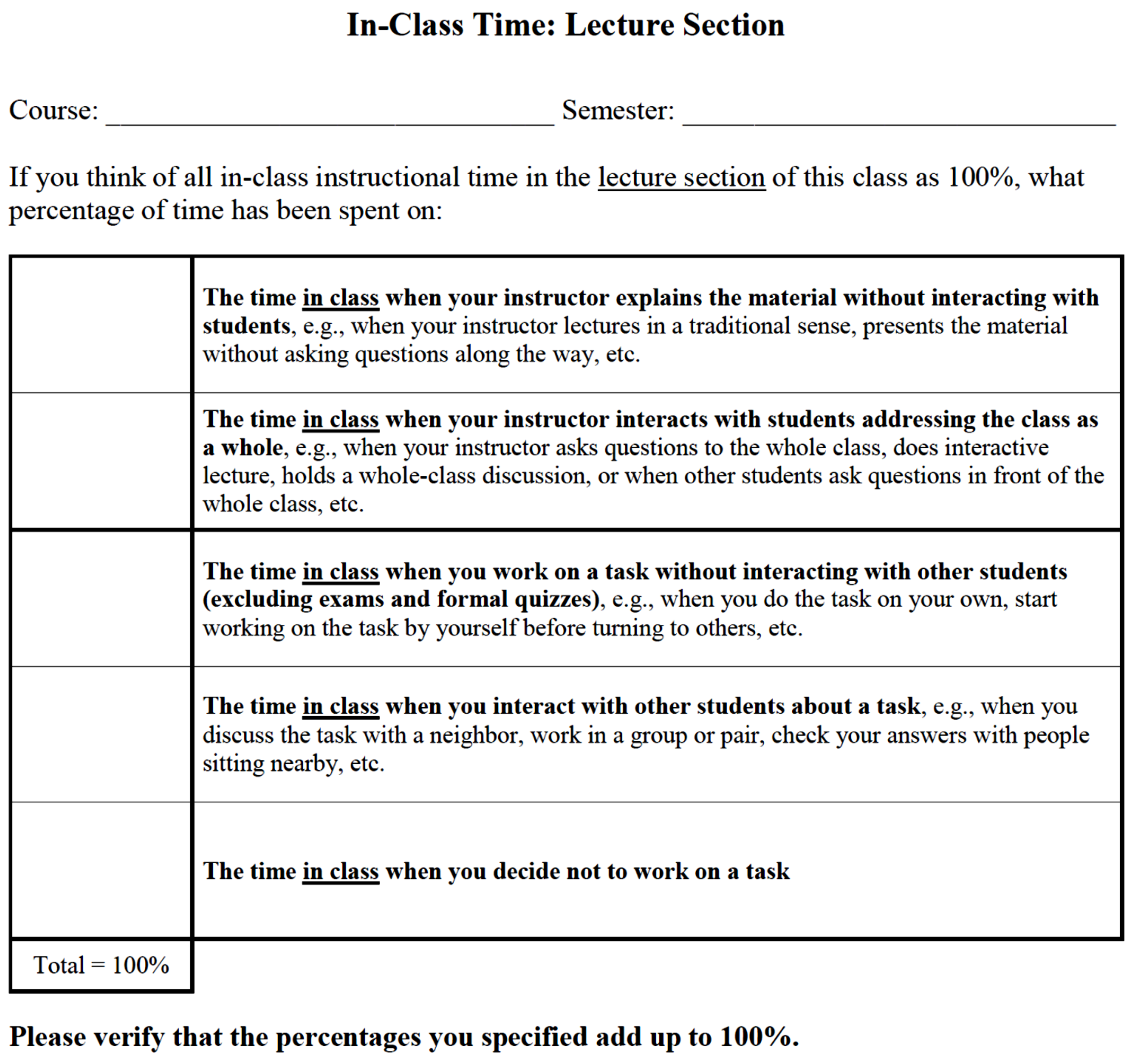
9. I formulate ideas clearly.
10. *I avoid philosophical discussions. (r)*

Public Speaking Anxiety:

1. My thoughts become confused and jumbled when I am speaking in front of the whole class.
2. Certain parts of my body feel very tense and rigid while I am speaking in front of the whole class.
3. My heart beats very fast while I am speaking in front of the whole class.
4. While speaking in front of the whole class, I get so nervous I forget facts I really know.
5. *I breathe faster just before I need to speak in front of the whole class.*

Appendix L

Student Survey



In-Class Time: Lecture Section

Course: ______________________________ Semester: ______________________________

If you think of all in-class instructional time in the lecture section of this class as 100%, what percentage of time has been spent on:

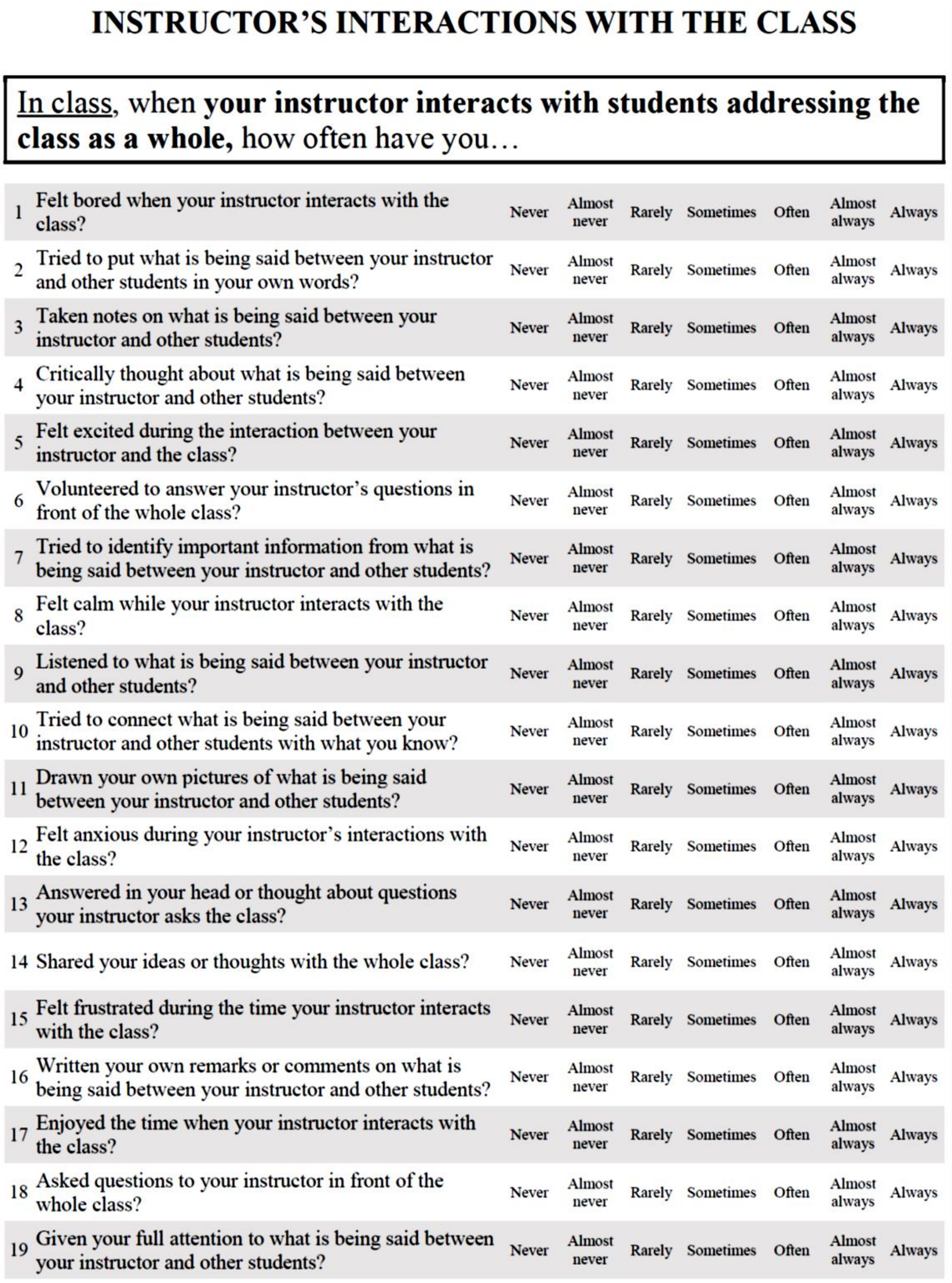
	The time in class when your instructor explains the material without interacting with students, e.g., when your instructor lectures in a traditional sense, presents the material without asking questions along the way, etc.
	The time in class when your instructor interacts with students addressing the class as a whole, e.g., when your instructor asks questions to the whole class, does interactive lecture, holds a whole-class discussion, or when other students ask questions in front of the whole class, etc.
	The time in class when you work on a task without interacting with other students (excluding exams and formal quizzes), e.g., when you do the task on your own, start working on the task by yourself before turning to others, etc.
	The time in class when you interact with other students about a task, e.g., when you discuss the task with a neighbor, work in a group or pair, check your answers with people sitting nearby, etc.
	The time in class when you decide not to work on a task
Total = 100%	

Please verify that the percentages you specified add up to 100%.

INSTRUCTOR’S PRESENTATIONS

In class, when **your instructor explains the material without interacting with students,** how often have you…

1	Enjoyed the time when your instructor is explaining the material?	Never	Almost never	Rarely	Sometimes	Often	Almost always	Always
2	Taken notes on what your instructor is explaining?	Never	Almost never	Rarely	Sometimes	Often	Almost always	Always
3	Given your full attention to what your instructor is explaining?	Never	Almost never	Rarely	Sometimes	Often	Almost always	Always
4	Felt frustrated during your instructor’s explanations?	Never	Almost never	Rarely	Sometimes	Often	Almost always	Always
5	Drawn your own pictures of your instructor’s explanations?	Never	Almost never	Rarely	Sometimes	Often	Almost always	Always
6	Tried to identify important information from what your instructor is explaining?	Never	Almost never	Rarely	Sometimes	Often	Almost always	Always
7	Read what your instructor is writing or showing in class (e.g., instructor’s notes, PowerPoint slides, etc.)?	Never	Almost never	Rarely	Sometimes	Often	Almost always	Always
8	Felt bored when your instructor is explaining the material?	Never	Almost never	Rarely	Sometimes	Often	Almost always	Always
9	Tried to put what your instructor is explaining in your own words?	Never	Almost never	Rarely	Sometimes	Often	Almost always	Always
10	Listened to your instructor’s explanations?	Never	Almost never	Rarely	Sometimes	Often	Almost always	Always
11	Felt excited during your instructor’s explanations?	Never	Almost never	Rarely	Sometimes	Often	Almost always	Always
12	Critically thought about your instructor’s explanations?	Never	Almost never	Rarely	Sometimes	Often	Almost always	Always
13	Written your own remarks or comments on your instructor’s explanations?	Never	Almost never	Rarely	Sometimes	Often	Almost always	Always
14	Felt calm during your instructor’s explanations?	Never	Almost never	Rarely	Sometimes	Often	Almost always	Always
15	Tried to connect what your instructor is explaining with what you know?	Never	Almost never	Rarely	Sometimes	Often	Almost always	Always
16	Felt anxious when your instructor is explaining the material?	Never	Almost never	Rarely	Sometimes	Often	Almost always	Always



INSTRUCTOR'S INTERACTIONS WITH THE CLASS

In class, when **your instructor interacts with students addressing the class as a whole,** how often have you…

1	Felt bored when your instructor interacts with the class?	Never	Almost never	Rarely	Sometimes	Often	Almost always	Always
2	Tried to put what is being said between your instructor and other students in your own words?	Never	Almost never	Rarely	Sometimes	Often	Almost always	Always
3	Taken notes on what is being said between your instructor and other students?	Never	Almost never	Rarely	Sometimes	Often	Almost always	Always
4	Critically thought about what is being said between your instructor and other students?	Never	Almost never	Rarely	Sometimes	Often	Almost always	Always
5	Felt excited during the interaction between your instructor and the class?	Never	Almost never	Rarely	Sometimes	Often	Almost always	Always
6	Volunteered to answer your instructor's questions in front of the whole class?	Never	Almost never	Rarely	Sometimes	Often	Almost always	Always
7	Tried to identify important information from what is being said between your instructor and other students?	Never	Almost never	Rarely	Sometimes	Often	Almost always	Always
8	Felt calm while your instructor interacts with the class?	Never	Almost never	Rarely	Sometimes	Often	Almost always	Always
9	Listened to what is being said between your instructor and other students?	Never	Almost never	Rarely	Sometimes	Often	Almost always	Always
10	Tried to connect what is being said between your instructor and other students with what you know?	Never	Almost never	Rarely	Sometimes	Often	Almost always	Always
11	Drawn your own pictures of what is being said between your instructor and other students?	Never	Almost never	Rarely	Sometimes	Often	Almost always	Always
12	Felt anxious during your instructor's interactions with the class?	Never	Almost never	Rarely	Sometimes	Often	Almost always	Always
13	Answered in your head or thought about questions your instructor asks the class?	Never	Almost never	Rarely	Sometimes	Often	Almost always	Always
14	Shared your ideas or thoughts with the whole class?	Never	Almost never	Rarely	Sometimes	Often	Almost always	Always
15	Felt frustrated during the time your instructor interacts with the class?	Never	Almost never	Rarely	Sometimes	Often	Almost always	Always
16	Written your own remarks or comments on what is being said between your instructor and other students?	Never	Almost never	Rarely	Sometimes	Often	Almost always	Always
17	Enjoyed the time when your instructor interacts with the class?	Never	Almost never	Rarely	Sometimes	Often	Almost always	Always
18	Asked questions to your instructor in front of the whole class?	Never	Almost never	Rarely	Sometimes	Often	Almost always	Always
19	Given your full attention to what is being said between your instructor and other students?	Never	Almost never	Rarely	Sometimes	Often	Almost always	Always

INDIVIDUAL IN-CLASS WORK

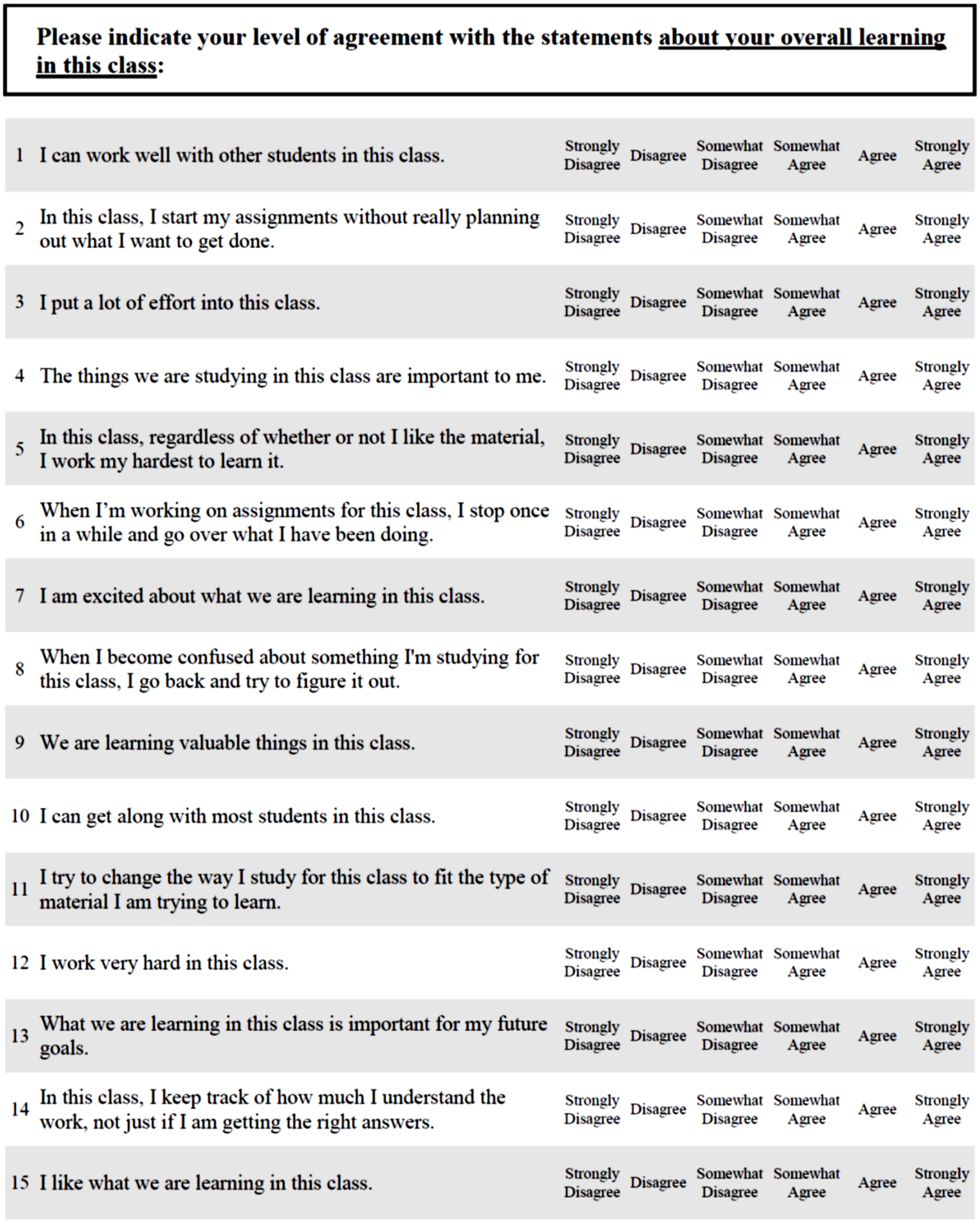
In class, when **you work on a task without interacting with other students (excluding exams and formal quizzes)**, how often have you…

1	Felt anxious doing the task in class on your own?	Never	Almost never	Rarely	Sometimes	Often	Almost always	Always
2	Written down in detail your task solution or thinking about the task?	Never	Almost never	Rarely	Sometimes	Often	Almost always	Always
3	Tried to recall from memory the content needed to solve/answer the task?	Never	Almost never	Rarely	Sometimes	Often	Almost always	Always
4	Felt bored when you work on the task in class by yourself?	Never	Almost never	Rarely	Sometimes	Often	Almost always	Always
5	Tried to identify the most important information from the task?	Never	Almost never	Rarely	Sometimes	Often	Almost always	Always
6	Thought about different ways of solving or answering the task even if you already have an answer?	Never	Almost never	Rarely	Sometimes	Often	Almost always	Always
7	Checked that your work or answer on the task fits with the task instructions/question?	Never	Almost never	Rarely	Sometimes	Often	Almost always	Always
8	Enjoyed working on the task in class on your own?	Never	Almost never	Rarely	Sometimes	Often	Almost always	Always
9	Tried to make sure you know why you use particular strategies or reasoning to solve or answer the task?	Never	Almost never	Rarely	Sometimes	Often	Almost always	Always
10	Felt calm doing the task in class by yourself?	Never	Almost never	Rarely	Sometimes	Often	Almost always	Always
11	Looked at your notes or other resources (e.g., Internet)?	Never	Almost never	Rarely	Sometimes	Often	Almost always	Always
12	Critically thought about your task solution, answer, or solution attempts?	Never	Almost never	Rarely	Sometimes	Often	Almost always	Always
13	Tried to keep the task instructions/question in mind while solving or answering the task?	Never	Almost never	Rarely	Sometimes	Often	Almost always	Always
14	Felt excited while working on the task in class by yourself?	Never	Almost never	Rarely	Sometimes	Often	Almost always	Always
15	Re-read the task before trying to solve or answer it?	Never	Almost never	Rarely	Sometimes	Often	Almost always	Always
16	Tried to put the task instructions/question in your own words?	Never	Almost never	Rarely	Sometimes	Often	Almost always	Always
17	Written down more than one way of solving or of thinking about the task even if you already have an answer?	Never	Almost never	Rarely	Sometimes	Often	Almost always	Always
18	Felt frustrated working on your own in class?	Never	Almost never	Rarely	Sometimes	Often	Almost always	Always

GROUP IN-CLASS WORK

In class, when **you interact with other students about a task,** how often have you…

1	Felt excited during the time you interact with other students about the task?	Never	Almost never	Rarely	Sometimes	Often	Almost always	Always
2	Tried to identify relevant information from what other students are saying about the task?	Never	Almost never	Rarely	Sometimes	Often	Almost always	Always
3	Asked other students about their solutions, answers, or thinking about the task?	Never	Almost never	Rarely	Sometimes	Often	Almost always	Always
4	Critically thought about other students' thinking about the task, their solutions, or answers?	Never	Almost never	Rarely	Sometimes	Often	Almost always	Always
5	Felt anxious when talking to other students about the task?	Never	Almost never	Rarely	Sometimes	Often	Almost always	Always
6	Tried to put what other students are saying about the task in your own words?	Never	Almost never	Rarely	Sometimes	Often	Almost always	Always
7	Justified your thinking about the task when speaking with other students?	Never	Almost never	Rarely	Sometimes	Often	Almost always	Always
8	Felt bored when you interact with other students about the task?	Never	Almost never	Rarely	Sometimes	Often	Almost always	Always
9	Given your full attention to what other students are saying about the task?	Never	Almost never	Rarely	Sometimes	Often	Almost always	Always
10	Checked with other students to see if your answers, solutions, or approaches match theirs?	Never	Almost never	Rarely	Sometimes	Often	Almost always	Always
11	Felt calm while talking to other students about the task?	Never	Almost never	Rarely	Sometimes	Often	Almost always	Always
12	Tried to use other students' ideas, solutions, or answers in your thinking about the task?	Never	Almost never	Rarely	Sometimes	Often	Almost always	Always
13	Taken notes on other students' thinking about the task or on their solution/answer?	Never	Almost never	Rarely	Sometimes	Often	Almost always	Always
14	Enjoyed interacting with other students about the task?	Never	Almost never	Rarely	Sometimes	Often	Almost always	Always
15	Tried to connect other students' thinking, solutions, or answers on the task to your own?	Never	Almost never	Rarely	Sometimes	Often	Almost always	Always
16	Shared your thinking about the task with other students?	Never	Almost never	Rarely	Sometimes	Often	Almost always	Always
17	Compared your and other students' ways of thinking about the task?	Never	Almost never	Rarely	Sometimes	Often	Almost always	Always
18	Felt frustrated when you interact with other students about the task?	Never	Almost never	Rarely	Sometimes	Often	Almost always	Always



Please indicate your level of agreement with the statements about your overall learning in this class:

1	I can work well with other students in this class.	Strongly Disagree	Disagree	Somewhat Disagree	Somewhat Agree	Agree	Strongly Agree
2	In this class, I start my assignments without really planning out what I want to get done.	Strongly Disagree	Disagree	Somewhat Disagree	Somewhat Agree	Agree	Strongly Agree
3	I put a lot of effort into this class.	Strongly Disagree	Disagree	Somewhat Disagree	Somewhat Agree	Agree	Strongly Agree
4	The things we are studying in this class are important to me.	Strongly Disagree	Disagree	Somewhat Disagree	Somewhat Agree	Agree	Strongly Agree
5	In this class, regardless of whether or not I like the material, I work my hardest to learn it.	Strongly Disagree	Disagree	Somewhat Disagree	Somewhat Agree	Agree	Strongly Agree
6	When I'm working on assignments for this class, I stop once in a while and go over what I have been doing.	Strongly Disagree	Disagree	Somewhat Disagree	Somewhat Agree	Agree	Strongly Agree
7	I am excited about what we are learning in this class.	Strongly Disagree	Disagree	Somewhat Disagree	Somewhat Agree	Agree	Strongly Agree
8	When I become confused about something I'm studying for this class, I go back and try to figure it out.	Strongly Disagree	Disagree	Somewhat Disagree	Somewhat Agree	Agree	Strongly Agree
9	We are learning valuable things in this class.	Strongly Disagree	Disagree	Somewhat Disagree	Somewhat Agree	Agree	Strongly Agree
10	I can get along with most students in this class.	Strongly Disagree	Disagree	Somewhat Disagree	Somewhat Agree	Agree	Strongly Agree
11	I try to change the way I study for this class to fit the type of material I am trying to learn.	Strongly Disagree	Disagree	Somewhat Disagree	Somewhat Agree	Agree	Strongly Agree
12	I work very hard in this class.	Strongly Disagree	Disagree	Somewhat Disagree	Somewhat Agree	Agree	Strongly Agree
13	What we are learning in this class is important for my future goals.	Strongly Disagree	Disagree	Somewhat Disagree	Somewhat Agree	Agree	Strongly Agree
14	In this class, I keep track of how much I understand the work, not just if I am getting the right answers.	Strongly Disagree	Disagree	Somewhat Disagree	Somewhat Agree	Agree	Strongly Agree
15	I like what we are learning in this class.	Strongly Disagree	Disagree	Somewhat Disagree	Somewhat Agree	Agree	Strongly Agree

16	When something that I am studying for this class gets difficult, I spend extra time and effort trying to understand it.	Strongly Disagree	Disagree	Somewhat Disagree	Somewhat Agree	Agree	Strongly Agree
17	I find it easy to start a conversation with other students in this class.	Strongly Disagree	Disagree	Somewhat Disagree	Somewhat Agree	Agree	Strongly Agree
18	Before starting an assignment for this class, I try to figure out the best way to do it.	Strongly Disagree	Disagree	Somewhat Disagree	Somewhat Agree	Agree	Strongly Agree
19	I see how I can apply what we are learning in this class to real life.	Strongly Disagree	Disagree	Somewhat Disagree	Somewhat Agree	Agree	Strongly Agree
20	What we are learning in this class is fascinating to me.	Strongly Disagree	Disagree	Somewhat Disagree	Somewhat Agree	Agree	Strongly Agree
21	For assignments in this class, I double check my work to make sure I am doing it right.	Strongly Disagree	Disagree	Somewhat Disagree	Somewhat Agree	Agree	Strongly Agree
22	What we are studying in this class is useful for me to know.	Strongly Disagree	Disagree	Somewhat Disagree	Somewhat Agree	Agree	Strongly Agree
23	If what I am working on for this class is difficult to understand, I change the way I learn the material.	Strongly Disagree	Disagree	Somewhat Disagree	Somewhat Agree	Agree	Strongly Agree
24	I can explain my point of view to other students in this class.	Strongly Disagree	Disagree	Somewhat Disagree	Somewhat Agree	Agree	Strongly Agree
25	I find the things we study in this class interesting.	Strongly Disagree	Disagree	Somewhat Disagree	Somewhat Agree	Agree	Strongly Agree
26	I try to adapt how I do assignments for this class to fit with what the teacher wants or expects.	Strongly Disagree	Disagree	Somewhat Disagree	Somewhat Agree	Agree	Strongly Agree
27	I find the content of this class personally meaningful.	Strongly Disagree	Disagree	Somewhat Disagree	Somewhat Agree	Agree	Strongly Agree
28	In this class, I try to learn all of the testable material "inside and out," even if it is boring.	Strongly Disagree	Disagree	Somewhat Disagree	Somewhat Agree	Agree	Strongly Agree
29	Before I begin to study for this class, I think about what I want to get done.	Strongly Disagree	Disagree	Somewhat Disagree	Somewhat Agree	Agree	Strongly Agree

Please indicate your level of agreement with the statements about yourself in general:

1	When I have a choice, I try to work in a group instead of by myself.	Strongly Disagree	Disagree	Somewhat Disagree	Somewhat Agree	Agree	Strongly Agree
2	I think quickly.	Strongly Disagree	Disagree	Somewhat Disagree	Somewhat Agree	Agree	Strongly Agree
3	I breathe faster just before I need to speak in front of the whole class.	Strongly Disagree	Disagree	Somewhat Disagree	Somewhat Agree	Agree	Strongly Agree
4	I have difficulty understanding abstract ideas.	Strongly Disagree	Disagree	Somewhat Disagree	Somewhat Agree	Agree	Strongly Agree
5	I prefer to work on a team rather than individual tasks.	Strongly Disagree	Disagree	Somewhat Disagree	Somewhat Agree	Agree	Strongly Agree
6	I can handle a lot of information.	Strongly Disagree	Disagree	Somewhat Disagree	Somewhat Agree	Agree	Strongly Agree
7	My heart beats very fast while I am speaking in front of the whole class.	Strongly Disagree	Disagree	Somewhat Disagree	Somewhat Agree	Agree	Strongly Agree
8	I formulate ideas clearly.	Strongly Disagree	Disagree	Somewhat Disagree	Somewhat Agree	Agree	Strongly Agree
9	I personally enjoy working with others.	Strongly Disagree	Disagree	Somewhat Disagree	Somewhat Agree	Agree	Strongly Agree
10	I learn things slowly.	Strongly Disagree	Disagree	Somewhat Disagree	Somewhat Agree	Agree	Strongly Agree
11	My thoughts become confused and jumbled when I am speaking in front of the whole class.	Strongly Disagree	Disagree	Somewhat Disagree	Somewhat Agree	Agree	Strongly Agree
12	I like to solve complex problems.	Strongly Disagree	Disagree	Somewhat Disagree	Somewhat Agree	Agree	Strongly Agree
13	Working in a group is better than working alone.	Strongly Disagree	Disagree	Somewhat Disagree	Somewhat Agree	Agree	Strongly Agree
14	I avoid difficult reading material.	Strongly Disagree	Disagree	Somewhat Disagree	Somewhat Agree	Agree	Strongly Agree
15	Certain parts of my body feel very tense and rigid while I am speaking in front of the whole class.	Strongly Disagree	Disagree	Somewhat Disagree	Somewhat Agree	Agree	Strongly Agree
16	Given the choice, I would rather do a job where I can work alone rather than do a job where I have to work with others in a group.	Strongly Disagree	Disagree	Somewhat Disagree	Somewhat Agree	Agree	Strongly Agree
17	I am quick to understand things.	Strongly Disagree	Disagree	Somewhat Disagree	Somewhat Agree	Agree	Strongly Agree
18	I have a rich vocabulary.	Strongly Disagree	Disagree	Somewhat Disagree	Somewhat Agree	Agree	Strongly Agree
19	I prefer to do my own work and let others do theirs.	Strongly Disagree	Disagree	Somewhat Disagree	Somewhat Agree	Agree	Strongly Agree
20	While speaking in front of the whole class, I get so nervous I forget facts I really know.	Strongly Disagree	Disagree	Somewhat Disagree	Somewhat Agree	Agree	Strongly Agree
21	I avoid philosophical discussions.	Strongly Disagree	Disagree	Somewhat Disagree	Somewhat Agree	Agree	Strongly Agree
22	I like to interact with others when working on projects.	Strongly Disagree	Disagree	Somewhat Disagree	Somewhat Agree	Agree	Strongly Agree

Demographic Questionnaire

1	What grade do you expect to get in this class?	A	B	C	D	F		
2	How much have you learned in this class?	Nothing	Very little	Some	Quite a bit	A lot		
3	How much **could** you have learned in this class in the ideal circumstances?	As much as I learned	Somewhat more than I learned	Substantially more than I learned				
4	How much did you know about the course content before taking this class?	Nothing	Very little	Some	Quite a bit	A lot		
5	How many times were you absent from the lecture section of this class?	0	1-4	5-10	More than 10			
6	What type of course is it for you?	Required for major / minor	Elective	General Education (Mason Core)	Other: ____			
7	What is your current/intended major?							
8	What is your current GPA?							
9	What is your current classification?	Freshman (0-29 credits completed)	Sophomore (30-59 credits completed)	Junior (60-89 credits completed)	Senior (90 or more credits completed)	Other: ____		
10	Are you in the Honors Program?	Yes	No					
11	What is your status this semester?	Full-time (12 credits or more)	Part-time (less than 12 credits)	Other: ____				
12	Are you a domestic or international student?	Domestic, In-state	Domestic, Out-of-state	International				
13	Is English your native language?	Yes	No	Other: ____				
14	What is your age?							
15	What is your gender?	Male	Female	Other	Prefer not to answer			
16	What is your race / ethnicity? (Please check all that apply).	White	Black or African-American	Hispanic / Latinx	Asian	American Indian or Pacific Islander	Other	Prefer not to answer

Thank you!

Appendix M

Item Descriptive Statistics and Frequencies

Item (abbreviated)	N	Mean	STD	Skewness	Kurtosis	1	2	3	4	5	6	7
BEHAVIORAL engagement in LECTURE (LB): In class, when your instructor explains the material without interacting with students, how often have you…												
LB7_read	1143	5.85	1.23	-1.22	1.80	13	7	27	96	249	307	444
LB10_listen	1138	5.88	1.08	-0.76	0.14	1	3	22	97	267	343	405
LB2_notes	1143	5.15	1.69	-0.74	-0.23	51	47	89	180	224	231	321
LB5_pictures	1141	3.77	1.58	-0.13	-0.58	140	107	191	332	223	103	45
LB13_remarks	1134	3.76	1.70	0.00	-0.74	158	112	202	285	199	105	73
COGNITIVE engagement in LECTURE (LC): In class, when your instructor explains the material without interacting with students, how often have you…												
LC3_attention	1139	5.47	1.20	-0.58	0.06	3	14	46	168	317	338	253
LC6_identify	1141	5.37	1.23	-0.64	0.63	11	14	37	185	365	294	235
LC15_connect	1137	5.28	1.24	-0.50	0.21	8	17	48	212	352	286	214
LC12_critical	1141	4.73	1.33	-0.28	0.03	20	36	123	294	365	183	120
LC9_ownwords	1137	4.47	1.37	-0.24	0.05	36	50	144	349	319	148	91
EMOTIONAL engagement in LECTURE (LE): In class, when your instructor explains the material without interacting with students, how often have you…												
LE1_enjoyed	1143	4.36	1.47	-0.14	-0.20	47	70	167	342	289	119	109
LE11_excited	1140	3.69	1.49	0.12	-0.23	107	114	277	346	170	75	51
LE14_calm	1139	5.49	1.37	-0.67	-0.06	10	15	67	172	281	239	355
LE4_frustrated_rec	1142	4.76	1.54	-0.31	-0.38	38	50	114	319	238	194	189
LE16_anxious_rec	1139	5.28	1.59	-0.66	-0.36	26	37	96	204	202	218	356
LE8_bored_rec	1139	4.20	1.48	0.00	-0.29	49	94	180	388	210	123	95
BEHAVIORAL engagement in WHOLE-CLASS INTERACTION (WB): In class, when your instructor interacts with students addressing the class as a whole, how often have you…												
WB6_volunteer	1233	3.04	1.65	0.32	-0.76	332	156	229	279	155	49	33
WB14_shared	1232	2.85	1.52	0.43	-0.52	324	208	270	261	112	38	19
WB18_asked	1235	2.90	1.64	0.46	-0.63	363	179	227	262	127	45	32
WB9_listen	1234	5.62	1.13	-0.62	0.35	5	5	27	156	353	360	328
WB3_notes	1232	4.29	1.73	-0.24	-0.74	107	98	179	260	267	178	143
WB11_pictures	1234	3.77	1.70	-0.02	-0.73	174	112	220	313	220	118	77
WB16_remarks	1230	3.21	1.62	0.27	-0.65	255	172	259	281	164	59	40
COGNITIVE engagement in WHOLE-CLASS INTERACTION (WC): In class, when your instructor interacts with students addressing the class as a whole, how often have you…												
WC13_answeredhead	1235	5.36	1.18	-0.51	0.39	7	14	39	206	407	327	235

Item (abbreviated)	N	Mean	STD	Skewness	Kurtosis	1	2	3	4	5	6	7
WC19_attention	1234	5.27	1.28	-0.47	-0.09	8	19	70	240	343	310	244
WC7_identify	1235	4.91	1.29	-0.47	0.47	22	33	80	286	443	223	148
WC10_connect	1234	5.20	1.27	-0.59	0.60	19	13	56	244	393	297	212
WC4_critical	1232	4.65	1.35	-0.37	0.28	34	46	113	340	408	172	119
WC2_ownwords	1238	4.31	1.31	-0.18	0.22	38	67	175	413	361	110	74
EMOTIONAL engagement in WHOLE-CLASS INTERACTION (WE): In class, when your instructor interacts with students addressing the class as a whole, how often have you…												
WE17_enjoyed	1229	4.62	1.46	-0.24	-0.16	37	60	122	372	312	172	154
WE5_excited	1233	3.83	1.47	-0.01	-0.14	105	96	265	408	215	85	59
WE8_calm	1235	5.62	1.28	-0.80	0.49	11	10	48	147	338	275	406
WE15_frustrated_rec	1235	5.50	1.48	-0.76	-0.10	16	29	74	201	238	240	437
WE12_anxious_rec	1230	5.63	1.44	-0.97	0.48	19	23	60	148	267	236	477
WE1_bored_rec	1238	4.67	1.46	-0.12	-0.39	29	46	163	361	287	172	180
BEHAVIORAL engagement in INDIVIDUAL WORK (IB): In class, when you work on a task without interacting with other students (excluding exams and formal quizzes), how often have you...												
IB15_reread	1049	5.39	1.22	-0.56	0.35	8	8	49	151	347	254	232
IB11_looked	1048	5.27	1.39	-0.78	0.58	23	21	50	173	296	255	230
IB7_checked	1047	5.47	1.18	-0.51	0.20	4	12	28	156	341	260	246
IB2_write	1046	4.51	1.46	-0.33	-0.19	38	63	124	274	292	159	96
IB17_wrotedifways	1049	3.63	1.56	0.24	-0.41	99	157	234	282	153	65	59
COGNITIVE engagement in INDIVIDUAL WORK (IC): In class, when you work on a task without interacting with other students (excluding exams and formal quizzes), how often have you...												
IC3_recall	1045	5.53	1.11	-0.60	0.83	6	6	16	131	357	294	235
IC13_keepinmind	1043	5.51	1.15	-0.46	0.05	2	11	23	146	346	262	253
IC9_why	1047	4.97	1.35	-0.40	0.10	19	19	88	238	329	193	161
IC6_thoughtdifways	1046	4.35	1.61	-0.11	-0.62	50	83	183	238	244	122	126
IC5_identify	1044	5.45	1.18	-0.61	0.49	4	18	26	142	350	274	230
IC12_critical	1046	5.22	1.21	-0.42	0.30	9	10	47	207	356	236	181
IC16_ownwords	1046	4.52	1.53	-0.23	-0.34	44	52	149	270	263	141	127
EMOTIONAL engagement in INDIVIDUAL WORK (IE): In class, when you work on a task without interacting with other students (excluding exams and formal quizzes), how often have you...												
IE8_enjoyed	1049	4.08	1.44	-0.03	-0.09	55	87	169	369	216	83	70
IE14_excited	1047	3.46	1.42	0.19	-0.07	113	133	273	332	114	51	31
IE10_calm	1048	4.85	1.42	-0.29	-0.16	21	35	94	271	302	154	171
IE18_frustrated_rec	1033	4.48	1.60	-0.10	-0.55	45	58	158	303	188	130	151
IE1_anxious_rec	1049	4.63	1.71	-0.17	-0.79	50	51	175	251	176	131	215
IE4_bored_rec	1045	4.29	1.50	-0.02	-0.28	46	62	178	330	219	104	106
BEHAVIORAL engagement in GROUP WORK (GB): In class, when you interact with other students about a task, how often have you...												
GB3_asked	1187	5.25	1.37	-0.78	0.59	22	30	62	172	369	290	242
GB7_justified	1189	5.01	1.25	-0.52	0.78	22	19	54	277	429	233	155

Item (abbreviated)	N	Mean	STD	Skewness	Kurtosis	1	2	3	4	5	6	7
GB10_checked	1189	5.56	1.29	-0.92	0.99	16	13	40	143	324	312	341
GB16_shared	1186	5.10	1.29	-0.50	0.38	16	26	64	237	410	239	194
GB13_notes	1186	3.73	1.67	0.02	-0.62	163	115	207	338	194	94	75
COGNITIVE engagement in GROUP WORK (GC): In class, when you interact with other students about a task, how often have you...												
GC9_attention	1185	5.40	1.11	-0.47	0.39	4	10	35	170	421	332	213
GC17_compared	1185	4.96	1.36	-0.57	0.48	30	29	75	253	404	224	170
GC12_use	1185	4.96	1.29	-0.59	0.71	25	28	60	274	412	247	139
GC2_identify	1192	5.15	1.20	-0.60	1.02	17	13	51	214	463	269	165
GC15_connect	1186	4.96	1.35	-0.50	0.48	28	28	64	284	402	203	177
GC4_critical	1189	5.00	1.31	-0.50	0.41	22	24	70	275	387	242	169
GC6_ownwords	1188	4.47	1.38	-0.35	0.23	48	44	140	357	359	149	91
EMOTIONAL engagement in GROUP WORK (GE): In class, when you interact with other students about a task, how often have you...												
GE14_enjoyed	1187	5.04	1.46	-0.55	0.09	34	25	86	259	324	223	236
GE1_excited	1194	4.29	1.50	-0.20	-0.15	69	66	168	374	287	126	104
GE11_calm	1187	5.54	1.31	-0.76	0.35	10	20	41	170	313	275	358
GE18_frustrated_rec	1189	5.37	1.41	-0.67	-0.02	14	28	64	216	263	283	321
GE5_anxious_rec	1191	5.20	1.58	-0.62	-0.25	31	41	94	205	282	195	343
GE8_bored_rec	1188	5.00	1.41	-0.37	-0.09	21	32	88	282	347	195	223

Note. Numbers in the abbreviated items are the order, in which the items were administered in an item block. The item block includes all items for a particular instruction type (4 item blocks in total).

Appendix N

Correlations for items within instruction type

Correlations for items within Lecture:

#	Item	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	LB7_read															
2	LB10_listen	**0.514**														
3	LB2_notes	**0.368**	**0.430**													
4	LB5_pictures	**0.172**	**0.234**	**0.287**												
5	LB13_remarks	**0.186**	**0.243**	**0.326**	**0.369**											
6	LC3_attention	**0.429**	**0.672**	**0.521**	**0.266**	**0.284**										
7	LC6_identify	**0.440**	**0.523**	**0.327**	**0.298**	**0.220**	**0.456**									
8	LC15_connect	**0.346**	**0.484**	**0.210**	**0.184**	**0.280**	**0.406**	**0.465**								
9	LC12_critical	**0.348**	**0.467**	**0.248**	**0.247**	**0.415**	**0.441**	**0.443**	**0.545**							
10	LC9_ownwords	**0.233**	**0.299**	**0.192**	**0.392**	**0.401**	**0.257**	**0.386**	**0.378**	**0.446**						
11	LE1_enjoyed	**0.268**	**0.332**	**0.176**	**0.175**	**0.195**	**0.400**	**0.229**	**0.289**	**0.350**	**0.159**					
12	LE11_excited	**0.180**	**0.276**	**0.178**	**0.199**	**0.283**	**0.326**	**0.159**	**0.271**	**0.371**	**0.218**	**0.679**				
13	LE14_calm	**0.199**	**0.245**	0.012	-0.032	-0.004	**0.208**	**0.158**	**0.289**	**0.188**	0.003	**0.349**	**0.267**			
14	LE4_frustrated_rec	**0.115**	**0.115**	-0.034	-0.041	0.008	**0.150**	0.076	**0.176**	**0.160**	0.004	**0.446**	**0.354**	**0.459**		
15	LE16_anxious_rec	0.055	**0.079**	**-0.081**	**-0.080**	-0.062	**0.088**	0.012	**0.100**	0.058	**-0.093**	**0.277**	**0.179**	**0.578**	**0.542**	
16	LE8_bored_rec	**0.254**	**0.378**	**0.224**	**0.168**	**0.201**	**0.456**	**0.270**	**0.285**	**0.327**	**0.152**	**0.635**	**0.550**	**0.303**	**0.514**	**0.324**

Note. Correlations that are statistically significant (p < 0.01) are in bold. Highlighted are correlations within hypothesized subscales.

Correlations for items within Whole-Class Interaction:

#	Item	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
1	WB6_volunteer																		
2	WB14_shared	**0.746**																	
3	WB18_asked	**0.656**	**0.655**																
4	WB9_listen	**0.106**	**0.143**	**0.161**															
5	WB3_notes	**0.106**	**0.156**	**0.156**	**0.348**														
6	WB11_pictures	**0.153**	**0.219**	**0.183**	**0.296**	**0.409**													
7	WB16_remarks	**0.170**	**0.234**	**0.220**	**0.217**	**0.477**	**0.470**												
8	WC13_answeredhead	0.069	**0.097**	**0.091**	**0.329**	**0.129**	**0.137**	**0.120**											
9	WC19_attention	**0.161**	**0.165**	**0.160**	**0.584**	**0.386**	**0.303**	**0.245**	**0.299**										
10	WC7_identify	**0.193**	**0.225**	**0.207**	**0.470**	**0.392**	**0.356**	**0.352**	**0.344**	**0.448**									
11	WC10_connect	**0.154**	**0.182**	**0.162**	**0.579**	**0.349**	**0.411**	**0.300**	**0.333**	**0.496**	**0.574**								
12	WC4_critical	**0.158**	**0.213**	**0.214**	**0.445**	**0.478**	**0.380**	**0.370**	**0.294**	**0.426**	**0.568**	**0.570**							
13	WC2_ownwords	**0.117**	**0.161**	**0.149**	**0.229**	**0.311**	**0.371**	**0.344**	**0.157**	**0.200**	**0.383**	**0.405**	**0.401**						
14	WE17_enjoyed	**0.241**	**0.229**	**0.205**	**0.384**	**0.174**	**0.271**	**0.203**	**0.216**	**0.373**	**0.330**	**0.318**	**0.357**	**0.170**					
15	WE5_excited	**0.264**	**0.255**	**0.214**	**0.302**	**0.239**	**0.349**	**0.263**	**0.151**	**0.314**	**0.343**	**0.319**	**0.401**	**0.225**	**0.648**				
16	WE8_calm	**0.132**	**0.101**	0.066	**0.294**	0.027	0.037	-0.058	**0.173**	**0.201**	**0.167**	**0.234**	**0.156**	0.048	**0.407**	**0.218**			
17	WE15_frustrated_rec	0.031	0.005	-0.051	**0.163**	-0.044	-0.019	**-0.112**	**0.098**	**0.175**	0.044	**0.096**	**0.080**	-0.008	**0.438**	**0.234**	**0.452**		
18	WE12_anxious_rec	0.012	-0.037	-0.066	**0.097**	**-0.155**	**-0.128**	**-0.162**	**0.078**	0.050	-0.046	0.025	**-0.040**	**-0.097**	**0.263**	0.043	**0.531**	**0.551**	
19	WE1_bored_rec	**0.133**	**0.125**	**0.110**	**0.316**	**0.119**	**0.198**	**0.145**	**0.187**	**0.382**	**0.239**	**0.259**	**0.271**	**0.086**	**0.607**	**0.501**	**0.297**	**0.449**	**0.255**

Note. Correlations that are statistically significant (*p* < 0.01) are in bold. Highlighted are correlations within hypothesized subscales.

Correlations for items within Individual Work:

#	Item	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
1	IB15_reread																	
2	IB11_looked	**0.176**																
3	IB7_checked	**0.366**	**0.087**															
4	IB2_write	**0.228**	**0.136**	**0.231**														
5	IB17_wrotedifways	**0.268**	0.052	**0.229**	**0.293**													
6	IC3_recall	**0.272**	0.050	**0.358**	**0.282**	0.077												
7	IC13_keepinmind	**0.405**	**0.160**	**0.472**	**0.262**	**0.172**	**0.382**											
8	IC9_why	**0.338**	0.063	**0.440**	**0.349**	**0.341**	**0.310**	**0.442**										
9	IC6_thoughtdifways	**0.280**	**0.085**	**0.368**	**0.259**	**0.553**	**0.200**	**0.264**	**0.415**									
10	IC5_identify	**0.340**	**0.162**	**0.414**	**0.281**	**0.192**	**0.403**	**0.511**	**0.416**	**0.372**								
11	IC12_critical	**0.400**	**0.256**	**0.452**	**0.336**	**0.336**	**0.307**	**0.602**	**0.527**	**0.423**	**0.449**							
12	IC16_ownwords	**0.457**	**0.125**	**0.272**	**0.236**	**0.389**	**0.139**	**0.241**	**0.314**	**0.282**	**0.318**	**0.331**						
13	IE8_enjoyed	0.071	-0.012	**0.191**	**0.142**	**0.181**	**0.096**	**0.224**	**0.328**	**0.227**	**0.146**	**0.245**	**0.096**					
14	IE14_excited	**0.141**	0.014	**0.167**	**0.196**	**0.314**	0.059	**0.168**	**0.263**	**0.274**	**0.135**	**0.228**	**0.211**	**0.559**				
15	IE10_calm	0.020	-0.060	**0.156**	0.004	0.042	0.029	**0.210**	**0.215**	**0.093**	**0.114**	**0.212**	0.003	**0.460**	**0.337**			
16	IE18_frustrated_rec	-0.037	**-0.089**	0.073	-0.009	-0.016	-0.019	**0.114**	**0.156**	0.059	0.020	**0.101**	-0.047	**0.378**	**0.262**	**0.583**		
17	IE1_anxious_rec	-0.067	-0.065	0.043	-0.074	-0.029	**-0.108**	0.065	**0.101**	0.009	-0.017	**0.091**	-0.051	**0.353**	**0.199**	**0.658**	**0.681**	
18	IE4_bored_rec	**0.092**	0.019	**0.170**	**0.126**	**0.062**	0.032	**0.221**	**0.227**	0.060	**0.086**	**0.204**	0.050	**0.423**	**0.302**	**0.285**	**0.379**	**0.267**

Note. Correlations that are statistically significant (*p* < 0.01) are in bold. Highlighted are correlations within hypothesized subscales.

Correlations for items within Group Work:

#	Item	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
1	GB3_asked																	
2	GB7_justified	**0.421**																
3	GB10_checked	**0.558**	**0.436**															
4	GB16_shared	**0.514**	**0.513**	**0.543**														
5	GB13_notes	**0.342**	**0.237**	**0.303**	**0.304**													
6	GC9_attention	**0.380**	**0.376**	**0.390**	**0.412**	**0.259**												
7	GC17_compared	**0.489**	**0.458**	**0.485**	**0.703**	**0.393**	**0.400**											
8	GC12_use	**0.489**	**0.371**	**0.505**	**0.416**	**0.397**	**0.359**	**0.411**										
9	GC2_identify	**0.561**	**0.476**	**0.427**	**0.499**	**0.293**	**0.514**	**0.504**	**0.441**									
10	GC15_connect	**0.522**	**0.478**	**0.497**	**0.618**	**0.435**	**0.430**	**0.616**	**0.531**	**0.573**								
11	GC4_critical	**0.628**	**0.470**	**0.431**	**0.516**	**0.350**	**0.475**	**0.542**	**0.466**	**0.643**	**0.567**							
12	GC6_ownwords	**0.381**	**0.519**	**0.326**	**0.383**	**0.343**	**0.302**	**0.449**	**0.386**	**0.494**	**0.503**	**0.485**						
13	GE14_enjoyed	**0.457**	**0.361**	**0.425**	**0.569**	**0.320**	**0.382**	**0.508**	**0.376**	**0.467**	**0.596**	**0.435**	**0.297**					
14	GE1_excited	**0.401**	**0.318**	**0.324**	**0.467**	**0.296**	**0.292**	**0.416**	**0.256**	**0.490**	**0.448**	**0.427**	**0.286**	**0.671**				
15	GE11_calm	**0.329**	**0.305**	**0.355**	**0.406**	**0.116**	**0.317**	**0.287**	**0.303**	**0.282**	**0.347**	**0.282**	**0.177**	**0.480**	**0.331**			
16	GE18_frustrated_rec	**0.169**	**0.132**	**0.168**	**0.237**	-0.018	**0.164**	**0.145**	**0.105**	**0.137**	**0.203**	**0.141**	0.047	**0.339**	**0.243**	**0.419**		
17	GE5_anxious_rec	**0.143**	**0.161**	**0.194**	**0.256**	-0.032	**0.136**	**0.166**	**0.087**	**0.075**	**0.164**	**0.097**	0.010	**0.313**	**0.189**	**0.576**	**0.459**	
18	GE8_bored_rec	**0.299**	**0.225**	**0.253**	**0.318**	**0.150**	**0.377**	**0.280**	**0.200**	**0.333**	**0.367**	**0.271**	**0.145**	**0.512**	**0.463**	**0.350**	**0.471**	**0.344**

Note. Correlations that are statistically significant (*p* < 0.01) are in bold. Highlighted are correlations within hypothesized subscales.

Appendix O

Other 7-factor models

The 7-factor model with Enjoyed and Frustrated

Item (abbreviated)	Factor 1: LB+LC	Factor 2: WB $_{\text{passive}}$ + WC	Factor 3: WB $_{\text{active}}$	Factor 4: IB+IC	Factor 5: GB + GC	Factor 6: E	Factor 7: “Diffi-culty”	R^2
LB7_read	**0.549**	-0.040	-0.027	**0.191**	0.014	0.008	-0.047	**0.386**
	0.055	0.045	0.022	**0.074**	0.027	0.022	0.054	**0.039**
LB10_listen	**0.698**	0.055	0.006	**0.233**	-0.038	0.011	-0.043	**0.678**
	0.054	0.038	0.020	**0.087**	0.027	0.031	0.049	**0.021**
LB2_notes	**0.582**	-0.023	0.013	-0.016	0.039	-0.010	**0.199**	**0.376**
	0.058	0.049	0.041	0.072	0.040	0.096	**0.066**	**0.040**
LB5_pictures	**0.177**	0.062	0.027	**0.146**	0.017	0.036	**0.323**	**0.224**
	0.081	0.071	0.044	**0.071**	0.034	0.083	**0.074**	**0.039**
LB13_remarks	0.143	0.017	**0.079**	**0.200**	0.069	0.109	**0.379**	**0.312**
	0.097	0.074	**0.036**	**0.088**	0.043	0.093	**0.108**	**0.056**
LC3_attention	**0.707**	-0.009	**0.050**	0.116	-0.004	0.142	**0.094**	**0.633**
	0.062	0.036	**0.024**	0.083	0.031	0.073	**0.035**	**0.025**
LC6_identify	**0.462**	0.081	-0.003	**0.316**	-0.016	**-0.058**	-0.020	**0.442**
	0.058	0.082	0.033	**0.075**	0.030	**0.027**	0.039	**0.031**
LC15_connect	**0.259**	0.125	0.053	**0.438**	0.007	0.027	-0.093	**0.454**
	0.055	0.075	0.034	**0.050**	0.029	0.067	0.060	**0.026**
LC12_critical	**0.282**	0.082	0.064	**0.426**	-0.035	**0.085**	-0.118	**0.455**
	0.071	0.092	0.033	**0.055**	0.036	**0.033**	0.079	**0.035**
LC9_ownwords	0.126	0.107	0.060	**0.347**	0.019	-0.040	**0.230**	**0.307**
	0.079	0.091	0.038	**0.075**	0.048	0.054	**0.068**	**0.032**
LE1_enjoyed	**0.253**	0.006	0.012	0.130	**-0.064**	**0.576**	0.135	**0.525**
	0.074	0.041	0.026	0.075	**0.025**	**0.080**	0.169	**0.044**
LE14_calm	0.102	0.036	0.013	**0.153**	-0.020	0.396	**-0.457**	**0.473**
	0.054	0.050	0.020	**0.048**	0.029	0.203	**0.188**	**0.034**
LE4_frustrated_rec	-0.003	-0.054	-0.019	0.033	0.005	**0.735**	-0.213	**0.597**
	0.048	0.038	0.022	0.039	0.028	**0.097**	0.264	**0.040**
LE8_bored_rec	**0.328**	0.050	-0.041	-0.032	-0.017	**0.699**	0.121	**0.673**
	0.070	0.042	0.029	0.064	0.031	**0.084**	0.182	**0.052**
WB6_volunteer	**0.073**	-0.016	**0.872**	-0.008	0.009	-0.005	-0.034	**0.751**
	0.029	0.031	**0.022**	0.020	0.022	0.019	0.026	**0.034**
WB14_shared	-0.032	**0.083**	**0.835**	-0.007	**0.045**	-0.011	0.005	**0.743**

Item (abbreviated)	Factor 1: LB+LC	Factor 2: WB passive + WC	Factor 3: WB active	Factor 4: IB+IC	Factor 5: GB + GC	Factor 6: E	Factor 7: “Diffi-culty”	R^2
	0.021	**0.039**	**0.018**	0.018	**0.021**	0.017	0.017	**0.024**
WB18_asked	**0.055**	0.038	**0.742**	0.008	0.042	-0.065	0.062	**0.589**
	0.024	0.029	**0.026**	0.021	0.023	0.034	0.032	**0.032**
WB9_listen	**0.249**	**0.594**	-0.018	-0.044	0.064	0.050	**-0.177**	**0.568**
	0.080	**0.052**	0.024	0.080	0.039	0.065	**0.049**	**0.033**
WB3_notes	**0.098**	**0.523**	-0.002	-0.068	0.010	-0.060	**0.230**	**0.394**
	0.041	**0.057**	0.031	0.049	0.041	0.063	**0.044**	**0.033**
WB11_pictures	-0.048	**0.449**	0.020	0.069	-0.034	0.108	**0.330**	**0.403**
	0.049	**0.046**	0.023	0.060	0.032	0.072	**0.093**	**0.038**
WB16_remarks	-0.046	**0.385**	**0.070**	0.054	-0.017	0.024	**0.400**	**0.396**
	0.065	**0.064**	**0.033**	0.063	0.036	0.096	**0.067**	**0.037**
WC13_answeredhead	**0.158**	**0.179**	-0.015	**0.234**	**0.110**	-0.015	**-0.123**	**0.249**
	0.061	**0.062**	0.032	**0.061**	**0.040**	0.059	**0.043**	**0.026**
WC19_attention	**0.319**	**0.426**	0.016	-0.008	0.066	**0.127**	-0.007	**0.496**
	0.051	**0.060**	0.021	0.043	0.042	**0.031**	0.028	**0.025**
WC7_identify	0.030	**0.653**	0.034	0.163	-0.019	**-0.064**	0.000	**0.549**
	0.055	**0.062**	0.022	0.097	0.028	**0.031**	0.036	**0.030**
WC10_connect	0.060	**0.672**	-0.020	0.162	-0.001	-0.006	-0.052	**0.598**
	0.060	**0.073**	0.024	0.084	0.024	0.060	0.032	**0.033**
WC4_critical	-0.035	**0.664**	-0.002	0.123	0.000	0.041	**0.085**	**0.548**
	0.035	**0.048**	0.026	0.075	0.025	0.023	**0.039**	**0.028**
WC2_ownwords	-0.065	**0.431**	0.005	0.109	0.027	-0.042	**0.178**	**0.281**
	0.043	**0.057**	0.032	0.076	0.043	0.038	**0.046**	**0.027**
WE17_enjoyed	0.082	**0.263**	**0.088**	-0.062	**0.102**	**0.538**	-0.003	**0.496**
	0.055	**0.047**	**0.021**	0.053	**0.038**	**0.036**	0.139	**0.033**
WE8_calm	0.035	**0.170**	**0.061**	0.076	**0.058**	0.336	**-0.479**	**0.441**
	0.050	**0.049**	**0.028**	0.063	**0.028**	0.209	**0.161**	**0.036**
WE15_frustrated_rec	**0.211**	0.097	-0.017	-0.113	**0.102**	**0.669**	0.063	**0.558**
	0.065	0.050	0.018	0.059	**0.041**	**0.068**	0.157	**0.052**
WE1_bored_rec	-0.069	0.049	-0.003	0.019	-0.028	**0.544**	-0.473	**0.547**
	0.046	0.057	0.027	0.036	0.031	**0.244**	0.267	**0.029**
IB15_reread	0.036	0.002	**-0.110**	**0.451**	**0.201**	-0.038	0.097	**0.314**
	0.050	0.042	**0.032**	**0.039**	**0.046**	0.042	0.055	**0.029**
IB11_looked	0.038	0.035	**-0.102**	0.107	**0.148**	-0.066	0.092	**0.073**
	0.056	0.073	**0.035**	0.055	**0.052**	0.049	0.058	**0.019**
IB7_checked	0.130	0.018	-0.016	**0.529**	**0.127**	-0.036	-0.057	**0.408**
	0.070	0.056	0.026	**0.053**	**0.060**	0.045	0.040	**0.030**
IB2_write	0.117	0.003	0.016	**0.332**	0.070	0.017	**0.261**	**0.260**
	0.061	0.042	0.039	**0.073**	0.043	0.073	**0.063**	**0.033**
IB17_wrotedifways	-0.060	-0.053	**0.088**	**0.409**	0.023	0.111	**0.423**	**0.393**
	0.080	0.055	**0.037**	**0.091**	0.026	0.112	**0.133**	**0.042**
IC3_recall	**0.175**	0.065	**-0.108**	**0.384**	**0.114**	**-0.174**	-0.105	**0.296**
	0.084	0.053	**0.038**	**0.069**	**0.051**	**0.050**	0.074	**0.038**

Item (abbreviated)	Factor 1: LB+LC	Factor 2: WB passive + WC	Factor 3: WB active	Factor 4: IB+IC	Factor 5: GB + GC	Factor 6: E	Factor 7: “Diffi-culty”	R^2
IC13_keepinmind	0.125	-0.018	**-0.087**	**0.639**	**0.132**	-0.047	**-0.127**	**0.524**
	0.076	0.039	**0.034**	**0.057**	**0.051**	0.072	**0.050**	**0.040**
IC9_why	0.018	0.020	0.018	**0.654**	0.019	0.036	0.076	**0.486**
	0.042	0.054	0.025	**0.034**	0.040	0.026	0.091	**0.031**
IC6_thoughtdifways	-0.086	0.025	**0.078**	**0.531**	-0.015	0.043	0.190	**0.351**
	0.071	0.061	**0.034**	**0.065**	0.043	0.049	0.107	**0.052**
IC5_identify	0.052	0.036	**-0.082**	**0.573**	**0.141**	-0.087	-0.019	**0.417**
	0.072	0.057	**0.037**	**0.049**	**0.060**	0.047	0.040	**0.036**
IC12_critical	0.051	0.042	-0.024	**0.683**	0.081	0.023	0.058	**0.578**
	0.048	0.038	0.021	**0.042**	0.045	0.036	0.092	**0.032**
IC16_ownwords	**-0.147**	0.047	-0.050	**0.431**	**0.136**	0.052	**0.271**	**0.325**
	0.064	0.044	0.029	**0.061**	**0.046**	0.061	**0.108**	**0.035**
IE8_enjoyed	0.006	-0.012	**0.105**	**0.426**	**-0.248**	**0.244**	-0.029	**0.315**
	0.048	0.061	**0.036**	**0.056**	**0.044**	**0.073**	0.166	**0.040**
IE10_calm	-0.090	-0.006	**0.117**	**0.372**	**-0.169**	0.258	-0.398	**0.424**
	0.062	0.049	**0.038**	**0.064**	**0.043**	0.219	0.219	**0.047**
IE18_frustrated_rec	-0.013	-0.088	-0.036	0.008	**0.106**	**0.525**	**-0.515**	**0.563**
	0.028	0.057	0.028	0.024	**0.032**	**0.245**	**0.242**	**0.029**
IE4_bored_rec	0.012	0.023	**-0.078**	**-0.094**	**0.130**	**0.666**	-0.340	**0.559**
	0.028	0.048	**0.027**	**0.034**	**0.034**	**0.139**	0.244	**0.027**
GB3_asked	**0.091**	-0.004	0.019	**-0.154**	**0.768**	**-0.074**	0.010	**0.577**
	0.043	0.032	0.022	**0.037**	**0.031**	**0.036**	0.069	**0.036**
GB7_justified	-0.020	0.003	**0.057**	**0.225**	**0.535**	0.028	-0.037	**0.431**
	0.047	0.049	**0.026**	**0.051**	**0.042**	0.054	0.056	**0.033**
GB10_checked	0.052	-0.047	0.011	-0.019	**0.706**	-0.094	**-0.095**	**0.490**
	0.041	0.039	0.028	0.036	**0.029**	0.058	**0.046**	**0.036**
GB16_shared	-0.033	-0.063	**0.167**	**0.099**	**0.737**	0.034	-0.041	**0.613**
	0.039	0.044	**0.033**	**0.050**	**0.030**	0.059	0.068	**0.026**
GB13_notes	0.000	0.100	-0.035	-0.006	**0.413**	0.000	**0.359**	**0.378**
	0.064	0.059	0.035	0.073	**0.036**	0.098	**0.055**	**0.046**
GC9_attention	**0.191**	0.032	-0.034	0.061	**0.509**	0.038	-0.043	**0.396**
	0.053	0.044	0.025	0.048	**0.046**	0.040	0.031	**0.029**
GC17_compared	**-0.084**	-0.030	**0.072**	**0.158**	**0.699**	0.016	0.078	**0.581**
	0.037	0.047	**0.029**	**0.055**	**0.028**	0.033	0.073	**0.035**
GC12_use	0.002	**0.070**	**-0.075**	-0.073	**0.662**	**-0.118**	-0.038	**0.457**
	0.036	**0.035**	**0.027**	0.045	**0.023**	**0.041**	0.057	**0.031**
GC2_identify	0.119	0.004	-0.001	0.046	**0.675**	0.022	0.040	**0.548**
	0.063	0.053	0.025	0.038	**0.027**	0.037	0.038	**0.026**
GC15_connect	-0.054	0.072	0.013	0.105	**0.721**	0.020	0.068	**0.640**
	0.028	0.056	0.017	0.055	**0.026**	0.026	0.052	**0.022**
GC4_critical	0.050	-0.011	0.008	0.053	**0.711**	-0.001	**0.102**	**0.574**
	0.061	0.049	0.025	0.040	**0.028**	0.038	**0.033**	**0.029**
GC6_ownwords	-0.079	0.066	0.005	**0.178**	**0.487**	-0.007	**0.159**	**0.393**

Item (abbreviated)	Factor 1: LB+LC	Factor 2: WB $_{\text{passive}}$ + WC	Factor 3: WB $_{\text{active}}$	Factor 4: IB+IC	Factor 5: GB + GC	Factor 6: E	Factor 7: “Diffi-culty”	R^2
	0.050	0.059	0.031	**0.058**	**0.032**	0.040	**0.051**	**0.034**
GE14_enjoyed	-0.031	0.061	**0.086**	-0.062	**0.663**	0.094	-0.032	**0.470**
	0.046	0.060	**0.028**	0.041	**0.028**	0.058	0.075	**0.035**
GE11_calm	-0.007	0.023	0.049	0.023	**0.484**	0.142	**-0.383**	**0.393**
	0.044	0.057	0.026	0.045	**0.042**	0.180	**0.113**	**0.036**
GE18_frustrated_rec	**-0.134**	-0.054	0.024	**0.198**	**-0.109**	**0.534**	-0.344	**0.507**
	0.051	0.044	0.036	**0.056**	**0.033**	**0.193**	0.283	**0.029**
GE8_bored_rec	**0.148**	-0.038	-0.028	**0.182**	-0.074	**0.406**	-0.069	**0.275**
	0.058	0.062	0.040	**0.051**	0.037	**0.063**	0.167	**0.034**

Note. For each factor, standardized loadings are presented. Standard errors are presented in the second line. Highlighted in yellow are loadings for items that represent a substantive factor. In bold are statistically significant loadings (*p* < 0.05).

Factors correlations for the 7-factor model with Enjoyed and Frustrated

	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	Factor 6	Factor 7
Factor 1: LB+LC	-						
Factor 2: WB $_{\text{passive}}$ + WC	**0.456** **(0.077)**	-					
Factor 3: WB $_{\text{active}}$	0.019 (0.046)	**0.227** **(0.056)**	-				
Factor 4: IB+IC	**0.313** **(0.073)**	**0.439** **(0.088)**	**0.209** **(0.043)**	-			
Factor 5: GB + GC	**0.248** **(0.048)**	**0.438** **(0.050)**	**0.136** **(0.032)**	**0.344** **(0.047)**	-		
Factor 6: E	**0.148** **(0.053)**	0.209 (0.108)	**0.185** **(0.046)**	**0.262** **(0.049)**	0.002 (0.045)	-	
Factor 7: “Difficulty”	0.011 (0.116)	0.186 (0.120)	0.141 (0.101)	0.029 (0.063)	**0.099** **(0.048)**	-0.045 (0.043)	-

Note. In bold are statistically significant correlations (*p* < 0.05). Standard errors are presented in parentheses.

The 7-factor model with Enjoyed and Anxious

Item (abbreviated)	Factor 1: LB+LC	Factor 2: WB $_{passive}$ + WC	Factor 3: WB $_{active}$	Factor 4: IB+IC	Factor 5: GB + GC	Factor 6: E	Factor 7: “Difficulty”	R^2
LB7_read	**0.558**	-0.059	-0.032	0.160	0.024	0.019	-0.039	**0.380**
	0.065	0.041	0.022	0.092	0.028	0.035	0.045	**0.040**
LB10_listen	**0.701**	0.036	0.015	**0.224**	-0.039	-0.006	-0.071	**0.674**
	0.054	0.032	0.019	**0.105**	0.027	0.030	0.049	**0.024**
LB2_notes	**0.636**	-0.036	-0.004	-0.103	0.070	**-0.114**	0.179	**0.405**
	0.064	0.048	0.049	0.099	0.046	**0.048**	0.147	**0.044**
LB5_pictures	**0.236**	0.040	-0.009	0.033	0.051	-0.028	**0.397**	**0.271**
	0.066	0.053	0.044	0.034	0.035	0.047	**0.063**	**0.044**
LB13_remarks	**0.233**	-0.004	0.025	0.042	**0.108**	0.021	**0.532**	**0.413**
	0.075	0.050	0.034	0.058	**0.036**	0.050	**0.064**	**0.068**
LC3_attention	**0.757**	-0.001	0.046	0.048	0.009	0.044	0.081	**0.645**
	0.045	0.050	0.027	0.078	0.039	0.034	0.075	**0.029**
LC6_identify	**0.450**	0.052	0.010	**0.322**	-0.026	-0.054	-0.037	**0.435**
	0.067	0.070	0.037	**0.097**	0.037	0.039	0.032	**0.032**
LC15_connect	**0.259**	0.082	0.047	**0.417**	0.010	**0.119**	0.002	**0.441**
	0.054	0.060	0.038	**0.054**	0.034	**0.050**	0.051	**0.027**
LC12_critical	**0.315**	0.048	0.041	**0.350**	-0.022	**0.106**	**0.223**	**0.468**
	0.064	0.071	0.037	**0.054**	0.046	**0.049**	**0.059**	**0.039**
LC9_ownwords	**0.151**	0.070	0.037	**0.271**	0.036	-0.043	**0.305**	**0.328**
	0.076	0.071	0.043	**0.055**	0.053	0.050	**0.051**	**0.037**
LE1_enjoyed	**0.326**	0.071	0.002	0.067	-0.070	**0.392**	0.226	**0.402**
	0.073	0.100	0.048	0.100	0.060	**0.115**	0.161	**0.090**
LE14_calm	**0.146**	0.021	-0.032	0.097	0.018	**0.594**	-0.153	**0.478**
	0.057	0.060	0.024	0.053	0.030	**0.042**	0.091	**0.036**
LE16_anxious_rec	-0.078	**-0.101**	0.030	0.039	-0.039	**0.744**	0.051	**0.559**
	0.057	**0.049**	0.047	0.066	0.048	**0.074**	0.045	**0.077**
LE8_bored_rec	**0.409**	0.143	-0.034	-0.067	-0.034	**0.390**	0.196	**0.439**
	0.081	0.104	0.054	0.080	0.056	**0.113**	0.144	**0.100**
WB6_volunteer	0.054	-0.008	**0.888**	-0.001	-0.008	0.004	-0.063	**0.768**
	0.029	0.018	**0.026**	0.019	0.021	0.023	0.035	**0.036**
WB14_shared	-0.049	**0.086**	**0.833**	-0.005	0.035	0.010	0.007	**0.738**
	0.028	**0.035**	**0.018**	0.020	0.020	0.015	0.020	**0.023**
WB18_asked	0.036	0.035	**0.739**	0.001	0.038	-0.056	0.038	**0.584**
	0.021	0.034	**0.030**	0.020	0.031	0.042	0.034	**0.033**
WB9_listen	**0.222**	**0.576**	0.002	0.053	0.051	0.064	**-0.163**	**0.572**
	0.076	**0.060**	0.025	0.078	0.041	0.045	**0.053**	**0.032**
WB3_notes	**0.129**	**0.484**	-0.005	-0.080	0.028	**-0.133**	**0.245**	**0.412**
	0.052	**0.061**	0.035	0.073	0.047	**0.042**	**0.103**	**0.031**
WB11_pictures	0.011	**0.427**	0.000	0.000	-0.009	0.023	**0.423**	**0.436**
	0.035	**0.044**	0.026	0.050	0.035	0.028	**0.053**	**0.031**
WB16_remarks	0.010	**0.357**	0.048	-0.018	0.003	**-0.090**	**0.474**	**0.453**
	0.048	**0.059**	0.032	0.038	0.030	**0.044**	**0.071**	**0.043**

Item (abbreviated)	Factor 1: LB+LC	Factor 2: WB $_{passive}$ + WC	Factor 3: WB $_{active}$	Factor 4: IB+IC	Factor 5: GB + GC	Factor 6: E	Factor 7: “Difficulty”	R^2
WC13_answeredhead	**0.121**	**0.162**	-0.006	**0.288**	**0.088**	0.049	-0.103	**0.248**
	0.061	**0.069**	0.036	**0.052**	**0.042**	0.043	0.055	**0.024**
WC19_attention	**0.321**	**0.431**	0.032	0.046	0.049	0.053	-0.018	**0.498**
	0.062	**0.073**	0.025	0.047	0.040	0.053	0.057	**0.027**
WC7_identify	-0.022	**0.620**	**0.065**	**0.272**	-0.052	**-0.054**	-0.016	**0.549**
	0.041	**0.063**	**0.023**	**0.095**	0.032	**0.025**	0.026	**0.031**
WC10_connect	0.038	**0.625**	-0.009	**0.216**	0.000	**0.052**	0.014	**0.585**
	0.043	**0.070**	0.022	**0.085**	0.033	**0.021**	0.025	**0.035**
WC4_critical	-0.045	**0.629**	0.010	**0.172**	-0.006	0.034	**0.133**	**0.541**
	0.039	**0.044**	0.025	**0.070**	0.028	0.030	**0.048**	**0.027**
WC2_ownwords	-0.067	**0.399**	0.005	0.114	0.030	-0.058	**0.214**	**0.285**
	0.041	**0.046**	0.033	0.079	0.042	0.042	**0.063**	**0.027**
WE17_enjoyed	0.123	**0.327**	**0.098**	-0.022	0.073	**0.331**	0.069	**0.361**
	0.072	**0.078**	**0.042**	0.057	0.047	**0.099**	0.114	**0.049**
WE8_calm	0.051	**0.161**	0.036	0.082	**0.076**	**0.499**	**-0.223**	**0.416**
	0.050	**0.045**	0.028	0.069	**0.035**	**0.064**	**0.096**	**0.034**
WE12_anxious_rec	0.075	0.109	-0.039	-0.081	**0.446**	**0.210**	-0.056	**0.278**
	0.068	0.063	0.036	0.058	**0.067**	**0.089**	0.087	**0.044**
WE1_bored_rec	0.006	0.042	**-0.063**	-0.074	0.031	**0.724**	-0.072	**0.530**
	0.046	0.046	**0.029**	0.051	0.030	**0.041**	0.073	**0.036**
IB15_reread	0.025	0.006	**-0.082**	**0.475**	**0.155**	-0.061	0.056	**0.320**
	0.048	0.042	**0.038**	**0.056**	**0.053**	0.041	0.119	**0.029**
IB11_looked	0.025	0.030	**-0.089**	**0.130**	**0.136**	**-0.093**	0.042	**0.072**
	0.060	0.072	**0.039**	**0.058**	**0.059**	**0.044**	0.063	**0.018**
IB7_checked	0.091	0.012	0.019	**0.577**	0.081	0.008	-0.088	**0.427**
	0.061	0.046	0.028	**0.042**	0.053	0.028	0.094	**0.030**
IB2_write	**0.142**	0.005	0.022	**0.289**	0.054	-0.068	**0.241**	**0.257**
	0.054	0.046	0.040	**0.067**	0.051	0.044	**0.078**	**0.031**
IB17_wrotedifways	0.005	-0.050	0.066	**0.296**	0.019	0.011	**0.478**	**0.387**
	0.053	0.061	0.040	**0.126**	0.054	0.036	**0.113**	**0.046**
IC3_recall	0.111	0.061	-0.049	**0.495**	0.045	**-0.170**	**-0.221**	**0.341**
	0.063	0.050	0.038	**0.072**	0.043	**0.040**	**0.076**	**0.042**
IC13_keepinmind	0.075	-0.030	-0.047	**0.718**	0.065	0.022	-0.147	**0.563**
	0.051	0.040	0.033	**0.041**	0.040	0.029	0.146	**0.037**
IC9_why	0.006	0.012	0.031	**0.639**	-0.011	**0.093**	0.100	**0.483**
	0.029	0.049	0.028	**0.064**	0.038	**0.036**	0.143	**0.033**
IC6_thoughtdifways	-0.079	0.033	**0.095**	**0.507**	-0.051	0.017	0.189	**0.337**
	0.056	0.063	**0.046**	**0.079**	0.054	0.039	0.148	**0.047**
IC5_identify	0.001	0.033	-0.041	**0.641**	0.083	-0.057	-0.071	**0.449**
	0.061	0.054	0.040	**0.046**	0.057	0.034	0.118	**0.034**
IC12_critical	0.036	0.023	-0.007	**0.682**	0.045	0.080	0.095	**0.581**
	0.037	0.030	0.024	**0.053**	0.036	0.043	0.141	**0.030**
IC16_ownwords	**-0.124**	0.051	-0.047	**0.396**	**0.108**	-0.010	**0.304**	**0.318**

Item (abbreviated)	Factor 1: LB+LC	Factor 2: WB$_{\text{passive}}$ + WC	Factor 3: WB$_{\text{active}}$	Factor 4: IB+IC	Factor 5: GB + GC	Factor 6: E	Factor 7: "Diffi-culty"	R^2
	0.053	0.042	0.029	**0.097**	**0.052**	0.033	**0.107**	**0.036**
IE8_enjoyed	0.047	-0.041	0.059	**0.318**	**-0.217**	**0.391**	0.185	**0.356**
	0.062	0.068	0.038	**0.063**	**0.045**	**0.068**	0.113	**0.050**
IE10_calm	-0.066	-0.080	0.034	**0.245**	-0.081	**0.680**	-0.011	**0.555**
	0.059	0.046	0.047	**0.065**	0.048	**0.082**	0.033	**0.070**
IE1_anxious_rec	-0.024	0.028	-0.062	**-0.143**	**0.373**	**0.458**	-0.052	**0.294**
	0.046	0.033	0.044	**0.050**	**0.044**	**0.063**	0.112	**0.038**
IE4_bored_rec	0.071	0.088	-0.079	-0.067	**0.112**	**0.542**	-0.126	**0.353**
	0.053	0.073	0.036	0.045	**0.040**	**0.095**	0.117	**0.055**
GB3_asked	0.079	0.002	0.022	-0.080	**0.737**	**-0.134**	-0.038	**0.568**
	0.057	0.039	0.023	0.066	**0.047**	**0.040**	0.048	**0.037**
GB7_justified	-0.020	-0.016	0.043	**0.243**	**0.511**	0.076	0.054	**0.426**
	0.061	0.054	0.031	**0.096**	**0.051**	0.041	0.069	**0.033**
GB10_checked	0.036	-0.071	0.003	0.027	**0.696**	-0.048	-0.072	**0.481**
	0.053	0.040	0.035	0.051	**0.035**	0.044	0.088	**0.036**
GB16_shared	-0.009	**-0.100**	0.122	0.071	**0.759**	**0.122**	0.102	**0.630**
	0.038	**0.044**	0.033	0.086	**0.029**	**0.039**	0.074	**0.030**
GB13_notes	0.062	0.094	-0.050	-0.052	**0.408**	**-0.152**	**0.373**	**0.405**
	0.048	0.063	0.032	0.062	**0.036**	**0.044**	**0.063**	**0.041**
GC9_attention	**0.174**	0.054	-0.026	0.120	**0.476**	-0.003	-0.069	**0.398**
	0.061	0.051	0.029	0.066	**0.058**	0.055	0.059	**0.028**
GC17_compared	-0.054	-0.059	0.037	0.127	**0.702**	0.049	**0.201**	**0.592**
	0.039	0.037	0.026	0.117	**0.048**	0.031	**0.061**	**0.036**
GC12_use	-0.020	0.058	**-0.065**	0.020	**0.624**	**-0.146**	-0.070	**0.449**
	0.054	0.036	**0.030**	0.057	**0.030**	**0.037**	0.053	**0.029**
GC2_identify	0.103	0.037	0.014	0.136	**0.612**	-0.084	-0.017	**0.542**
	0.092	0.079	0.033	0.115	**0.066**	0.054	0.080	**0.027**
GC15_connect	-0.041	0.066	-0.004	0.120	**0.709**	0.006	**0.125**	**0.642**
	0.028	0.046	0.019	0.101	**0.037**	0.028	**0.048**	**0.023**
GC4_critical	0.051	-0.001	0.005	0.101	**0.666**	**-0.071**	0.095	**0.557**
	0.094	0.068	0.027	0.130	**0.065**	**0.032**	0.065	**0.031**
GC6_ownwords	-0.078	0.066	0.004	0.215	**0.440**	**-0.073**	**0.181**	**0.391**
	0.074	0.071	0.034	0.116	**0.057**	**0.040**	**0.089**	**0.034**
GE14_enjoyed	0.001	0.058	0.054	-0.098	**0.705**	**0.131**	0.061	**0.511**
	0.056	0.072	0.030	0.052	**0.031**	**0.055**	0.093	**0.034**
GE11_calm	-0.002	-0.006	0.017	0.001	**0.540**	**0.364**	-0.193	**0.445**
	0.070	0.070	0.035	0.074	**0.059**	**0.084**	0.125	**0.039**
GE5_anxious_rec	-0.062	-0.057	0.050	-0.060	**0.364**	**0.421**	-0.180	**0.311**
	0.068	0.075	0.042	0.092	**0.075**	**0.084**	0.140	**0.041**
GE8_bored_rec	**0.213**	-0.018	-0.058	0.095	-0.041	**0.402**	0.088	**0.254**
	0.058	0.069	0.043	0.059	0.043	**0.078**	0.093	**0.050**

Note. For each factor, standardized loadings are presented. Standard errors are presented in the second line. Highlighted in yellow are loadings for items that represent a substantive factor. In bold are statistically significant loadings ($p < 0.05$).

Factors correlations for the 7-factor model with Enjoyed and Anxious

	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	Factor 6	Factor 7
Factor 1: LB+LC	-						
Factor 2: WB $_{\text{passive}}$ + WC	**0.489**	-					
	(0.067)						
Factor 3: WB $_{\text{active}}$	0.069	**0.194**	-				
	(0.059)	**(0.041)**					
Factor 4: IB+IC	**0.403**	**0.385**	**0.186**	-			
	(0.083)	**(0.102)**	**(0.052)**				
Factor 5: GB + GC	**0.229**	**0.406**	**0.152**	**0.376**	-		
	(0.043)	**(0.051)**	**(0.037)**	**(0.070)**			
Factor 6: E	**0.143**	0.075	**0.140**	0.145	-0.068	-	
	(0.028)	(0.039)	**(0.040)**	(0.079)	(0.054)		
Factor 7: “Difficulty”	0.075	**0.204**	**0.267**	**0.201**	0.051	**-0.140**	-
	(0.092)	**(0.099)**	**(0.032)**	**(0.044)**	(0.079)	**(0.002)**	

Note. In bold are statistically significant loadings ($p < 0.05$). Standard errors are presented in parentheses.

The 7-factor model with Excited and Anxious

Item (abbreviated)	Factor 1: LB+LC	Factor 2: WB $_{passive}$ + WC	Factor 3: WB $_{active}$	Factor 4: IB+IC	Factor 5: GB + GC	Factor 6: E	Factor 7: “Diffi-culty”	R^2
LB7_read	**0.557**	-0.063	-0.029	**0.175**	0.022	0.014	-0.021	**0.383**
	0.058	0.040	0.021	**0.077**	0.029	0.038	0.034	**0.040**
LB10_listen	**0.704**	0.033	0.014	**0.240**	-0.046	-0.006	-0.039	**0.563**
	0.047	0.030	0.019	**0.082**	0.026	0.029	0.035	**0.034**
LB2_notes	**0.620**	-0.043	-0.005	-0.080	0.076	**-0.138**	0.188	**0.403**
	0.055	0.042	0.044	0.093	0.046	**0.054**	0.106	**0.044**
LB5_pictures	**0.215**	0.049	-0.013	0.050	0.057	-0.066	**0.377**	**0.259**
	0.064	0.050	0.044	0.044	0.037	0.052	**0.071**	**0.043**
LB13_remarks	**0.211**	-0.011	0.006	0.057	**0.127**	-0.007	**0.553**	**0.429**
	0.065	0.040	0.027	0.048	**0.033**	0.055	**0.067**	**0.063**
LC3_attention	**0.735**	0.010	0.045	0.069	0.009	0.024	0.109	**0.635**
	0.040	0.045	0.028	0.068	0.038	0.027	0.058	**0.029**
LC6_identify	**0.445**	0.059	0.011	**0.332**	-0.027	-0.056	-0.036	**0.436**
	0.058	0.063	0.039	**0.082**	0.036	0.036	0.033	**0.031**
LC15_connect	**0.269**	0.082	0.043	**0.416**	0.010	**0.129**	0.022	**0.445**
	0.048	0.056	0.037	**0.047**	0.034	**0.043**	0.043	**0.028**
LC12_critical	**0.306**	0.055	0.028	**0.359**	-0.022	**0.104**	**0.258**	**0.481**
	0.056	0.068	0.038	**0.050**	0.046	**0.048**	**0.051**	**0.037**
LC9_ownwords	**0.144**	0.073	0.028	**0.276**	0.042	-0.059	**0.299**	**0.328**
	0.071	0.067	0.038	**0.057**	0.052	0.042	**0.053**	**0.034**
LE11_excited	**0.221**	**0.187**	0.033	-0.005	-0.046	**0.285**	**0.376**	**0.379**
	0.056	**0.084**	0.052	0.064	0.055	**0.067**	**0.124**	**0.075**
LE14_calm	**0.160**	0.029	-0.037	0.089	0.016	**0.621**	-0.083	**0.488**
	0.053	0.057	0.023	0.046	0.029	**0.038**	0.065	**0.039**
LE16_anxious_rec	-0.063	**-0.082**	0.028	0.027	-0.028	**0.752**	0.086	**0.564**
	0.041	**0.037**	0.039	0.061	0.044	**0.052**	0.050	**0.055**
LE8_bored_rec	**0.389**	**0.164**	-0.037	-0.058	-0.027	**0.341**	**0.240**	**0.411**
	0.061	**0.083**	0.053	0.052	0.046	**0.066**	**0.114**	**0.069**
WB6_volunteer	**0.051**	-0.006	**0.889**	-0.001	-0.003	0.003	-0.048	**0.773**
	0.024	0.018	**0.025**	0.018	0.020	0.025	0.037	**0.035**
WB14_shared	-0.048	**0.091**	**0.829**	-0.007	0.039	0.004	0.007	**0.734**
	0.029	**0.033**	**0.017**	0.020	0.023	0.014	0.018	**0.023**
WB18_asked	0.034	0.028	**0.735**	0.007	0.042	-0.063	0.042	**0.583**
	0.023	0.029	**0.031**	0.022	0.033	0.040	0.034	**0.033**
WB9_listen	**0.246**	**0.562**	0.004	0.043	0.053	0.069	**-0.144**	**0.563**
	0.065	**0.054**	0.024	0.037	0.041	0.041	**0.041**	**0.034**
WB3_notes	**0.120**	**0.485**	-0.006	-0.072	0.033	**-0.157**	**0.215**	**0.402**
	0.052	**0.060**	0.037	0.077	0.052	**0.045**	**0.096**	**0.033**
WB11_pictures	0.000	**0.430**	-0.009	0.003	0.005	-0.012	**0.415**	**0.436**
	0.031	**0.040**	0.027	0.034	0.035	0.033	**0.052**	**0.032**
WB16_remarks	-0.004	**0.353**	0.042	-0.005	0.016	**-0.133**	**0.439**	**0.435**
	0.049	**0.053**	0.033	0.047	0.033	**0.050**	**0.080**	**0.046**

Item (abbreviated)	Factor 1: LB+LC	Factor 2: WB passive + WC	Factor 3: WB active	Factor 4: IB+IC	Factor 5: GB + GC	Factor 6: E	Factor 7: “Diffi-culty”	R^2
WC13_answeredhead	**0.134**	**0.162**	-0.003	**0.274**	**0.092**	0.057	**-0.105**	**0.248**
	0.052	**0.057**	0.036	**0.048**	**0.042**	0.040	**0.051**	**0.024**
WC19_attention	**0.330**	**0.420**	0.029	0.039	0.061	0.051	0.011	**0.494**
	0.055	**0.065**	0.024	0.035	0.043	0.041	0.051	**0.028**
WC7_identify	-0.005	**0.622**	**0.067**	**0.254**	-0.044	**-0.057**	-0.036	**0.552**
	0.035	**0.053**	**0.022**	**0.076**	0.031	**0.024**	0.027	**0.028**
WC10_connect	0.062	**0.623**	-0.009	**0.193**	0.012	**0.050**	0.010	**0.588**
	0.041	**0.062**	0.023	**0.062**	0.035	**0.023**	0.024	**0.033**
WC4_critical	-0.043	**0.640**	0.008	**0.162**	-0.002	0.020	**0.118**	**0.549**
	0.037	**0.038**	0.024	**0.057**	0.031	0.030	**0.050**	**0.027**
WC2_ownwords	-0.071	**0.408**	0.003	0.111	0.035	-0.076	**0.181**	**0.283**
	0.045	**0.048**	0.035	0.079	0.045	0.042	**0.074**	**0.028**
WE5_excited	0.010	**0.387**	**0.116**	-0.014	0.007	0.130	**0.238**	**0.311**
	0.058	**0.063**	**0.052**	0.059	0.041	0.070	**0.102**	**0.035**
WE8_calm	0.071	**0.156**	0.034	0.079	**0.069**	**0.520**	**-0.163**	**0.408**
	0.046	**0.038**	0.027	0.051	**0.033**	**0.050**	**0.063**	**0.034**
WE12_anxious_rec	0.076	0.106	-0.031	-0.069	**0.429**	**0.188**	-0.045	**0.252**
	0.067	0.057	0.038	0.063	**0.067**	**0.072**	0.068	**0.041**
WE1_bored_rec	0.017	0.056	**-0.066**	-0.079	0.030	**0.737**	-0.015	**0.536**
	0.046	0.038	**0.029**	0.055	0.031	**0.031**	0.032	**0.035**
IB15_reread	0.014	0.002	**-0.082**	**0.491**	**0.142**	-0.058	0.050	**0.327**
	0.042	0.037	**0.037**	**0.041**	**0.051**	0.035	0.070	**0.028**
IB11_looked	0.017	0.028	**-0.085**	**0.142**	**0.126**	**-0.097**	0.022	**0.071**
	0.057	0.065	**0.040**	**0.057**	**0.058**	**0.045**	0.058	**0.017**
IB7_checked	0.096	0.007	0.025	**0.577**	0.079	0.018	-0.095	**0.427**
	0.056	0.045	0.027	**0.040**	0.051	0.031	0.059	**0.030**
IB2_write	**0.126**	-0.002	0.022	**0.304**	0.060	**-0.087**	**0.227**	**0.256**
	0.053	0.039	0.041	**0.058**	0.046	**0.043**	**0.060**	**0.032**
IB17_wrotedifways	-0.030	-0.045	0.055	**0.320**	0.022	-0.017	**0.483**	**0.408**
	0.048	0.052	0.042	**0.074**	0.050	0.028	**0.076**	**0.041**
IC3_recall	**0.121**	0.058	-0.040	**0.490**	0.034	**-0.152**	**-0.237**	**0.338**
	0.057	0.046	0.034	**0.059**	0.042	**0.040**	**0.049**	**0.042**
IC13_keepinmind	0.082	-0.033	-0.037	**0.721**	0.056	0.036	**-0.164**	**0.566**
	0.047	0.038	0.030	**0.036**	0.036	0.035	**0.082**	**0.035**
IC9_why	0.010	0.014	0.036	**0.635**	-0.008	**0.082**	0.085	**0.471**
	0.031	0.048	0.028	**0.044**	0.030	**0.031**	0.084	**0.037**
IC6_thoughtdifways	-0.093	0.034	**0.093**	**0.519**	-0.052	0.004	0.182	**0.344**
	0.052	0.058	**0.046**	**0.058**	0.050	0.035	0.105	**0.047**
IC5_identify	0.002	0.031	-0.033	**0.651**	0.068	-0.049	-0.098	**0.454**
	0.051	0.049	0.038	**0.041**	0.052	0.031	0.072	**0.032**
IC12_critical	0.033	0.025	-0.001	**0.694**	0.036	0.075	0.077	**0.583**
	0.033	0.027	0.023	**0.037**	0.036	0.041	0.081	**0.030**
IC16_ownwords	**-0.141**	0.042	-0.055	**0.417**	0.103	-0.022	**0.300**	**0.335**

Item (abbreviated)	Factor 1: LB+LC	Factor 2: WB$_{passive}$ + WC	Factor 3: WB$_{active}$	Factor 4: IB+IC	Factor 5: GB + GC	Factor 6: E	Factor 7: “Diffi-culty”	R^2
	0.047	0.038	0.029	**0.063**	0.053	0.032	**0.066**	**0.037**
IE14_excited	0.027	-0.057	**0.105**	**0.247**	**-0.116**	**0.265**	**0.353**	**0.308**
	0.040	0.057	**0.053**	**0.067**	**0.056**	**0.053**	**0.119**	**0.055**
IE10_calm	-0.045	-0.068	0.032	**0.237**	-0.081	**0.701**	0.031	**0.570**
	0.048	0.039	0.038	**0.058**	0.046	**0.057**	0.039	**0.054**
IE1_anxious_rec	-0.019	0.038	-0.055	**-0.134**	**0.357**	**0.450**	-0.045	**0.276**
	0.041	0.032	0.038	**0.051**	**0.041**	**0.055**	0.074	**0.039**
IE4_bored_rec	0.073	0.092	**-0.078**	-0.054	**0.098**	**0.527**	-0.074	**0.318**
	0.047	0.054	**0.031**	0.042	**0.035**	**0.063**	0.078	**0.045**
GB3_asked	**0.073**	0.003	0.020	**-0.114**	**0.761**	**-0.112**	-0.051	**0.579**
	0.036	0.030	0.023	**0.050**	**0.039**	**0.031**	0.057	**0.035**
GB7_justified	-0.021	-0.011	0.041	**0.215**	**0.531**	**0.094**	0.045	**0.430**
	0.044	0.039	0.029	**0.062**	**0.040**	**0.040**	0.049	**0.032**
GB10_checked	0.033	-0.070	0.006	0.008	**0.706**	-0.027	-0.096	**0.485**
	0.052	0.042	0.037	0.044	**0.035**	0.041	0.106	**0.036**
GB16_shared	-0.015	-0.092	**0.118**	0.058	**0.762**	**0.129**	0.092	**0.620**
	0.032	0.039	**0.029**	0.050	**0.026**	**0.037**	0.078	**0.028**
GB13_notes	0.043	0.082	-0.059	-0.046	**0.421**	**-0.169**	**0.358**	**0.401**
	0.054	0.051	0.035	0.042	**0.034**	**0.046**	**0.072**	**0.041**
GC9_attention	**0.173**	0.053	-0.026	0.111	**0.479**	0.007	-0.068	**0.396**
	0.054	0.046	0.027	0.058	**0.054**	0.044	0.048	**0.028**
GC17_compared	-0.063	-0.053	0.031	0.114	**0.710**	0.050	**0.181**	**0.586**
	0.033	0.037	0.026	0.067	**0.040**	0.031	**0.067**	**0.036**
GC12_use	-0.024	0.055	**-0.063**	0.002	**0.632**	**-0.126**	-0.101	**0.449**
	0.043	0.036	**0.032**	0.041	**0.026**	**0.036**	0.075	**0.029**
GC2_identify	0.097	0.038	0.013	0.097	**0.641**	-0.069	-0.021	**0.555**
	0.068	0.056	0.032	0.089	**0.054**	0.041	0.047	**0.025**
GC15_connect	**-0.052**	0.070	-0.004	0.116	**0.703**	0.002	**0.100**	**0.627**
	0.025	0.047	0.020	0.062	**0.030**	0.022	**0.051**	**0.023**
GC4_critical	0.042	0.002	-0.002	0.061	**0.699**	**-0.054**	**0.090**	**0.573**
	0.067	0.049	0.027	0.090	**0.053**	**0.025**	**0.035**	**0.031**
GC6_ownwords	-0.092	0.073	-0.002	**0.191**	**0.462**	-0.065	**0.162**	**0.397**
	0.052	0.056	0.030	**0.075**	**0.043**	0.037	**0.052**	**0.034**
GE1_excited	0.017	0.056	**0.108**	-0.092	**0.560**	0.073	**0.138**	**0.368**
	0.052	0.047	**0.038**	0.072	**0.042**	0.053	**0.049**	**0.030**
GE11_calm	0.007	-0.006	0.019	0.005	**0.522**	**0.397**	-0.171	**0.438**
	0.054	0.054	0.029	0.058	**0.054**	**0.060**	0.089	**0.037**
GE5_anxious_rec	-0.047	-0.060	0.052	-0.053	**0.346**	**0.452**	-0.154	**0.311**
	0.050	0.055	0.038	0.071	**0.067**	**0.058**	0.094	**0.040**
GE8_bored_rec	**0.218**	-0.013	-0.053	**0.109**	-0.049	**0.359**	0.118	**0.229**
	0.050	0.064	0.046	**0.048**	0.040	**0.060**	0.087	**0.039**

Note. For each factor, standardized loadings are presented. Standard errors are presented in the second line. Highlighted in yellow are loadings for items that represent a substantive factor. In bold are statistically significant loadings ($p < 0.05$).

Factors correlations for the 7-factor model with Excited and Anxious

	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	Factor 6	Factor 7
Factor 1: LB+LC	-						
Factor 2: WB $_{passive}$ + WC	**0.466** **(0.057)**	-					
Factor 3: WB $_{active}$	0.055 (0.047)	**0.195** **(0.039)**	-				
Factor 4: IB+IC	**0.375** **(0.065)**	**0.401** **(0.067)**	**0.182** **(0.039)**	-			
Factor 5: GB + GC	**0.232** **(0.045)**	**0.411** **(0.054)**	**0.151** **(0.041)**	**0.408** **(0.064)**	-		
Factor 6: E	**0.131** **(0.028)**	0.048 (0.041)	**0.125** **(0.042)**	0.105 (0.062)	0.091 (0.050)	-	
Factor 7: “Difficulty”	0.058 (0.059)	**0.221** **(0.051)**	**0.285** **(0.031)**	**0.216** **(0.039)**	0.051 (0.051)	**-0.141** **(0.039)**	-

Note. In bold are statistically significant loadings ($p < 0.05$). Standard errors are presented in parentheses.

Appendix P

Weights for each subscale

Subscales and Items	Weights
Behavioral/Cognitive Engagement in Lecture	
LB7_read	0.70
LB2_notes	0.90
LB5_pictures	1.35
LB13_remarks	1.35
LC3_attention	0.80
LC6_identify	0.85
LC15_connect	0.85
LC12_critical	1.05
LC9_ownwords	1.15
Behavioral/Cognitive Engagement in Whole-Class Interaction	
WB3_notes	1.05
WB11_pictures	1.20
WB16_remarks	1.35
WC19_attention	0.75
WC7_identify	0.85
WC10_connect	0.80
WC4_critical	0.95
WC2_ownwords	1.05
Behavioral/Cognitive Engagement in Individual Work	
IB15_reread	0.85
IB7_checked	0.85
IB2_write	1.15
IB17_wrotedifways	1.45
IC3_recall	0.80
IC13_keepinmind	0.80
IC9_why	1.00
IC5_identify	0.85
IC6_thoughtdifways	1.20
IC12_critical	0.90
IC16_ownwords	1.15
Behavioral/Cognitive Engagement in Group Work	
GB3_asked	0.90
GB7_justified	1.00

Subscales and Items	Weights
GB10_checked	0.80
GB13_notes	1.40
GC12_use	1.00
GC17_compared	1.00
GC9_attention	0.85
GC2_identify	0.90
GC15_connect	1.00
GC4_critical	1.00
GC6_ownwords	1.15
Active Behavioral Engagement in Whole-Class Interaction	
WB6_volunteer	1.00
WB14_shared	1.00
WB18_asked	1.00
Emotional Engagement (in any instruction type)	
Excited	1.25
Calm	0.80
Not Frustrated	0.90
Not Bored	1.05

Appendix Q

Correlations between weighted engagement composites

Correlations between weighted subscale engagement composites

	LBC	LE	WBC passive	WE	IBC	IE	GBC	GE
LBC								
LE	**0.345**							
WBC passive	**0.687**	**0.278**						
WE	**0.294**	**0.741**	**0.323**					
IBC	**0.577**	**0.257**	**0.513**	**0.274**				
IE	**0.192**	**0.589**	**0.090**	**0.450**	**0.259**			
GBC	**0.419**	0.071	**0.467**	**0.167**	**0.489**	-0.022		
GE	**0.212**	**0.317**	**0.239**	**0.451**	**0.278**	**0.248**	**0.497**	
WB active	**0.245**	**0.151**	**0.286**	**0.196**	**0.234**	**0.161**	**0.198**	**0.175**

Note. LBC = Behavioral/Cognitive Engagement in Lecture; LE = Emotional Engagement in Lecture; WBC passive = Behavioral (passive)/Cognitive Engagement in Whole-Class Interaction; WE = Emotional Engagement in Whole-Class Interaction; IBC = Behavioral/Cognitive Engagement in Individual Work; IE = Emotional Engagement in Individual Work; GBC = Behavioral/Cognitive Engagement in Group Work; GE = Emotional Engagement in Group Work; WB active = Active Behavioral Engagement in Whole-Class Interaction. Statistically significant correlations (*p* < 0.01) are in bold.

Correlations between weighted dimension and instruction type engagement composites

	L	W	I	BC
Instruction type composites:				
Engagement in Lecture (L)				
Engagement in Whole-Class Interaction (W)	**0.693**			
Engagement in Individual Work	**0.624**	**0.513**		
Engagement in Group Work (G)	**0.352**	**0.451**	**0.339**	
Dimension composites				
Behavioral/Cognitive Engagement (BC)				
Emotional Engagement (E)				**0.309**

Note. Statistically significant correlations (*p* < 0.01) are in bold.

References

AERA, APA, & NCME. (2014). *Standards for educational and psychological testing*. American Educational Research Association.

Al-Musalli, A. M. (2015). Taxonomy of lecture note-taking skills and subskills. *International Journal of Listening*, *29*(3), 134–147. https://doi.org/10.1080/10904018.2015.1011643

An ILA definition of listening. (1995). *ILA Listening Post*, *53*(1), 4–5.

Appleton, J. J., Christenson, S. L., Kim, D., & Reschly, A. L. (2006). Measuring cognitive and psychological engagement: Validation of the Student Engagement Instrument. *Journal of School Psychology*, *44*(5), 427–445. https://doi.org/10.1016/j.jsp.2006.04.002

Archambault, I., Janosz, M., Fallu, J.-S., & Pagani, L. S. (2009). Student engagement and its relationship with early high school dropout. *Journal of Adolescence*, *32*(3), 651–670. https://doi.org/10.1016/j.adolescence.2008.06.007

Aryadoust, V., Goh, C. C. M., & Kim, L. O. (2012). Developing and validating an Academic Listening Questionnaire. *Psychological Test and Assessment Modeling; Lengerich*, *54*(3), 227–256.

Assunção, H., Lin, S.-W., Sit, P.-S., Cheung, K.-C., Harju-Luukkainen, H., Smith, T., Maloa, B., Campos, J. Á. D. B., Ilic, I. S., Esposito, G., Francesca, F. M., &

Marôco, J. (2020). University Student Engagement Inventory (USEI): Transcultural validity evidence across four continents. *Frontiers in Psychology, 10*. https://doi.org/10.3389/fpsyg.2019.02796

Awang Hashim, R., & Murad Sani, A. (2008). A confirmatory factor analysis of a newly integrated multidimensional school engagement scale. *Malaysian Journal of Learning & Instruction, 5*, 21–40.

Babcock, P. (2010). Real costs of nominal grade inflation? New evidence from student course evaluations. *Economic Inquiry, 48*(4), 983–996. https://doi.org/10.1111/j.1465-7295.2009.00245.x

Bainter, S. A., & Bollen, K. A. (2014). Interpretational confounding or confounded interpretations of causal indicators? *Measurement: Interdisciplinary Research and Perspectives*, *12*(4), 125–140. https://doi.org/10.1080/15366367.2014.968503

Barrett, L. F., & Russell, J. A. (1998). Independence and bipolarity in the structure of current affect. *Journal of Personality and Social Psychology, 74*(4), 967–984. https://doi.org/10.1037/0022-3514.74.4.967

Bartlett, F. C. (1932). *Remembering: A study in experimental and social psychology*. Cambridge University Press.

Bennett, N., & Dunne, E. (1991). The nature and quality of talk in co-operative classroom groups. *Learning and Instruction*, *1*(2), 103–118. https://doi.org/10.1016/0959-4752(91)90021-Y

Bollen, K. A., & Bauldry, S. (2011). Three Cs in measurement models: Causal indicators, composite indicators, and covariates. *Psychological Methods*, *16*(3), 265–284. https://doi.org/10.1037/a0024448

Bollen, K. A., & Diamantopoulos, A. (2017). In defense of causal-formative indicators: A minority report. *Psychological Methods*, *22*(3), 581–596. https://doi.org/10.1037/met0000056

Bollen, K. A., & Lennox, R. (1991). Conventional wisdom on measurement: A structural equation perspective. *Psychological Bulletin*, *110*(2), 305–314. https://doi.org/10.1037/0033-2909.110.2.305

Bonner, S. M. (2013). Mathematics strategy use in solving test items in varied formats. *Journal of Experimental Education*, *81*(3), 409–428. https://doi.org/10.1080/00220973.2012.727886

Borsboom, D., Mellenbergh, G. J., & Van Heerden, J. (2003). The theoretical status of latent variables. *Psychological Review*, *110*(2), 203–219.

Boyle, G. J. (1991). Does item homogeneity indicate internal consistency or item redundancy in psychometric scales? *Personality and Individual Differences*, *12*(3), 291–294. https://doi.org/10.1016/0191-8869(91)90115-R

Brewer, M. B., & Lui, L. N. (1996). Use of sorting tasks to assess cognitive structures. In N. Schwartz & S. Sudman (Eds.), *Answering questions: Methodology for determining cognitive and communicative processes in survey research* (pp. 373–385). Jossey-Bass.

Broughton, S., Sinatra, G., & Nussbaum, E. (2013). “Pluto has been a planet my whole life!” Emotions, attitudes, and conceptual change in elementary students’ learning about Pluto’s reclassification. *Research in Science Education, 43*(2), 529–550. https://doi.org/10.1007/s11165-011-9274-x

Burch, G. F., Heller, N. A., Burch, J. J., Freed, R., & Steed, S. A. (2015). Student engagement: Developing a conceptual framework and survey instrument. *Journal of Education for Business*, *90*(4), 224–229. https://doi.org/10.1080/08832323.2015.1019821

Burt, R. S. (1976). Interpretational confounding of unobserved variables in structural equation models. *Sociological Methods & Research*, *5*(1), 3–52. https://doi.org/10.1177/004912417600500101

Canpolat, M., Kuzu, S., Yildirim, B., & Canpolat, S. (2015). Active listening strategies of academically successful university students. *Eurasian Journal of Educational Research*, *60*, 163–180. https://doi.org/10.14689/ejer.2015.60.10

Carraher, D., & Schliemann, A. (2002). The transfer dilemma. *Journal of the Learning Sciences*, *11*(1), 1–24. https://doi.org/10.1207/S15327809JLS1101_1

Cavanagh, R.F., & Kennish, P. (2009). *Quantifying student engagement in classroom learning: Student learning capabilities and the expectations of their learning*. annual international conference of the Australian Association for Research in Education, Canberra.

Cavanagh, Robert F. (2015). A unified model of student engagement in classroom learning and classroom learning environment: One measure and one underlying construct. *Learning Environments Research*, *18*(3), 349–361. https://doi.org/10.1007/s10984-015-9188-z

Charland, P., Léger, P.-M., Sénécal, S., Courtemanche, F., Mercier, J., Skelling, Y., & Labonté-Lemoyne, E. (2015). Assessing the multiple dimensions of engagement to characterize learning: A neurophysiological perspective. *Journal of Visualized Experiments*, *101*. https://doi.org/10.3791/52627

Chen, B., Zaebst, D., & Seel, L. (2005). A macro to calculate kappa statistics for categorizations by multiple raters. *Proceeding of the 30th Annual SAS Users Group International Conference*, 155–30.

Chi, M. T. H. (2000). Self-explaining: The dual processes of generating inference and repairing mental models. In R. Glaser (Ed.), *Advances in instructional psychology: Educational design and cognitive science* (Vol. 5, pp. 161–238). Lawrence Erlbaum Associates.

Christenson, S. L., Reschly, A. L., & Wylie, C. (2012). *Handbook of research on student engagement*. Springer Science & Business Media.

Cizek, G. J., Bowen, D., & Church, K. (2010). Sources of validity evidence for educational and psychological tests: A follow-up study. *Educational and Psychological Measurement*, *70*(5), 732–743. https://doi.org/10.1177/0013164410379323

Clark, L. A., & Watson, D. (1996). Constructing validity: Basic issues in objective scale development. *Psychological Assessment*, *7*(3), 309–19. https://doi.org/10.1037/1040-3590.7.3.309

Comrey, A. L., & Lee, H. B. (1992). *A first course in factor analysis, 2nd ed*. Lawrence Erlbaum Associates, Inc.

Connell, J. P. (1990). Context, self, and action: A motivational analysis of self-system processes across the life-span. In D. Cicchetti (Ed.), *The Self in Transition: Infancy to Childhood* (pp. 61–97). University of Chicago Press.

Connell, J. P., & Wellborn, J. G. (1991). Competence, autonomy, and relatedness: A motivational analysis of self-system processes. *Self-Processes and Development*, *23*, 43–77.

Cooper, L. O., & Buchanan, T. (2010). Listening competency on campus: A psychometric analysis of student listening. *International Journal of Listening*, *24*(3), 141–163. https://doi.org/10.1080/10904018.2010.508681

Csikszentmihalyi, M. (2000). *Beyond boredom and anxiety: Experiencing flow in work and play*. Wiley.

Csikszentmihalyi, Mihaly. (1990). *Flow: The psychology of optimal experience*. HarperPerennial.

Csikszentmihalyi, Mihaly. (1997). *Finding flow: The psychology of engagement with everyday life* (1st ed..). BasicBooks.

Dermitzaki, I., Leondari, A., & Goudas, M. (2009). Relations between young students' strategic behaviours, domain-specific self-concept, and performance in a problem-solving situation. *Learning and Instruction*, *19*(2), 144–157. https://doi.org/10.1016/j.learninstruc.2008.03.002

Derry, S. J. (1996). Cognitive schema theory in the constructivist debate. *Educational Psychologist, 31*(3/4), 163.

DeYoung, C. G., Quilty, L. C., & Peterson, J. B. (2007). Between facets and domains: 10 aspects of the Big Five. *Journal of Personality and Social Psychology, 93*(5), 880–896. https://doi.org/10.1037/0022-3514.93.5.880

Dimitrov, D. M. (2012). *Statistical methods for validation of assessment scale data in counseling and related fields*. http://site.ebrary.com/lib/georgemason/Doc?id=10964414

diSessa, A. A. (1993). Toward an epistemology of physics. *Cognition and Instruction, 10*(2/3), 105–225.

Douglas, H. E., Bore, M., & Munro, D. (2016). Coping with university education: The relationships of time management behaviour and work engagement with the five factor model aspects. *Learning and Individual Differences, 45*, 268–274. https://doi.org/10.1016/j.lindif.2015.12.004

Edwards, J. R. (2011). The fallacy of formative measurement. *Organizational Research Methods, 14*(2), 370–388. https://doi.org/10.1177/1094428110378369

Ekman, P. (2016). What scientists who study emotion agree about. *Perspectives on Psychological Science, 11*(1), 31–34. https://doi.org/10.1177/1745691615596992

Ekman, P., & Cordaro, D. (2011). What is meant by calling emotions basic. *Emotion Review, 3*(4), 364–370. https://doi.org/10.1177/1754073911410740

Elliot, A. J., McGregor, H. A., & Gable, S. (1999). Achievement goals, study strategies, and exam performance: A mediational analysis. *Journal of Educational Psychology, 91*(3), 549–563. https://doi.org/10.1037/0022-0663.91.3.549

F1rst Gen Mason. (n.d.). F1rst Gen Mason. Retrieved January 5, 2018, from http://www.f1rstgen.org/about-us.html

FA09: Pell Grant Report. (n.d.). Retrieved January 5, 2018, from http://research.schev.edu/fair/pell_dom_report.asp

Fabrigar, L. R., & Wegener, D. T. (2011). *Exploratory factor analysis: Exploratory factor analysis*. Oxford University Press, Incorporated. http://ebookcentral.proquest.com/lib/gmu/detail.action?docID=1036308

Fall, A.-M., & Roberts, G. (2012). High school dropouts: Interactions between social context, self-perceptions, school engagement, and student dropout. *Journal of Adolescence*, *35*(4), 787–798. https://doi.org/10.1016/j.adolescence.2011.11.004

Fassinger, P. A. (1995). Understanding classroom interaction: Students' and professors' contributions to students' silence. *The Journal of Higher Education*, *66*(1), 82–96. https://doi.org/10.2307/2943952

Felix, R., & Garcia-Vega, J. (2012). Quality of life in Mexico: A formative measurement approach. *Applied Research in Quality of Life*, *7*(3), 223–238. https://doi.org/10.1007/s11482-011-9164-4

Finn, J. D., & Zimmer, K. S. (2012). Student engagement: What is it? Why does it matter? In S. L. Christenson, A. L. Reschly, & C. Wylie (Eds.), *Handbook of* *research on student engagement* (pp. 97–131). Springer US. https://doi.org/10.1007/978-1-4614-2018-7_5

Fleiss, J. L. (1981). *Statistical methods for rates and proportions* (2nd edition). Wiley-Interscience.

Fontana, P. C., Cohen, S. D., & Wolvin, A. D. (2015). Understanding listening competency: A systematic review of research scales. *International Journal of Listening*, *29*(3), 148–176. https://doi.org/10.1080/10904018.2015.1015226

Ford, W. S. Z., Wolvin, A. D., & Chung, S. (2000). Students' self-perceived listening competencies in the basic speech communication course. *International Journal of Listening, 14*(1), 1.

Forsyth, B., Rothgeb, J. M., & Willis, G. B. (2004). Does pretesting make a difference? An experimental test. In *Methods for testing and evaluating survey questionnaires* (pp. 525–546). Wiley-Blackwell. https://doi.org/10.1002/0471654728.ch25

Fredricks, J. A., Blumenfeld, P. C., Friedel, J., & Paris, A. (2003, March 11). *School engagement. Paper presented at the Indicators of Positive Development Conference, Washington, DC.*

Fredricks, J. A., Blumenfeld, P. C., & Paris, A. H. (2004). School engagement: Potential of the concept, state of the evidence. *Review of Educational Research, 74*, 59–109. https://doi.org/10.3102/00346543074001059

Gasiewski, J. A., Eagan, M. K., Garcia, G. A., Hurtado, S., & Chang, M. J. (2012). From gatekeeping to engagement: A multicontextual, mixed method study of student academic engagement in introductory STEM courses. *Research in Higher Education*, *53*(2), 229–261. https://doi.org/10.1007/s11162-011-9247-y

Gehlbach, H., & Brinkworth, M. E. (2011). Measure twice, cut down error: A process for enhancing the validity of survey scales. *Review of General Psychology*, *15*, 380–387. https://doi.org/10.1037/a0025704

George Mason University Strategic Plan. (n.d.). Retrieved January 5, 2018, from https://strategicplan.gmu.edu/

Gibson, W. A. (1960). Nonlinear factors in two dimensions. *Psychometrika*, *25*(4), 381–392. https://doi.org/10.1007/BF02289755

Gick, M. L. (1986). Problem-solving strategies. *Educational Psychologist*, *21*(1/2), 99.

Gick, M. L., & Holyoak, K. J. (1983). Schema induction and analogical transfer. *Cognitive Psychology*, *15*(1), 1–38. https://doi.org/10.1016/0010-0285(83)90002-6

Gillies, R. M. (2004). The effects of cooperative learning on junior high school students during small group learning. *Learning and Instruction*, *14*(2), 197–213. https://doi.org/10.1016/S0959-4752(03)00068-9

Gillies, R. M., & Khan, A. (2009). Promoting reasoned argumentation, problem-solving and learning during small-group work. *Cambridge Journal of Education*, *39*(1), 7–27. https://doi.org/10.1080/03057640802701945

Goetz, T., Zirngibl, A., Pekrun, R., & Hall, N. (2003). Emotions, learning and achievement from an educational-psychological perspective. In P. Mayring & C.

Von Rhoeneck (Eds.), *Learning emotions: The influence of affective factors on classroom learning* (pp. 9–28). Peter Lang.

Green, D. P., Goldman, S. L., & Salovey, P. (1993). Measurement error masks bipolarity in affect ratings. *Journal of Personality and Social Psychology*, *64*(6), 1029–1041. https://doi.org/10.1037/0022-3514.64.6.1029

Greene, B. A., Miller, R. B., Crowson, H. M., Duke, B. L., & Akey, K. L. (2004). Predicting high school students' cognitive engagement and achievement: Contributions of classroom perceptions and motivation. *Contemporary Educational Psychology*, *29*(4), 462–482. https://doi.org/10.1016/j.cedpsych.2004.01.006

Greeno, J. (1977). Process of understanding in problem solving. In N. J. Castellan, D. B. Pisoni, & G. R. Potts (Eds.), *Cognitive theory* (Vol. 2, pp. 43–83). Lawrence Erlbaum Associates, Inc.

Gunuc, S., & Kuzu, A. (2015). Student engagement scale: Development, reliability and validity. *Assessment & Evaluation in Higher Education*, *40*(4), 587–610. https://doi.org/10.1080/02602938.2014.938019

Halone, K. K., Cunconan, T. M., Coakley, C. G., & Wolvin, A. D. (1998). Toward the establishment of general dimensions underlying the listening process. *International Journal of Listening*, *12*(1), 12–28. https://doi.org/10.1080/10904018.1998.10499016

Handelsman, M. M., Briggs, W. L., Sullivan, N., & Towler, A. (2005). A measure of college student course engagement. *The Journal of Educational Research*, *98*(3), 184–192. https://doi.org/10.3200/JOER.98.3.184-192

Hegarty, M., Mayer, R. E., & Monk, C. A. (1995). Comprehension of arithmetic word problems: A comparison of successful and unsuccessful problem solvers. *Journal of Educational Psychology*, *87*(1), 18–32. https://doi.org/10.1037/0022-0663.87.1.18

Hospel, V., Galand, B., & Janosz, M. (2016). Multidimensionality of behavioural engagement: Empirical support and implications. *International Journal of Educational Research*, *77*, 37–49. https://doi.org/10.1016/j.ijer.2016.02.007

Howell, R. D., & Breivik, E. (2016). Causal indicator models have nothing to do with measurement. *Measurement: Interdisciplinary Research and Perspectives*, *14*(4), 167–169. https://doi.org/10.1080/15366367.2016.1251271

Howell, R. D., Breivik, E., & Wilcox, J. B. (2007). Reconsidering formative measurement. *Psychological Methods*, *12*(2), 205–218. https://doi.org/10.1037/1082-989X.12.2.205

Http://www.appliedmissingdata.com/littles-mcar-test.sas. (n.d.). Retrieved April 6, 2020, from http://www.appliedmissingdata.com/littles-mcar-test.sas

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 1–55. https://doi.org/10.1080/10705519909540118

Huang, J. L., Bowling, N. A., Liu, M., & Li, Y. (2015). Detecting insufficient effort responding with an infrequency scale: Evaluating validity and participant reactions. *Journal of Business and Psychology*, *30*(2), 299–311. https://doi.org/10.1007/s10869-014-9357-6

Huang, J. L., Curran, P. G., Keeney, J., Poposki, E. M., & DeShon, R. P. (2012). Detecting and deterring insufficient effort responding to surveys. *Journal of Business and Psychology*, *27*(1), 99–114. https://doi.org/10.1007/s10869-011-9231-8

Imhof, M. (1998). What makes a good listener? Listening behavior in instructional settings. *International Journal of Listening*, *12*(1), 81–105. https://doi.org/10.1080/10904018.1998.10499020

Imhof, M. (2010). What is going on in the mind of a listener? The cognitive psychology of listening. In A. D. Wolvinessor (Ed.), *Listening and human communication in the 21st century* (pp. 97–126). Wiley-Blackwell. https://doi.org/10.1002/9781444314908.ch4

Inda-Caro, M., Maulana, R., Fernández-García, C.-M., Peña-Calvo, J.-V., Rodríguez-Menéndez, M. del C., & Helms-Lorenz, M. (2018). Validating a model of effective teaching behaviour and student engagement: Perspectives from Spanish students. *Learning Environments Research*. https://doi.org/10.1007/s10984-018-9275-z

Ing, M., & Victorino, C. (2016). Differences in classroom engagement of Asian American engineering students. *Journal of Engineering Education, 105*(3), 431–451. https://doi.org/10.1002/jee.20126

Izard, C. E., Dougherty, F. E., Bloxom, B. M., & Kotsch, W. E. (1971). *The Differential Emotions Scale: A method of measuring the subjective experience of discrete emotions*. Vanderbilt University Press.

Jones, E. S., Mullen, R., & Hardy, L. (2019). Measurement and validation of a three factor hierarchical model of competitive anxiety. *Psychology of Sport and Exercise*, *43*, 34–44. https://doi.org/10.1016/j.psychsport.2018.12.011

Kane, M. T. (2006). Validation. In R. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). American Council on Education and Praeger.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, *50*(1), 1–73. https://doi.org/10.1111/jedm.12000

Kane, M. T. (2015). Validation strategies: Delineating and validating proposed interpretations and uses of test scores. In *Handbook of test development* (pp. 80–96). Routledge.

Kardash, C. M., & Amlund, J. T. (1991). Self-reported learning strategies and learning from expository text. *Contemporary Educational Psychology*, *16*(2), 117–138. https://doi.org/10.1016/0361-476X(91)90032-G

Kennish, P., & Cavanagh, R. F. (2011). The engagement in classroom learning of year 10 and 11 Western Australian students. In Robert F. Cavanagh & R. F. Waugh

(Eds.), *Applications of Rasch Measurement in Learning Environments Research* (pp. 281–300). SensePublishers. https://doi.org/10.1007/978-94-6091-493-5_13

Khatri, P., & Gupta, P. (2019). Development and validation of employee wellbeing scale – a formative measurement model. *International Journal of Workplace Health Management*, *12*(5), 352–368. https://doi.org/10.1108/IJWHM-12-2018-0161

Kong, Q.-P., Wong, N.-Y., & Lam, C.-C. (2003). Student engagement in mathematics: Development of instrument and validation of construct. *Mathematics Education Research Journal*, *15*(1), 4–21. https://doi.org/10.1007/BF03217366

Kosko, K. W. (2014). What students say about their mathematical thinking when they listen. *School Science & Mathematics*, *114*(5), 214–223. https://doi.org/10.1111/ssm.12070

Kyle, G., & Jun, J. (2015). An alternate conceptualization of the leisure constraints measurement model. *Journal of Leisure Research*, *47*(3), 337–357. https://doi.org/10.1080/00222216.2015.11950364

Lam, S., Jimerson, S., Wong, B. P. H., Kikas, E., Shin, H., Veiga, F. H., Hatzichristou, C., Polychroni, F., Cefai, C., Negovan, V., Stanculescu, E., Yang, H., Liu, Y., Basnett, J., Duck, R., Farrell, P., Nelson, B., & Zollneritsch, J. (2014). Understanding and measuring student engagement in school: The results of an international study from 12 countries. *School Psychology Quarterly*, *29*(2), 213. https://doi.org/10.1037/spq0000057

Lam, S., Wong, B. P., Yang, H., & Liu, Y. (2012). Understanding student engagement with a contextual model. In *Handbook of research on student engagement* (pp.

403–419). Springer. http://link.springer.com/chapter/10.1007/978-1-4614-2018-7_19

Larson, R., & Csikszentmihalyi, M. (2014). The experience sampling method. In M. Csikszentmihalyi (Ed.), *Flow and the foundations of positive psychology: The collected works of Mihaly Csikszentmihalyi* (pp. 21–34). Springer Netherlands. https://doi.org/10.1007/978-94-017-9088-8_2

Lau, S., & Roeser, R. W. (2002). Cognitive abilities and motivational processes in high school students' situational engagement and achievement in science. *Educational Assessment*, *8*(2), 139. https://doi.org/10.1207/S15326977EA0802_04

Lee, J.-S. (2014). The relationship between student engagement and academic performance: Is it a myth or reality? *The Journal of Educational Research*, *107*(3), 177–185. https://doi.org/10.1080/00220671.2013.807491

Lee, N., & Cadogan, J. W. (2013). Problems with formative and higher-order reflective variables. *Journal of Business Research*, *66*(2), 242–247. https://doi.org/10.1016/j.jbusres.2012.08.004

Lee, N., & Chamberlain, L. (2016). Pride and prejudice and causal indicators. *Measurement: Interdisciplinary Research and Perspectives*, *14*(3), 105–109. https://doi.org/10.1080/15366367.2016.1227681

Lerdpornkulrat, T., Koul, R., & Poondej, C. (2018). Relationship between perceptions of classroom climate and institutional goal structures and student motivation, engagement and intention to persist in college. *Journal of Further and Higher Education*, *42*(1), 102–115. https://doi.org/10.1080/0309877X.2016.1206855

Li, Y., & Lerner, R. M. (2013). Interrelations of behavioral, emotional, and cognitive school engagement in high school students. *Journal of Youth and Adolescence, 42*(1), 20–32. http://dx.doi.org/10.1007/s10964-012-9857-5

Linnenbrink-Garcia, L., Durik, A. M., Conley, A. M., Barron, K. E., Tauer, J. M., Karabenick, S. A., & Harackiewicz, J. M. (2010). Measuring situational interest in academic domains. *Educational and Psychological Measurement*, *70*(4), 647–671. https://doi.org/10.1177/0013164409355699

Linnenbrink-Garcia, L., Rogat, T. K., & Koskey, K. L. K. (2011). Affect and engagement during small group instruction. *Contemporary Educational Psychology*, *36*(1), 13–24. https://doi.org/10.1016/j.cedpsych.2010.09.001

Little, R. J. A. (1988). A test of missing completely at random for multivariate data with missing values. *Journal of the American Statistical Association*, *83*(404), 1198–1202. JSTOR. https://doi.org/10.2307/2290157

Liu, R.-D., Zhen, R., Ding, Y., Liu, Y., Wang, J., Jiang, R., & Xu, L. (2018). Teacher support and math engagement: Roles of academic self-efficacy and positive emotions. *Educational Psychology*, *38*(1), 3–16. https://doi.org/10.1080/01443410.2017.1359238

Lobato, J. (2012). The actor-oriented transfer perspective and its contributions to educational research and practice. *Educational Psychologist*, *47*(3), 232–247. https://doi.org/10.1080/00461520.2012.693353

Loevinger, J. (1954). The attenuation paradox in test theory. *Psychological Bulletin*, *51*(5), 493–504. https://doi.org/10.1037/h0058543

Marks, H. M. (2000). Student engagement in instructional activity: Patterns in the elementary, middle, and high school years. *American Educational Research Journal*, *37*(1), 153–184. https://doi.org/10.2307/1163475

Markus, K. A., & Borsboom, D. (2013). *Frontiers of test validity theory: Measurement, causation, and meaning*. Routledge.

Maroco, J., Maroco, A. L., Campos, J. A. D. B., & Fredricks, J. A. (2016). University student's engagement: Development of the University Student Engagement Inventory (USEI). *ResearchGate*, *29*(1). https://doi.org/10.1186/s41155-016-0042-8

Marshall, S. P. (1995). *Schemas in problem solving*. Cambridge University Press.

Mayer, R. E. (1982). The psychology of mathematical problem solving. In F. K. Lester & J. Garofalo (Eds.), *Mathematical problem solving: Issues in research* (pp. 1–13). The Franklin Institute Press.

Mayer, R. E. (1983). *Thinking, problem solving, cognition*. Freeman.

Mayer, R. E. (2012). Information processing. In K. R. Harris, S. Graham, & T. Urdan (Eds.), *APA educational psychology handbook: Vol. 1.Theories, constructs, and critical issues* (pp. 85–99). American Psychological Association.

Mazer, J. P. (2012). Development and validation of the student interest and engagement scales. *Communication Methods and Measures*, *6*(2), 99–125. https://doi.org/10.1080/19312458.2012.679244

McCrone, S. S. (2005). The development of mathematical discussions: An investigation in a fifth-grade classroom. *Mathematical Thinking and Learning*, *7*(2), 111–133. https://doi.org/10.1207/s15327833mtl0702_2

McCroskey, J. C. (1970). Measures of communication-bound anxiety. *Speech Monographs*, *37*(4), 269–277. https://doi.org/10.1080/03637757009375677

McDonald, R. P., & Ho, M.-H. R. (2002). Principles and practice in reporting structural equation analyses. *Psychological Methods*, *7*(1), 64–82. https://doi.org/10.1037/1082-989X.7.1.64

McIntosh, C. N. (2007). Rethinking fit assessment in structural equation modelling: A commentary and elaboration on Barrett (2007). *Personality and Individual Differences*, *42*(5), 859–867. https://doi.org/10.1016/j.paid.2006.09.020

McNeish, D., & Wolf, M. G. (2020). Thinking twice about sum scores. *Behavior Research Methods*. https://doi.org/10.3758/s13428-020-01398-0

Meade, A. W., & Craig, S. B. (2012). Identifying careless responses in survey data. *Psychological Methods*, *17*(3), 437. https://doi.org/10.1037/a0028085

Mehrabian, A. (1996). Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in temperament. *Current Psychology*, *14*(4), 261–292. https://doi.org/10.1007/BF02686918

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, *50*, 741–749. https://doi.org/10.1037/0003-066X.50.9.741

Meyers, L. S., Gamst, G., & Guarino, A. J. (2012). *Applied Multivariate Research: Design and Interpretation*. SAGE.

Miller, R. B., Greene, B. A., Montalvo, G. P., Ravindran, B., & Nichols, J. D. (1996). Engagement in academic work: The role of learning goals, future consequences, pleasing others, and perceived ability. *Contemporary Educational Psychology*, *21*(4), 388–422. https://doi.org/10.1006/ceps.1996.0028

Miserandino, M. (1996). Children who do well in school: Individual differences in perceived competence and autonomy in above-average children. *Journal of Educational Psychology*, *88*(2), 203–214. https://doi.org/10.1037/0022-0663.88.2.203

Nokes-Malach, T. J., & Richey, J. E. (2015). Knowledge transfer: New approaches to a controversial phenomenon. In R. Scott & S. Kosslyn (Eds.), *Emerging trends in the social and behavioral sciences* (pp. 1–15). John Wiley and Sons.

Nokes-Malach, Timothy J., & Mestre, J. P. (2013). Toward a model of transfer as sense-making. *Educational Psychologist*, *48*(3), 184–207. https://doi.org/10.1080/00461520.2013.807556

Nunnally, J. C. (1967). *Psychometric theory*. McGraw-Hill.

Nunnally, J. C. (1978). *Psychometric theory*. McGraw-Hill.

Office of Institutional Research and Effectiveness. (n.d.). Retrieved January 5, 2018, from https://irr2.gmu.edu/New/N_EnrollOff/EnrlStsDemo.cfm

Otten, S., Herbel-Eisenmann, B., Steele, M., Cirillo, M., & Bosman, H. (2011). *Students actively listening: A foundation for productive discourse in mathematics* *classrooms*. Annual Meeting of the American Educational Research Association, New Orleans, LA.

Patrick, H., Ryan, A. M., & Kaplan, A. (2007a). Early adolescents' perceptions of the classroom social environment, motivational beliefs, and engagement. *Journal of Educational Psychology*, *99*(1), 83–98. https://doi.org/10.1037/0022-0663.99.1.83

Patrick, H., Ryan, A. M., & Kaplan, A. (2007b). Early adolescents' perceptions of the classroom social environment, motivational beliefs, and engagement. *Journal of Educational Psychology*, *99*(1), 83.

Pekrun, R. (2006). The control-value theory of achievement emotions: Assumptions, corollaries, and implications for educational research and practice. *Educational Psychology Review*, *18*(4), 315–341. https://doi.org/10.1007/s10648-006-9029-9

Pekrun, R. (2016). Academic emotions. In K. R. Wentzel & D. B. Miele (Eds.), *Handbook of motivation at school* (2nd ed., pp. 120–144). Routledge.

Pekrun, R., Goetz, T., Frenzel, A. C., Barchfeld, P., & Perry, R. P. (2011). Measuring emotions in students' learning and performance: The Achievement Emotions Questionnaire (AEQ). *Contemporary Educational Psychology*, *36*(1), 36–48. https://doi.org/10.1016/j.cedpsych.2010.10.002

Pekrun, R., & Linnenbrink-Garcia, L. (2012). Academic emotions and student engagement. In *Handbook of research on student engagement* (pp. 259–282). Springer, Boston, MA. https://doi.org/10.1007/978-1-4614-2018-7_12

Pekrun, R., & Stephens, E. J. (2012). Academic emotions. In K. R. Harris, S. Graham, & T. Urdan (Eds.), *APA educational psychology handbook: Vol. 2. Individual* *differences and cultural and contextual factors* (pp. 3–31). American Psychological Association.

Pekrun, R., Vogl, E., Muis, K. R., & Sinatra, G. M. (2017). Measuring emotions during epistemic activities: The epistemically-related emotion scales. *Cognition and Emotion*, *31*(6), 1268–1276. https://doi.org/10.1080/02699931.2016.1204989

Piaget, J. (1952). *The origins of intelligence in children*. W W Norton & Co.

Pintrich, P. R., & de Groot, E. V. (1990). Motivational and self-regulated learning components of classroom academic performance. *Journal of Educational Psychology*, *82*(1), 33–40. https://doi.org/10.1037/0022-0663.82.1.33

Pintrich, P. R., Smith, D. A. F., Garcia, T., & McKeachie, W. J. (1991). *A manual for the use of the Motivated Strategies for Learning Questionnaire (MSLQ)*.

Pintrich, P. R., Smith, D. A. F., Garcia, T., & Mckeachie, W. J. (1993). Reliability and predictive validity of the Motivated Strategies for Learning Questionnaire (MSLQ). *Educational and Psychological Measurement*, *53*(3), 801–813. https://doi.org/10.1177/0013164493053003024

Pokay, P., & Blumenfeld, P. C. (1990). Predicting achievement early and late in the semester: The role of motivation and use of learning strategies. *Journal of Educational Psychology*, *82*, 41–50. https://doi.org/10.1037/0022-0663.82.1.41

Polya, G. (1957). *How to solve it*. Princeton University Press.

Powers, D. E. (1986). Academic demands related to listening skills. *Language Testing*, *3*(1), 1–38. https://doi.org/10.1177/026553228600300101

President's Council of Advisors on Science and Technology. (2012). *Engage to excel: Producing one million additional college graduates with degrees in Science, Technology, Engineering, and Mathematics*. Executive Office of the President. http://eric.ed.gov/?id=ED541511

Qureshi, A., Wall, H., Humphries, J., & Bahrami Balani, A. (2016). Can personality traits modulate student engagement with learning and their attitude to employability? *Learning and Individual Differences*, *51*, 349–358. https://doi.org/10.1016/j.lindif.2016.08.026

Ramirez, E. M. (2016). *Development and validation of an instrument to measure secondary teachers' self-efficacy in reading instruction (STERI) across the content areas* [Ph.D., George Mason University]. http://search.proquest.com.mutex.gmu.edu/pqdtlocal1006610/docview/1800304341/abstract/F6E7F3EEA15B4591PQ/1

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods*. SAGE.

Reeve, J. (2013). How students create motivationally supportive learning environments for themselves: The concept of agentic engagement. *Journal of Educational Psychology*, *105*(3), 579–595. https://doi.org/10.1037/a0032690

Reeve, J., & Tseng, C.-M. (2011). Agency as a fourth aspect of students' engagement during learning activities. *Contemporary Educational Psychology*, *36*(4), 257–267. https://doi.org/10.1016/j.cedpsych.2011.05.002

Reschly, A. L., & Christenson, S. L. (2006). Prediction of dropout among students with mild disabilities: A case for the inclusion of student engagement variables. *Remedial & Special Education*, *27*(5), 276–292.

Reschly, A. L., & Christenson, S. L. (2012). Jingle, jangle, and conceptual haziness: Evolution and future directions of the engagement construct. In S. L. Christenson, A. L. Reschly, & C. Wylie (Eds.), *Handbook of research on student engagement* (pp. 3–19). Springer US. https://doi.org/10.1007/978-1-4614-2018-7_1

Rhemtulla, M., Bork, R. van, & Borsboom, D. (2015). Calling models with causal indicators “measurement models” implies more than they can deliver. *Measurement: Interdisciplinary Research and Perspectives*, *13*(1), 59–62. https://doi.org/10.1080/15366367.2015.1016343

Rhemtulla, M., van Bork, R., & Borsboom, D. (2020). Worse than measurement error: Consequences of inappropriate latent variable measurement models. *Psychological Methods*, *25*(1), 30–45. https://doi.org/10.1037/met0000220

Richards, J. C. (1983). Listening comprehension: Approach, design, procedure. *TESOL Quarterly*, *17*(2), 219–240. https://doi.org/10.2307/3586651

Richmond, V. P., Gorham, J. S., & McCroskey, J. C. (1987). The relationship between selected immediacy behaviors and cognitive learning. *Annals of the International Communication Association*, *10*(1), 574–590.

Rimm-Kaufman, S. E., Baroody, A. E., Larsen, R. A. A., Curby, T. W., & Abry, T. (2015). To what extent do teacher–student interaction quality and student gender contribute to fifth graders’ engagement in mathematics learning? *Journal of Educational Psychology*, *107*(1), 170–185. https://doi.org/10.1037/a0037252

Rittle-Johnson, B., Loehr, A. M., & Durkin, K. (2017). Promoting self-explanation to improve mathematics learning: A meta-analysis and instructional design principles. *ZDM*, *49*(4), 599–611. https://doi.org/10.1007/s11858-017-0834-z

Rodrigues, M., Menezes, I., & Ferreira, P. D. (2018). Validating the formative nature of psychological empowerment construct: Testing cognitive, emotional, behavioral, and relational empowerment components. *Journal of Community Psychology*, *46*(1), 58–78. https://doi.org/10.1002/jcop.21916

Rosenberg, E. L. (1998). Levels of analysis and the organization of affect. *Review of General Psychology*, *2*(3), 247–270. https://doi.org/10.1037/1089-2680.2.3.247

Rosenzweig, C., Krawec, J., & Montague, M. (2011). Metacognitive strategy use of eighth-grade students with and without learning disabilities during mathematical problem solving: A think-aloud analysis. *Journal of Learning Disabilities*, *44*(6), 508–520. https://doi.org/10.1177/0022219410378445

Rovai, A. P., & Barnum, K. T. (2003). On-line course effectiveness: An analysis of student interactions and perceptions of learning. *Journal of Distance Education*, *18*(1), 57–73.

Russell, J. A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology*, *39*(6), 1161–1178. https://doi.org/10.1037/h0077714

Salmela-Aro, K., Moeller, J., Schneider, B., Spicer, J., & Lavonen, J. (2016). Integrating the light and dark sides of student engagement using person-oriented and situation-specific approaches. *Learning and Instruction*, *43*, 61–70. https://doi.org/10.1016/j.learninstruc.2016.01.001

Salmela-Aro, K., & Upadaya, K. (2012). The Schoolwork Engagement Inventory: Energy, dedication, and absorption (EDA). *European Journal of Psychological Assessment*, *28*(1), 60–67. https://doi.org/10.1027/1015-5759/a000091

Salmela-Aro, K., & Upadyaya, K. (2014). School burnout and engagement in the context of demands-resources model. *British Journal of Educational Psychology*, *84*(1), 137–151. https://doi.org/10.1111/bjep.12018

Sass, D. A., & Schmitt, T. A. (2010). A comparative investigation of rotation criteria within exploratory factor analysis. *Multivariate Behavioral Research*, *45*(1), 73–103. https://doi.org/10.1080/00273170903504810

Schaufeli, W. B., Bakker, A. B., & Salanova, M. (2006). The measurement of work engagement with a short questionnaire: A cross-national study. *Educational and Psychological Measurement*, *66*(4), 701–716. https://doi.org/10.1177/0013164405282471

Schaufeli, W. B., Salanova, M., González-romá, V., & Bakker, A. B. (2002). The measurement of engagement and burnout: A two sample confirmatory factor analytic approach. *Journal of Happiness Studies*, *3*(1), 71–92. https://doi.org/10.1023/A:1015630930326

Schmeck, R. R., Ribich, F., & Ramanaiah, N. (1977). Development of a self-report inventory for assessing individual differences in learning processes. *Applied*

Psychological Measurement, *1*(3), 413–431. https://doi.org/10.1177/014662167700100310

Sciarra, D. T., & Seirup, H. J. (2008). The multidimensionality of school engagement and math achievement among racial groups. *Professional School Counseling*, *11*(4), 218–228.

Shachar, H., & Sharan, S. (1994). Talking, relating, and achieving: Effects of cooperative learning and whole-class instruction. *Cognition and Instruction*, *12*(4), 313–353.

Shaw, J. D., Duffy, M. K., & Stark, E. M. (2000). Interdependence and preference for group work: Main and congruence effects on the satisfaction and performance of group members. *Journal of Management*, *26*(2), 259–279. https://doi.org/10.1177/014920630002600205

Shernoff, D. J. (2013). *Optimal learning environments to promote student engagement*. Springer. http://dx.doi.org/10.1007/978-1-4614-7089-2

Shernoff, D. J., Csikszentmihalyi, M., Schneider, B., & Shernoff, E. S. (2003). Student engagement in high school classrooms from the perspective of flow theory. *School Psychology Quarterly*, *18*(2), 158.

Shernoff, D. J., Ruzek, E. A., & Sinha, S. (2016). The influence of the high school classroom environment on learning as mediated by student engagement. *School Psychology International*, 0143034316666413. https://doi.org/10.1177/0143034316666413

Signature learning spaces. (2015). *Learning Environments*. http://learningenvironments.gmu.edu/signature-learning-spaces/

Sinatra, G. M., Heddy, B. C., & Lombardi, D. (2015). The challenges of defining and measuring student engagement in science. *Educational Psychologist*, *50*(1), 1–13. https://doi.org/10.1080/00461520.2014.1002924

Sinclair, M. F., Christenson, S. L., Lehr, C. A., & Anderson, A. R. (2003). Facilitating student engagement: Lessons learned from Check & Connect longitudinal studies. *The California School Psychologist, 8*(1), 29–41. https://doi.org/10.1007/BF03340894

Skemp, R. R. (1987). *The psychology of learning mathematics*. L. Erlbaum Associates.

Skinner, E. A., Furrer, C., Marchand, G., & Kindermann, T. (2008). Engagement and disaffection in the classroom: Part of a larger motivational dynamic? *Journal of Educational Psychology, 100*(4), 765–781. https://doi.org/10.1037/a0012840

Skinner, E. A., Kindermann, T. A., & Furrer, C. J. (2009). A motivational perspective on engagement and disaffection conceptualization and assessment of children's behavioral and emotional participation in academic activities in the classroom. *Educational and Psychological Measurement*, *69*(3), 493–525. https://doi.org/10.1177/0013164408323233

Skinner, E. A., & Pitzer, J. R. (2012). Developmental dynamics of student engagement, coping, and everyday resilience. In S. L. Christenson, A. L. Reschly, & C. Wylie (Eds.), *Handbook of research on student engagement* (pp. 21–44). Springer US. https://doi.org/10.1007/978-1-4614-2018-7_2

Skinner, E. A., Saxton, E., Currie, C., & Shusterman, G. (2017). A motivational account of the undergraduate experience in science: Brief measures of students' self-system appraisals, engagement in coursework, and identity as a scientist. *International Journal of Science Education, 39*(17), 2433–2459. https://doi.org/10.1080/09500693.2017.1387946

Stearns Center for Teaching and Learning. (n.d.). Stearns Center for Teaching and Learning. Retrieved January 5, 2018, from https://stearnscenter.gmu.edu/

Streiner, D. L. (2003). Starting at the beginning: An introduction to coefficient alpha and internal consistency. *Journal of Personality Assessment*, *80*(1), 99–103.

Sulea, C., van Beek, I., Sarbescu, P., Virga, D., & Schaufeli, W. B. (2015). Engagement, boredom, and burnout among students: Basic need satisfaction matters more than personality traits. *Learning and Individual Differences*, *42*, 132–138. https://doi.org/10.1016/j.lindif.2015.08.018

Sweller, J., & Cooper, G. A. (1985). The use of worked examples as a substitute for problem solving in learning algebra. *Cognition and Instruction*, *2*(1), 59–89.

Tabachnick, B. G., & Fidell, L. S. (2007). *Using multivariate statistics, 5th ed*. Allyn & Bacon/Pearson Education.

Thayer, R. E. (1986). Activation-deactivation adjective check list: Current overview and structural analysis. *Psychological Reports*, *58*(2), 607–614. https://doi.org/10.2466/pr0.1986.58.2.607

Thien, L. M. (2019). Assessing a second-order quality of school life construct using partial least squares structural equation modelling approach. *International Journal of Research & Method in Education*. https://www.tandfonline.com/doi/abs/10.1080/1743727X.2019.1662779

Thien, L. M., & Razak, N. A. (2013). Academic coping, friendship quality, and student engagement associated with student quality of school life: A partial least square analysis. *Social Indicators Research*, *112*(3), 679–708. https://doi.org/10.1007/s11205-012-0077-x

Thompson, K., Leintz, P., Nevers, B., & Witkowski, S. (2004). The integrative listening model: An approach to teaching and learning listening. *The Journal of General Education*, *53*(3/4), 225–246.

Thorndyke, P. W. (1984). Applications of schema theory in cognitive research. In J. R. Anderson & S. M. Kosslyn (Eds.), *Tutorials in learning and memory: Essays in honor of Gordon Bower* (pp. 167–191). Freeman & Company.

Tomkins, S. S., & McCarter, R. (1964). What and where are the primary affects? Some evidence for a theory. *Perceptual and Motor Skills*, *18*, 119–158. https://doi.org/10.2466/pms.1964.18.1.119

Tourangeau, Roger. (2000). *The psychology of survey response*. Cambridge University Press.

Uekawa, K., Borman, K., & Lee, R. (2007). Student engagement in U.S. urban high school mathematics and science classrooms: Findings on social organization, race, and ethnicity. *The Urban Review, 39,* 1–43. https://doi.org/10.1007/s11256-006-0039-1

Uzzaman, M. A., & Karim, A. K. M. R. (2016). The psychometric properties of school engagement scale in Bangladeshi culture. *Journal of the Indian Academy of Applied Psychology*, *42*(1), 143–153.

Valentine, J., & Painter, B. A. (2007). *Instructional Practices Inventory*. Columbia: University of Missouri, Center for School Improvement. https://mospace.umsystem.edu/xmlui/handle/10355/3566

Van Meter, P., Yokoi, L., & Pressley, M. (1994). College students' theory of note-taking derived from their perceptions of note-taking. *Journal of Educational Psychology*, *86*(3), 323–338. https://doi.org/10.1037/0022-0663.86.3.323

Vermunt, J. D. (1996). Metacognitive, cognitive and affective aspects of learning styles and strategies: A phenomenographic analysis. *Higher Education*, *31*(1), 25–50.

Vermunt, J. D. (1998). The regulation of constructive learning processes. *British Journal of Educational Psychology*, *68*(2), 149–171.

Voelkl, K. E. (1995). School warmth, student participation, and achievement. *The Journal of Experimental Education*, *63*(2), 127–138.

Walter, E. M., Henderson, C. R., Beach, A. L., & Williams, C. T. (2016). Introducing the postsecondary instructional practices survey (PIPS): A concise, interdisciplinary, and easy-to-score survey. *CBE Life Sciences Education*, *15*(4). https://doi.org/10.1187/cbe.15-09-0193

Wang, M.-T. (2010). *School climate support for student engagement during adolescence* [Ed.D., Harvard University]. https://search.proquest.com/pqdtglobal/docview/859272647/abstract/C26926056A754EFEPQ/1

Wang, M.-T., Fredricks, J. A., Ye, F., Hofkens, T. L., & Linn, J. S. (2016). The math and science engagement scales: Scale development, validation, and psychometric properties. *Learning and Instruction*, *43*, 16–26. https://doi.org/10.1016/j.learninstruc.2016.01.008

Wang, M.-T., & Holcombe, R. (2010). Adolescents’ perceptions of school environment, engagement, and academic achievement in middle school. *American Educational Research Journal*, *47*(3), 633–662. https://doi.org/10.3102/0002831209361209

Wang, M.-T., Willett, J. B., & Eccles, J. S. (2011). The assessment of school engagement: Examining dimensionality and measurement invariance by gender and race/ethnicity. *Journal of School Psychology*, *49*(4), 465–480. https://doi.org/10.1016/j.jsp.2011.04.001

Wang, Z., Bergin, C., & Bergin, D. A. (2014). Measuring engagement in fourth to twelfth grade classrooms: The Classroom Engagement Inventory. *School Psychology Quarterly: The Official Journal of the Division of School Psychology, American Psychological Association*, *29*(4), 517–535. https://doi.org/10.1037/spq0000050

Watson, D., & Tellegen, A. (1985). Toward a consensual structure of mood. *Psychological Bulletin*, *98*(2), 219–235. https://doi.org/10.1037/0033-2909.98.2.219

Watt, T., Rasmussen, A. K., Groenvold, M., Bjorner, J. B., Watt, S. H., Bonnema, S. J., Hegedüs, L., & Feldt-Rasmussen, U. (2008). Improving a newly developed patient-reported outcome for thyroid patients, using cognitive interviewing. *Quality of Life Research: An International Journal of Quality of Life Aspects of Treatment, Care and Rehabilitation*, *17*, 1009–1017. https://doi.org/10.1007/s11136-008-9364-z

Webb, N. M. (1980). Group process and learning in an interacting group. *The Quarterly Newsletter of the Institute for Comparative Human Cognition*, *2*(1), 10–15.

Webb, N. M. (1982). Group composition, group interaction, and achievement in cooperative small groups. *Journal of Educational Psychology*, *74*(4), 475–484. https://doi.org/10.1037/0022-0663.74.4.475

Webb, N. M. (1984). Sex differences in interaction and achievement in cooperative small groups. *Journal of Educational Psychology*, *76*(1), 33–44. https://doi.org/10.1037/0022-0663.76.1.33

Webb, N. M., Franke, M. L., Ing, M., Chan, A., De, T., Freund, D., & Battey, D. (2008). The role of teacher instructional practices in student collaboration. *Contemporary Educational Psychology*, *33*(3), 360–381. https://doi.org/10.1016/j.cedpsych.2008.05.003

Weinberg, A., Wiesner, E., & Fukawa-Connelly, T. (2014). Students' sense-making frames in mathematics lectures. *The Journal of Mathematical Behavior*, *33*, 168–179. https://doi.org/10.1016/j.jmathb.2013.11.005

Weinstein, C. E., Palmer, D. R., & Acee, T. W. (2016). *LASSI (Learning and Study Strategies Inventory): User's Manual* (3rd ed.). H&H Publishing Company, Inc.

Whitney, B. M., Cheng, Y., Brodersen, A. S., & Hong, M. R. (2019). The scale of student engagement in Statistics: Development and initial validation. *Journal of Psychoeducational Assessment*, *37*(5), 553–565. https://doi.org/10.1177/0734282918769983

Willis, G. B. (2015). *Analysis of the cognitive interview in questionnaire design*. Oxford University Press.

Willoughby, M. T., & Blair, C. B. (2016). Measuring executive function in early childhood: A case for formative measurement. *Psychological Assessment*, *28*(3), 319–330. https://doi.org/10.1037/pas0000152

Wilson, D., Jones, D., Bocell, F., Crawford, J., Kim, M., Veilleux, N., Floyd-Smith, T., Bates, R., & Plett, M. (2015). Belonging and academic engagement among undergraduate STEM students: A multi-institutional study. *Research in Higher Education*, *56*(7), 750–776. https://doi.org/10.1007/s11162-015-9367-x

Wilson, D., Jones, D., Kim, M. J., Allendoerfer, C., Bates, R., Crawford, J., Floyd-Smith, T., Plett, M., & Veilleux, N. (2014). The link between cocurricular activities and academic engagement in engineering education. *Journal of Engineering Education*, *103*(4), 625–651. https://doi.org/10.1002/jee.20057

Wolfe, E. W., & Smith, E. V. (2007). Instrument development tools and activities for measure validation using Rasch models: Part I - instrument development tools. *Journal of Applied Measurement*, *8*(1), 97–123.

Wolters, C. A. (2004). Advancing achievement goal theory: Using goal structures and goal orientations to predict students' motivation, cognition, and achievement. *Journal of Educational Psychology*, *96*(2), 236–250. https://doi.org/10.1037/0022-0663.96.2.236

Wolvin, A. D. (2010). Listening engagement: Intersecting theoretical perspectives. In A. D. Wolvinessor (Ed.), *Listening and human communication in the 21st century* (pp. 7–30). Wiley-Blackwell. https://doi.org/10.1002/9781444314908.ch1

Wong, R. M. F., Lawson, M. J., & Keeves, J. (2002). The effects of self-explanation training on students' problem solving in high-school mathematics. *Learning and Instruction*, *12*(2), 233–262. https://doi.org/10.1016/S0959-4752(01)00027-5

Xu, L., Diket, R., & Brewer, T. (2018). Predicting student performance via NAEP secondary art analysis using partial least squares SEM. *Arts Education Policy Review*, *119*(4), 231–242. https://doi.org/10.1080/10632913.2017.1327382

Yair, G. (2000). Educational battlefields in America: The tug-of-war over students' engagement with instruction. *Sociology of Education*, *73*, 247–269. https://doi.org/10.2307/2673233

Biography

Daria Gerasimova grew up in St. Petersburg, Russia. She graduated with a B.S. and an M.S. in Engineering and Technology from St. Petersburg State Polytechnical University in 2010 and 2012, respectively. From 2012-2014, she was a Mathematics teacher (Grades 5-9) at School #136 in St. Petersburg. In 2014, she moved to the U.S. to pursue her Ph.D. in Education. She is interested in psychological aspects of mathematics education and in applied research methods and measurement.