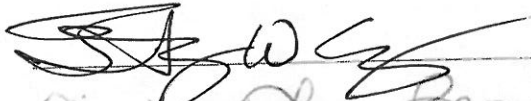

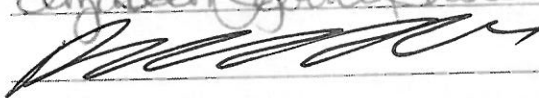


USING AUTOREGRESSIVE MOVING AVERAGE MODEL TO INVESTIGATE THE  
STABILITY OF DIMENSIONS OF EMOTIONAL SUPPORT

by

Noora Hamdan  
A Thesis  
Submitted to the  
Graduate Faculty  
of  
George Mason University  
in Partial Fulfillment of  
The Requirements for the Degree  
of  
Master of Arts  
Psychology

Committee:

Director

Jim Thompson  
James T. Cooper

Department Chairperson

Dean, College of Humanities  
and Social Sciences

Date: July 2, 2013

Summer Semester 2013  
George Mason University  
Fairfax, VA

Using Autoregressive Moving Average Model to Investigate the Stability of Dimensions  
of Emotional Support

A thesis submitted in partial fulfillment of the requirements for the degree of Master of  
Arts at George Mason University

By

Noora Hamdan  
Bachelor of Arts  
James Madison University, 2011

Director: Timothy W. Curby, Assistant Professor  
Department of Psychology

Summer Semester 2013  
George Mason University  
Fairfax, VA

## Table of Contents

List of Tables .....	v
List of Figures .....	vi
Abstract .....	vii
1. Introduction .....	1
2. Emotionally Supportive Teacher-child Interactions .....	5
3. Importance of Emotional Support Stability .....	12
4. Problems Associated with Observer Ratings .....	16
5. Quantification of Potential Rater Effects .....	20
6. Research Questions/Hypotheses .....	22
7. Method .....	24
8. Data Analysis .....	27
9. Results .....	31
Emotional Support Domain .....	34
Positive Climate .....	38
Negative Climate .....	40
Teacher Sensitivity .....	43
Regard for Student Perspective .....	45
10. Discussion .....	48
11. Implications .....	53
12. Limitations .....	56
13. Future Directions .....	58
List of References .....	59

## List of Tables

Table	Page
1. ARMA Stability Estimates of Dimensions of Emotional Support and Overall Emotional Support, With Associated Average Correlation of Errors .....	47

## List of Figures

Figure	Page
1. Hypothesized Emotional Support Stability.....	31
2. Hypothesized Autocorrelation Function of Residuals .....	33
3. Overall Emotional Support Score Over Time by Minute .....	35
4. Autocorrelation Function of Residuals of Emotional Support .....	36
5. Partial Autocorrelation Function of Residuals of Emotional Support .....	37
6. Positive Climate Score Over Time by Minute .....	38
7. Autocorrelation Function of Residuals of Positive Climate .....	39
8. Partial Autocorrelation Function of Residuals of Positive Climate.....	40
9. Negative Climate Score Over Time by Minute .....	41
10. Autocorrelation Function of Residuals of Negative Climate.....	42
11. Partial Autocorrelation Function of Residuals of Negative Climate .....	42
12. Teacher Sensitivity Score Over Time by Minute .....	43
13. Autocorrelation Function of Residuals of Teacher Sensitivity.....	44
14. Partial Autocorrelation Function of Residuals of Teacher Sensitivity .....	44
15. Regard for Student Perspective Score by Minute Over Time.....	45
16. Autocorrelation Function of Residuals of Regard for Student Perspective.....	46
17. Partial Autocorrelation Function of Residuals of Regard for Student Perspective .....	46

## **Abstract**

### **USING AUTOREGRESSIVE MOVING AVERAGE MODEL TO INVESTIGATE THE STABILITY OF DIMENSIONS OF EMOTIONAL SUPPORT**

Noora Hamdan, M.A.

George Mason University, 2013

Thesis Director: Dr. Timothy W. Curby

This study investigated the minute-to-minute stability of dimensions of teacher Emotional Support in pre-kindergarten classrooms. Developmental theory states that the proximal processes, in this case, the teachers' Emotional Supportive interactions, occur in the moment-to-moment interactions between a person and their environment and drive development. The present study examined the stability of those interactions, on the time scale in which they take place, to appropriately align our analytical methods with our theory of development. That is, how stable are minute-to-minute classroom Emotional Supports over time? Furthermore, when conducting observations of a lesson, raters may increasingly determine that a score on a dimension for a teacher fits as opposed to actually making independent ratings at each time point. In this way, ratings of teachers' Emotional Support may become more stable over time as raters increasingly make up their minds. Then the question is not only how stable are the ratings but also, to what

degree are ratings stabilizing over a lesson? In combination, we can better understand the experiences of children during a lesson.

Participants were randomly selected from publically-funded pre-kindergarten programs where the majority of children were eligible to enter kindergarten the following school year. Data were coded from videotapes of teachers using an adaptation of the Classroom Assessment Scoring System (CLASS) (Pianta, La Paro, & Hamre, 2008), whereby dimensions of Emotional Support (*positive climate, negative climate, teacher sensitivity, and regard for student perspectives*) were coded once every minute during a language arts lesson. To address the research questions, autoregressive moving average (ARMA) models were fit to each dimension of Emotional Support. ARMA modeling is based on the notion that repeated measurements are correlated across time and may be expressed as an autocorrelation function. The autoregressive portion of the model answers the question of how stable Emotional Support dimensions are. The correlations between error variances were examined to determine if the correlations are negative which suggests that raters were increasingly making up their minds about a given rating. Findings showed that overall Emotional Support and all individual Emotional Support dimensions showed low levels of stability. Results showed that raters are becoming more consistent in their ratings of Average Emotional Support and its dimensions over time.

These findings allow for a better understanding of rater effects in classroom observations. With the recent emphasis by federal agencies on the use of observation, understanding the extent of rater effects has large policy implications.

## **1. Introduction**

### **Investigation of the Stability of Teacher Emotional Support Using Autoregressive Moving Average Models**

Recent large studies have revealed that children from pre-k through elementary in the United States experience classroom environments which are, on average, high in Emotional Support (Pianta, Belsky, Houts & Morrison, 2007; Stuhlman & Pianta, 2009).

It has also been found that those teachers who are more emotionally supportive value social-emotional learning in their classroom and incorporate this learning into their daily interactions with students (Zinsser, Shewark, Denham, & Curby, under review).

Teachers with high levels of Emotional Support offer an important context for children's early school success (Morrison & Connor, 2002; Rutter & Maughan, 2002; Thompson & Happold, 2002). There is growing evidence that indicates Emotional Support fosters children's academic outcomes, especially for at-risk children (Brock et al., 2008; Brock & Curby, 2012a Doll et al., 2004; Hamre & Pianta, 2005; O Connor & McCartney, 2007; Stipek et al., 1995). These results in regards to student outcomes have been confirmed by the NICHD SEECYD in terms of being connected to student's learning engagement and literacy skills (NICHD ECCRN, 2003). This study will provide investigation of the stability of teacher Emotional Support which is observed in Pre-kindergarten classrooms. Classroom interactions, occurring between teachers and



students, are the most direct channels of influence on children's learning (Bronfenbrenner & Morris, 2006; Creemers, 1994; Nye, Konstantopoulos, & Hedges, 2004; Pianta, 1999). Further, recent observational research has shown that there are substantial differences between teachers in the quality of classroom interactions with students (Early et al., 2005; Hamre & Pianta, 2005). These differences have important implications for children, particularly with children who have higher quality emotional, organizational, and instructional supports which have shown to influence students' social, behavioral, and academic outcomes (Early et al., 2005; Hamre & Pianta, 2005).

Differences in stability and quality of teacher's interactions with students in the classroom can be seen within teachers and between teachers. Specifically, not only are there differences between teachers in terms of the quality of interactions with students, but there are also differences in teachers within a day with some aspects of instruction being more stable than others (Curby, Stuhlman, Grimm, Mashburn, Chomat-Mooney, & Downer, 2011). Somewhat relatedly, there are between-teacher differences in how stable the provision of emotional, organizational, and instructional support are in the classroom. These differences in stability, specifically in regards to Emotional Support, are important because pre-kindergarten children who experience more consistency in the Emotional Support environment over a day have been shown to have better academic and social outcomes in preschool and later in kindergarten (Curby et al, 2013). However, these examinations of stability have focused on the stability over a day in 30-minute blocks. What has yet to be assessed is the measurement of these classroom interactions as they occur on a minute-to-minute basis. Thus the first purpose of this study is to investigate

the minute-to-minute stability of teacher Emotional Support as provided to students in the classroom over time. This incremental, micro-level investigation of classroom interaction quality will allow for a novel description of the stability in teacher Emotional Supports, thus allowing for a better understanding about the dynamics of these teacher-student interactions.

The next aspect of this study involves looking at the potential influence of raters on observed Emotional Support stability. Stability may have little to do with the teacher and much to do with the raters. It could very well be that as raters make ratings minute after minute, they come to decide that a classroom is characterized at a certain level and deviations from that set point become smaller and smaller. Another way to conceptualize this question is does a rater increasingly make up their mind over the course of carrying out ratings of Emotional Support? In order to assess the degree to which raters are becoming more stable in their ratings, the average correlation of errors were calculated across adjacent time points. Therefore, the second purpose of this study is to investigate whether ratings of the dimensions of teacher Emotional Support become more stable over the course of a lesson.

The present study takes part of its significance from recent research that has shown that lack of consistency in Emotional Support (i.e., where a teacher's Emotional Support varies greatly over time) within a school day is a hindrance to child academic and social development (Curby et al., 2013). In other words, students who experience a lot of ups and downs in teachers' provision of Emotional Supports have worse outcomes

than children who experience a consistent provision of Emotional Support (Curby et al., 2013). Notably, the effect of consistency in Emotional Support was stronger than the effect of the mean level of Emotional Support which suggests that consistency of Emotional Support itself is a topic worthy of investigation. There are many aspects of consistency that remain unexplored. Older grades (e.g., third and fifth) showed that dimensions of Emotional Support were most consistent out of all aspects of classroom quality which the CLASS measures (including Classroom Organization and Instructional support) (Curby, Stuhlman et al., 2011), but is this true in pre-kindergarten? Also, the research that linked the stability of Emotional Support to pre-kindergarten children's outcomes measured the cycle-to-cycle stability of Emotional Support with each cycle consisting of a 20-minute observation and a 10-minute rating period (Curby et al., 2013). Therefore, the variability within a 20-minute observation is being lost because there is only one score for each dimension in that 20 minutes. This leaves the variability within a cycle uninvestigated because ratings were only made overall, at the completion of that 20 minute cycle and not throughout its observation. Lastly, we know very little about the role of raters in the degree of stability. It could be that over repeated ratings, raters begin to rate teachers more consistently on Emotional Support, where previous ratings influence subsequent ratings.

## **2. Emotionally Supportive Teacher- Child Interactions**

Emotionally supportive teachers are warm and kind. They are sensitive to the social and emotional needs of children in their classrooms, and are thoughtful about how they respond to children. They provide gentle guidance and take part in positive communication with students. They also show respect for children through respectful language, eye contact, and a warm and calm voice. Classroom observation measures where observers rate multiple aspects of teaching, including Emotional Support, are increasingly being implemented in both research and policy. Researchers are looking at how these measures relate to other possible measures of teacher quality like those based on student outcomes (Bill and Melinda Gates Foundation, 2012), and states and districts are more greatly introducing these measures into their accountability systems. In both cases, scores from typically only a few lessons are aggregated in order to derive measures of a teacher's quality of teaching.

How do these classroom observations breakdown the complex classroom environment like that involved in Emotional Support? One of the most popular classroom observation instruments used to get an understanding of the complexity of classroom environments and students' classroom experiences is the Classroom Assessment Scoring System (CLASS; Pianta, LaParo, & Hamre, 2008). In the CLASS, Emotional Support is made up of four measured dimensions: positive climate, negative

climate, teacher sensitivity, and regard for student perspectives. These dimensions provide an understanding of the quality of Emotional Support as provided by teachers to students in the classroom. Notably, the present study used this measure but used a different scoring protocol in that ratings were made every minute instead of approximately every 20 minutes. Ratings of each dimension at every minute were investigated and also averaged to obtain an Emotional Support score at every minute. Higher scores indicate a more emotionally supportive teacher. What follows is description of the previously mentioned CLASS dimensions of Emotional Support. All of these dimensions together give us an idea of a teacher's level of Emotional Support.

**Positive Climate.** One aspect of emotionally supportive teachers is the positive climate they create in their classroom. Teachers who offer high levels of positive climate take part in positive communication with students, and show respect for students via a warm and calm voice, eye contact, and respectful and courteous language (Pianta et al., 2008). For instance, in classrooms with high positive climate, teachers demonstrate positive affect toward students, encourage students to behave in a kind and caring manner towards one another, provide feedback in regards to behavior or school work which is encouraging, show knowledge and understanding of individual children's likes and dislikes, and create an environment where children can voice their viewpoints and ideas (Pianta et al., 2008).

**Negative Climate.** Emotionally supportive teachers are not likely to display many behaviors which are indicative of a negative climate. That is, they do not display

controlling behaviors, criticize students, employ punitive disciplinary approaches, or use sarcastic language (National Institute of Child Health and Human Development Early Child Care Research Network, NICHD-ECCRN, 2002; Pianta & Hamre, 2009). Therefore, this dimension measures the level of negativity (anger, aggression, or hostility) which exists in teacher-child classroom interactions. For this study, negative climate will be reversed coded, when aggregated into Emotional Support.

**Teacher Sensitivity.** Teachers who are sensitive are both aware of and responsive to their students' academic and emotional needs (National Institute of Child Health and Human Development Early Child Care Research Network, NICHD-ECCRN, 2002, Pianta et al., 2008, Pianta & Hamre, 2009 ). A sensitive teacher is one who is aware of students' needs and abilities, and appropriately matches her support to these varying student needs, providing extra support to those who require it .In this classroom, students demonstrate comfort in expressing their ideas to the teacher, approaching the teacher for support, and responding to the teacher.

**Regard For Student Perspectives.** This aspect of classroom quality reflects the degree to which teacher-student interactions emphasize student interests and points of view. Teachers who score high on this aspect of Emotional Support incorporate student interests and ideas into lessons and classroom activities (Pianta et al., 2008). They also offers support for student autonomy, provides opportunities for student expression, and allows freedom of movement during activities (Pianta et al., 2008).

Much of the variability which can be seen in classrooms is a function of the reality of the complexity of the classroom environment, where activities and the nature of teacher and child interactions are changing constantly. Understanding the possible mechanisms of this source of variability which might be seen in aspects of classroom quality can help in distinguishing this variability which manifests in ratings from that which might come from the rater.

The complexity which is inherent in the classroom suggests that teachers' goals are a function of the changes in the interactions which are taking place (Curby et al., 2011). That is, as the classroom environment is constantly changing, because of the given activity, so too are teachers' objectives. Day to day variability in classroom activities and variation in the context across different classrooms means that a small sample of lessons can vary widely on scores even when they are scored by the same rater. Part of the variability observed in classroom interactions can be because of what aspect of the classroom environment is being observed, where some aspects inherently involve more variability in interactions than others. That is, the stability of children's classroom experiences changes with the different aspects of classroom interactions which are being observed. For instance, in one study, stability or variability in the quality of classroom interactions was seen to be partially dependent on the domain which is being assessed. Students in third and fifth grade classrooms experience much stability in the quality of classroom interactions. However, substantial variability was observed, depending on the domain under observation (Curby, Grimm, & Pianta, 2010). Experiences of Classroom Organization and Emotional Support were most stable within a day and Curby et al.

(2011) found that in general, dimensions of Classroom Organization and Instructional Support are more sensitive to activity setting than dimensions of Emotional Support. This is perhaps because they speak more to the characteristic responses of a teacher than to just subjective reactions to what is happening in the classroom. That is, it seems less vulnerable to the changes in activities in the classroom and aspects of these activities. This shows that Emotional Support is a behavior which might be less dependent on what is happening in the classroom at any given time and perhaps a better gauge of the actual teacher. This is an indication that Emotional Support, as it is measured and what it encompasses, is different than Classroom Organization and Instructional Support. We might say that teacher-student interactions in terms of Emotional Support are least affected by elements of the classroom. So, a teacher's interactions with a student, in terms of Emotional Support will be least affected by what is actually going on in the classroom at that time when compared to the other Instructional and Organizational domains of the CLASS.

Variability and stability of classroom interaction quality might also depend on how the teacher's behavior is defined (Curby et al., 2011). Teaching behaviors have been defined in various ways, and thus, have acquired different results in investigation of the stability of these behaviors. For instance, one way teaching behaviors have been defined and investigated has been to look at frequencies of discrete teaching behaviors, such as, the use of certain phrases, the number of questions asked, etc. By examination of these metrics, classrooms seem to be relatively stable (Darling-Hammond, Wise, & Pease, 1983). Yet, a different conceptualization of teaching behaviors focuses on the amount of



time teachers dedicate to certain classroom activities (Meyer, Linn, & Hastings, 1991). Here, when rank-ordered, the behaviors across teachers prove to be fairly stable over the course of one academic year to the next, and from the morning to the afternoon. Also involved in this observed stability might be the emotional display rules expected of teachers where without necessarily knowing what these rules are, teachers do exert emotional labor on the job (Levine-Brown, 2011).

In regards to Emotional Support, variability and stability might also have to do with teacher years of experience. In investigation of student, self, and school level factors which influence teacher Emotional Support consistency, it has been shown that novice teachers, while perceiving themselves to be more emotionally supportive, are less emotionally consistent in the classroom (Bailey, Carlson, Brock, Curby, & LoCasale-Crouch, *in revision*). Middle career teachers who perceived themselves as providing more organizational support showed higher stability in Emotional Support. Consistency in the Emotional Support of late career teachers was found to not be influenced by student, self, or school level factors. These results show that external factors might have less influence on Emotional Support consistency, such that it might be the case that internal teacher characteristics influence Emotional Support consistency (Bailey et al., *in revision*).

Classroom observation measures have many sources of error when it comes to measures of the attributes of a teacher and their teaching, even when those attributes are precisely and narrowly defined. The results of this study suggest that the stability of teacher Emotional Support might also be due to a rater effect of a rater making up their

mind, such that previous ratings influence future ratings. That is, raters may increasingly make up their minds on a rating and generally stick with it throughout conducting ratings of teacher Emotional Support. It has been shown that despite efforts to retrain raters through calibration sessions and frequent feedback, raters change their judgments over time (Bock, 1995; Congdon & McQueen, 2000; McKinley & Boulet, 2004). Although, Cash, Hamre, Pianta, & Myers (2011) found that it is possible to train large numbers of raters to achieve calibration on the CLASS instrument. However, rater effects seemed to be central in predicting the degree of calibration, specifically, consistently so were rater beliefs about teachers and children.

Recall that the question of this study is not only how stable are the ratings of Emotional Support, but also, to what degree are ratings stabilizing over a lesson. In combination, we can better understand the experiences of children during a lesson. In order to assess the degree to which raters are becoming more stable in their ratings, the average correlation of errors were calculated across adjacent time points. It was hypothesized that residual variation among ratings of teacher Emotional Support will decrease over time. Looking at the correlated errors of the ARMA model allows us to determine whether the correlation residuals are increasing or decreasing over time. A decrease in correlation residuals indicates a negative autoregressive correlation, which means that a rater is becoming increasingly consistent in his ratings of Emotional Support. This means that the actual experience of a child may not be as stable as the stability estimate would suggest, and generally allows for a more accurate understanding of a child's experiences in the classroom.

### **3. Importance of Emotional Support Stability**

There are different ways of looking at stability. This would include looking at stability within and across teachers and over different time periods. Previous scholarship has looked at stability of classroom interactions over different time periods (over a school year and within days in the school year). The National Institute of Child Health Early Child Care Research Network (NICHD ECCRN, 2005) carried out a large scale study of classroom quality and teacher and student behavior showing that children experience high patterns of variability across classrooms.

In addition to examining the nature and quality of children's experiences, the stability in children's classroom experiences from first grade to third grade was examined. That is, stability was assessed from a child's perspective across different teachers. Detailed observations of a typical day in 800 third grade classrooms revealed that from first to third grade, children's experiences of global elements of the classroom, such as teacher sensitivity or positive climate, exhibited low stability. Thus, if a child were observed in a classroom rated high on one of these dimensions in first, it was unlikely that their classroom would also be high in that dimension during third grade. It would seem that children's experiences of global aspects of the classroom setting are not extremely stable across years. In comparing mean levels of classroom and teacher

features, across years, classrooms showed lower levels of negative climate as well as lower positive climate and teacher sensitivity and engagement than shown in first grade classrooms. A different way that stability has been looked at is within grade level. Again using data from the NICHD Study of Early Child Care and Youth Development, correlations of scores over the course of days in the quality of interactions in the classroom were high (Chomat-Mooney et al., 2008; NICHD ECCRN, 2005).

Finally, looking at within-day variability of classroom experiences, Curby et al. (2011) found that in third and fifth grade classrooms, students experience great stability in the quality of classroom interactions, with substantial variability dependent on domains. Low stability estimates indicate that children's experience of classrooms is least consistent within a day in terms of instruction, and more constant on aspects of classrooms organization and Emotional Support. In terms of between-grade comparisons, fifth graders experience less stability in classroom interactions than do third graders. This suggests general consistency over time in the behavior of teachers when conducting within-grade and within classroom analysis.

From the previously mentioned results, we can see that children experience different amounts of variability depending on whether we are looking at variability within the classroom, or comparing classrooms on some level. After conceptualizing children's experiences in this regard, it is appropriate to look at the outcomes related to these different amounts of variability which children seem to be experiencing, and for our purposes, specifically in regards to Emotional Support. Similar to differences seen in the

actual amount of variability that children experience in the classroom, research on child outcomes associated with that variability in teacher Emotional Support has garnered varied results. In regards to pre-kindergarten, Emotional Support consistency was related to better social and academic outcomes for children (Curby et al. 2013). Relations between pre-kindergarten teachers Emotional Support consistency and children's social competence and problem behaviors has been found to be mediated by closer and less conflictual relationships with children. (Brock & Curby, in revision a) Additionally, teacher Emotional Support consistency in third grade has been found to interact with child adaptability in predicting academic skills and social skills (Brock & Curby, in revision b).

Variability in Emotional Support, to the extent that it affects child outcomes, might do so differently, depending on teacher overall level of Emotional Support, such that, the most advantageous outcomes across social and cognitive domains of child development are associated with high and consistent levels of Emotional Support. In experiencing lower levels of Emotional Support, worse child outcomes are associated with consistent levels of that lower-level support, than are with more variable levels of that lower-level support (Zinsser, Bailey, Curby, & Denham, *in press*).

Aside from showing us that there are differences in the variability which children experience in the classroom, and in the outcomes associated with that variability, these findings further point to the appropriateness of investigating the different ways teachers interact with students, as opposed to solely attending to mean ratings of classroom

observations, where much of the variability that actually exists might be lost and the information previously mentioned could not have been gained. It also points to the necessity of investigating both the level and variability of classroom interactions together, in order to garner the most appropriate understanding of child experiences in classrooms.

#### **4. Problems Associated with Observer Ratings**

This study investigates the possibility that raters become more consistent in their ratings of Emotional Support over time, largely because their previous ratings are influencing their subsequent ratings. Research which has been done on classroom observation measures, and the potential for this type of error, where ratings that raters conduct include something other than information about what they are observing, while not extensive, has been broadly carried out. In this effort, it has been seen that ratings are a result of both the teacher and the rater.

Classroom observation measures have many sources of error when it comes to measures of the attributes of a teacher and their teaching, even when those attributes are exactly and narrowly defined. Researchers have over the past several years become concerned with problems associated with these ratings (Guilford, 1954). First, observers can only observe a short time period in video-based classroom research and one of a few lessons are videotaped, threatening the validity of these ratings (Casabianca, 2013). Research shows that the judgment processes of trained raters can lead to biased ratings, such that rater bias is defined as disagreement among raters which can be traced to different interpretations of rating scales or unique, idiosyncratic perceptions of the target in the question, or changes over time in a single rater's judgments (Hoyt, 2000).

Variation among raters and changes in raters' use of a measure's scale over time can contribute to error. While there have been efforts to retrain raters with calibration sessions and constant feedback, judgments of raters can still change over time (Bock, 1995; Congdon & McQueen, 2000; McKinley & Boulet, 2004) and in cases where calibration efforts have been deemed successful, rater beliefs about teachers and students affected the degree of calibration (see Cash, Hamre, Pianta, & Myers, 2011).

Investigations dealing with the type and extent of rater bias in classroom observations have been carried out regularly. A meta-analysis of Hoy and Kerns (1999) concludes that 37 percent of the variance in ratings is due to rater bias. In addition, they looked at moderators of rater bias and concluded that the highest risk rates are inferential ratings by raters with less than five hours of training. One cannot conclude that if raters receive sufficient training that ratings of Emotional Support can be conducted without problems however. Rater trainings which are concerned with complex objects (like Emotional Support) are not automatically effective in dealing with rater bias and accuracy as some have found (see Lumley & McNamara, 1995). Researchers doing video-based classroom studies with high inference ratings assume as a rule that training is effective when there is consensus about a joint theoretical understanding in the training group (Rakoczy & Pauli, 2006; Seidel, 2005). It remains unclear whether rater training really works as intended.

More recent research on rater bias has been done looking specifically at rater severity drift, central tendency, and rater experience/learning (Leckie & Baird, 2011; Myford & Wolfe, 2009). Casabianca and colleagues demonstrated that domain scores



from video observations of the Classroom Assessment Scoring System, Secondary (CLASS-S) followed a downward trend throughout the scoring period and then increased their scores later in the scoring period (2012). They found that during the scoring period, there was a one point decrease and then a one point increase on the 1 to 7 Likert scale. Casabianca and Lockwood further carried out a study looking at the nature of rater variation due to rating severity, including overall time trends and variations among raters (2013). The Emotional Support and Instructional Support domains appeared to be more sensitive to time, and the Classroom Organization domain appeared to stabilize soon after the start of scoring. Most raters gave higher scores to the Classroom Organization domain. There was however, much variation in the trends over scoring days in terms of the domain under investigation and the rater.

In this study, the stability of teacher Emotional Support might be due to a rater effect of a rater making up their mind, such that previous ratings influence future ratings. That is, raters may increasingly make up their minds on a rating and generally stick with it throughout conducting ratings of teacher Emotional Support. It has been shown that despite efforts to retrain raters through calibration sessions and frequent feedback, raters change their judgments over time (Bock, 1995; Congdon & McQueen, 2000; McKinley & Boulet, 2004). Although, Cash, Hamre, Pianta, & Myers (2011) found that it is possible to train large numbers of raters to achieve calibration on the CLASS. However, rater effects seemed to be central in predicting the degree of calibration, specifically and consistently were rater beliefs about teachers and children. For this study, it was hypothesized that residual variation among ratings of teacher Emotional Support would

decrease over time. This question was addressed by looking at the correlated errors of the ARMA model.

Knowing about the nature and extent of rater variation over time in scoring is crucial to practitioners and researchers in designing measurement systems for teaching that are minimally impacted by these sources of variance and for preventing them from becoming biases (Casabianca, 2013).

## **5. Quantification of potential rater effects in measuring teacher Emotional Support**

In order to assess how much of an influence that rater effects have on measuring a conceptually difficult construct, like Emotional Support, different statistical frameworks have been used. While every scientific study has to report whether the ratings they use for their conclusions are sufficiently reliable, the efficacy of observer ratings is not directly investigated (Praetorius, Lenske, and Helmke, 2012). Various coefficients have been developed in order to quantify reliability, however these only allow for the investigation of one type of reliability at a certain time point. Furthermore, there is not information provided about the amount and causes for bias with the reliability coefficients. For these reasons, Chronbach, Gleser, Nanda, and Rajaratnam (1972) developed generalizability theory (G theory). An advantage of G theory is that the resulting variance components can be used as inputs in order to estimate the reliability under multiple measurement conditions in a subsequent step (known as a decision study or D study). This information is convenient for research practice, allowing researchers to carry out more precise yet economical investigations.

Generalizability theory (or G theory) is a powerful framework by which to assess ratings, allowing the separation of multiple sources of error through variance components (Brennan, 2001; Shavelson & Webb, 1991) and serves as a means for investigating the dependability of behavioral measurements. Praetorius, Lenske, and Helmke (2012) used

generalizability analysis in order to determine how reliably and validly instructional quality is measured by observer ratings. They found that 16-44% of the variance in ratings could be attributed to instructional quality, whereas rater bias accounted for 12-40% of the variance. It seems that rater bias contributes very similarly in its accounting of the variance in ratings to actual instructional quality. These findings point to the appropriateness of carrying out the proposed study and the criticalness of not solely looking at the reliability of ratings but also their validity. This current research in the field of teacher quality shows that rater effects are not only real, but, in some cases, can be substantial. In getting at rater effects in a manner which has not been done before (through ARMA) we will be able to assess the power of this methodology and compare it to other frameworks which have been used to evaluate the influence of errors in ratings, like the powerful generalizability theory. These findings might suggest that we should treat observer ratings of its aspects in a more differentiated manner (Praetorius, Lenske, & Helmke, 2012).

## **6. Research Questions/Hypotheses**

The first aim of this study was to assess the minute-to-minute stability of aspects of teacher Emotional Support in pre-kindergarten classrooms. That is, how stable are minute-to-minute classroom Emotional Supports overtime? Developmental theory states that the proximal processes drive development. The present study examines the stability of those processes (teacher Emotional Support interactions) on the time scale in which they take place. Furthermore, raters may determine that a teacher fits a certain score profile as opposed to actually making independent ratings at each time point. Thus, the second aim of this study is to estimate the effect of the rater on the observed stability of teacher Emotional Support. Ratings of teachers' Emotional Support may become more stable over time as raters increasingly make up their minds. This study allows us to get an understanding of how both the rater and teacher might influence the stability of Emotional Support. Answering the questions posed in this study provides us insight into how teachers and raters change over time in the measurement of Emotional Support and helps us to better understand rater effects such as non-independent ratings over time.

Accordingly, I hypothesized that observed minute-to-minute classroom Emotional Support experiences of students would exhibit moderate to high levels of stability over time. Moderate to moderate high stability has been observed for each of the dimensions of Emotional Support over the first four hours of a school day (Curby et al., 2011),

potentially as they are characteristic of teachers' classroom responses. Teachers will demonstrate stability in these aspects which seem to be less dependent on what is going on in the classroom at a certain time, and are more characteristic teacher responses.

In terms of the dimensions of Emotional Support, I hypothesized that Positive Climate will demonstrate the least stability and Regard for Student Perspectives will be the most stable. Because Positive Climate is so encompassing of different aspects of Emotional Support, it is possible that it introduces greater opportunity for inconsistencies. In terms of Regard for Student Perspectives, it is possible that this dimension of Emotional Support is least dependent on what is going on in the classroom, and so is most stable in its manifestation as a teacher behavior.

Secondly I asked, are the ratings of these aspects of Emotional Support becoming more stable over time? I hypothesized that ratings of Emotional Support will become more stable over time. This hypothesis is based on the notion that raters increasingly make up their minds about the level of Emotional Support aspects being demonstrated by teachers. As has been reviewed, rater effects are real and the value of this study is that it allows us to investigate this important aspect of how raters are changing in their ratings of that Emotional Support over time, while, at the same time, understand how teacher Emotional Support is changing over time. Therefore, understanding of the nature of teacher Emotional Support and ratings of that support is gained through one statistical analysis.

## **7. Method**

Data was obtained from classrooms in which teachers participated in the National Center for Research on Early Childhood Education's Professional Development System study. Teachers were randomly selected to participate after they consented to the study conditions. Participants were selected from publically-funded pre-kindergarten programs where the majority of children were eligible to enter kindergarten the following school year, and did not have an individualized education program (IEP) at the start of the academic year. The participating teacher was the lead teacher and the majority of instruction provided was in English. The purpose of the NCRECE PDS study was to investigate whether the quality of teacher's interactions with children might be improved through use of a web-based intervention. All teachers included in the study provided videotapes of classroom instruction throughout the course of the school year, that is, approximately every other week. All intervention teachers took part in a four-step process with an assigned consultant throughout the course of the year. The four-step process in which teachers were involved proceeded as follows: Teachers videotaped themselves instructing students. A consultant reviewed the classroom observation videotape and subsequently posted short video clips as well as written prompts on a private, secure website, which would then be reviewed by the respective observed teacher. Teachers viewed the edited video of their instruction and responded to prompts with an online

journal. Lastly, the teacher and the consultant took part in an online videoconference in order to discuss the edited classroom video and other topics related to classroom performance and determined goals for possible future cycles.

This study used data from all teachers in the treatment condition of the Professional Development System (PDS), which were coded every minute, from one video segment from the spring (approximately 30 minutes) from each teacher. The entire PDS sample included approximately 350 pre-kindergarten teachers. For the purposes of this study, solely data on those classrooms where teachers participated in the intervention were utilized, such that the sample includes 72 teachers. The videos were coded on each dimension of Emotional Support (positive climate, negative climate, teacher sensitivity, and regard for student perspective) every minute, for a total of approximately 30 minutes of video coding of teacher instruction for 72 teachers.

Video segments from a literacy lesson were coded using the Classroom Assessment Scoring System (CLASS) (Pianta, La Paro, & Hamre, 2008). The CLASS was not used in the traditional way with ratings of multiple cycles over a 15 to 20 minute period. Instead, raters rated teachers on the four dimensions of Emotional Support every minute. Ratings of the dimensions of Emotional Support at every minute were individually investigated and also averaged to obtain an overall Emotional Support score at every minute.



Videos of teachers were coded using the Classroom Assessment Scoring System (CLASS) (Pianta, La Paro, & Hamre, 2008). The CLASS measures the quality of teacher-child interactions across three domains: Classroom Organization, Instructional Support, and Emotional Support. 10 observable dimensions are scored on a 7-point Likert scale from 1 *low* to 7 *high*. For the purposes of this study, focus will be dedicated to the four dimensions of the Emotional Support domain: *positive climate*, *negative climate*, *teacher sensitivity*, and *regard for student perspectives*.

Training for the CLASS coders took place until all raters were reliable. During the course of training, raters viewed and extensively discussed video segments of real life early childhood classrooms. Over the course of the reliability phase of training, raters were held to a gold standard composition of ratings. Raters were considered reliable when their agreement with the gold standard, plus or minus one scale-point, matched or exceeded 80%. For the purposes of this study, the ratings of only one rater will be analyzed who coded all video segments. This is a way to avoid the introduction of confounding variables, as all teachers in the sample were not rated by all raters. That is, this respective rater is perfectly reliable with himself.

## **8. Data Analysis**

In order to address the question of the minute-to-minute stability of classroom Emotional Support which children experience over time, ARMA models were fit to the data. Analysis and interpretation of rating error variance were conducted. In addition to looking at scores of aspects of emotional Support, average scores of Emotional Support were also calculated, given scores of these four dimensions, at every minute.

ARMA models are appropriate to use when there is solely one component of interest that has been repeatedly measured (i.e., classroom Emotional Support) (Hasan & Thaut, 1999). ARMA modeling is thus a method for parameterizing complex system dynamics. It is based on the notion that time points are correlated across time and may be expressed as an autocorrelation function. It is a representation of the best linear predictor of Emotional Support, from all previous measurements. In other words, the ARMA model utilizes the average across the observed data in the autoregressive portion of the model, as opposed to using the observed prior time point the way an Autoregressive Model (AR) does. This MA component provides us with the best guess for any given time point of Emotional Support.

The ARMA model parameters for Emotional Support were computed by recursive calculating of the autocorrelation function (ACF) of interresponse intervals, which

measures the correlation of signal  $x(t)$  (Emotional Support) with itself, as shifted by some delay in time.

Stability. The  $z$ -transformation for standardization is often used in order to analyze the stability of linear systems (Hasan & Thaut, 1999). The stability of the difference equations would thus be calculated for all the aspects of Emotional Support and average Emotional Support at every minute. This is the autoregressive component of the model and was used to evaluate the first research question about the stability of minute-to-minute experiences of children in these classrooms. It was predicted that minute-to-minute Emotional Support would exhibit moderate to high stability over time, with specifically Positive Climate exhibiting the least amount of stability and Regard for Student perspective showing the most stability.

In this study, at least some of the stability in Emotional Support might also be due to the rater increasingly making up their mind about the level of Emotional Support displayed. In this way previous ratings would influence current ratings. In order to assess the degree to which raters are becoming more stable in their ratings, the average correlation of errors were calculated across adjacent time points. This allowed us to determine whether the correlation residuals were increasing or decreasing over time. A decrease in correlation residuals is indicative of a negative autoregressive correlation, which means that a rater is becoming increasingly consistent in his ratings of Emotional Support. In addition, in looking at the Autocorrelation Function (ACF) plot and the Partial Autocorrelation Function (PACF) plot, it can be seen if there is any autocorrelation

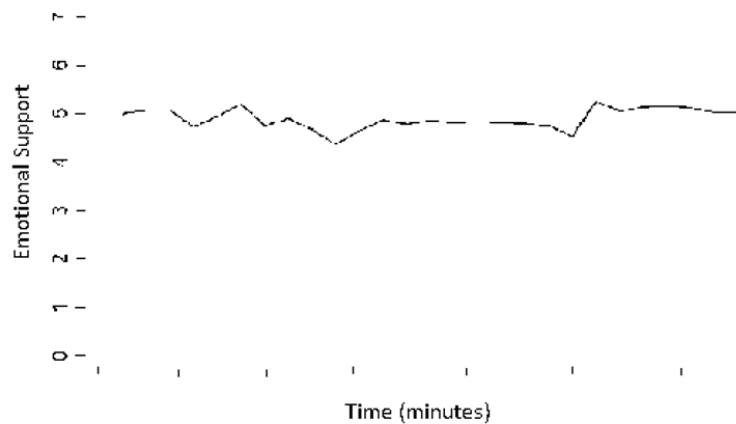
which exists in the time series (Gottman, 1981). The ACF plot is a bar chart of the coefficients of correlation between a time series and lags of itself, and the PACF plot shows the amount of correlation between a lag and itself that is not explained by previous lags (Nau, 2005). If a rating of teacher Emotional Support is influenced by previous ratings, the ACF plot would show a decline from a high correlation to a low correlation (a negative autocorrelation), while the PACF plot would show only a significant spike at lag 1 (2005).

It is in combining the information that we get from both parts of the model (i.e., the autoregressive portion and the moving average (correlated residuals) component that a powerful picture of the measurement of Emotional Support emerges. For example, let us imagine that the stability estimate generated by the autoregressive portion of the ARMA model was moderately strong (0.80 or above). If there was also a negative correlation between residuals, it would be an indication that some of this stability is due to the raters becoming more consistent in their ratings of Emotional Support over time and that the actual experience of a child may not be as stable as the stability estimate would suggest. On the other hand, if there were this same .80 stability estimate along with a positive correlation between residuals, it is an indication that raters were becoming less consistent in their ratings of Emotional Support over time and that the actual experience of a child may be more stable than the stability estimate would suggest. Although it is more difficult to think of why this might be, it could be due to raters questioning their prior ratings, where they might begin to intentionally vary subsequent ratings as a result. In both of these cases, the rater is taking prior information into account

when making each rating and not making an independent observation. Conversely, if there were this same 0.80 stability estimate along with a zero or near-zero correlation between residuals, it is an indication that teachers were rated moderately strong in Emotional Support, and raters were not becoming more or less consistent in their ratings of Emotional Support over time. This is the scenario which would be ideal, in that it demonstrates that prior information is not influencing ratings of Emotional Support over time (i.e., ratings are independent). Because the moving averages component of an ARMA model is the best guess for any given time point of Emotional Support, raters should not be becoming more or less consistent in their ratings over time. We should ideally see no autoregressive correlation, in one direction or another.

## 9. Results

The first aim of this study was to assess the minute-to-minute stability of aspects of teacher Emotional Support in pre-kindergarten classrooms. Autoregressive moving average (ARMA) analyses were carried out on the domain of Emotional Support as well as the four dimensions: Positive Climate, Negative Climate, Teacher Sensitivity, and Regard for Student Perspectives. Thus, a total of five ARMA models were fit to the data. It was predicted that minute-to-minute Emotional Support would exhibit moderate to high stability over time,

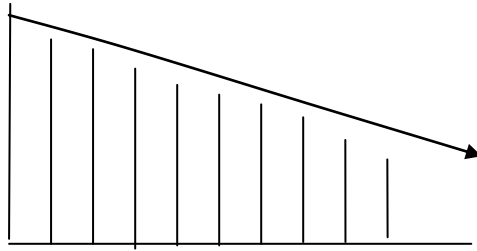


*Figure 1: Hypothesized Emotional Support Stability (by minute over time)*

Figure 1 line graph shows that the change at every minute is such where there is stability over time, i.e. the plot shows that there is not substantial change in Emotional Support and its dimensions which is associated with changes in time. This is approximately what would be obtained in modeling teacher Emotional Support and its individual dimensions at every minute over time, as determined from previous studies which have examined stability of classroom interactions in different capacities..

The second aim of this study was to estimate the effect of the rater on the observed stability of teacher Emotional Support. It was predicted that raters would become increasingly stable in their ratings of teacher Emotional Support over time. The moving average portion of the model is the focus of investigating the rater effect on the stability of teacher Emotional Support. As such, the Autocorrelation function plot (ACF) and Partial Autocorrelation plot (PACF) are of interest in investigating the rater effect on ratings, where the prior shows the coefficients of correlation between a time series and lags of itself, and the latter plot shows the amount of correlation between a lag and itself that is not explained by previous lags (Nau, 2005). Thus, the partial autocorrelation at a given lag is the difference between the actual correlation at that lag and the expected correlation due to the spreading of correlation at the first lag.

If the prediction about raters is correct along with the prediction about the stability of teacher Emotional Support and its dimensions, we would expect the following ACF outcome.



*Figure 2: Hypothesized autocorrelation function of residuals*

This hypothetical ACF plot demonstrates the expected relationship between the time lags within overall Emotional Support and its individual dimensions. If a rating of teacher Emotional Support is influenced by previous ratings, the autocorrelation function would show a decline from a high correlation to a low correlation. Thus, the ACF plot would show this strong autocorrelation (where at the first lag there is high autocorrelation, and then a slow decline over time).

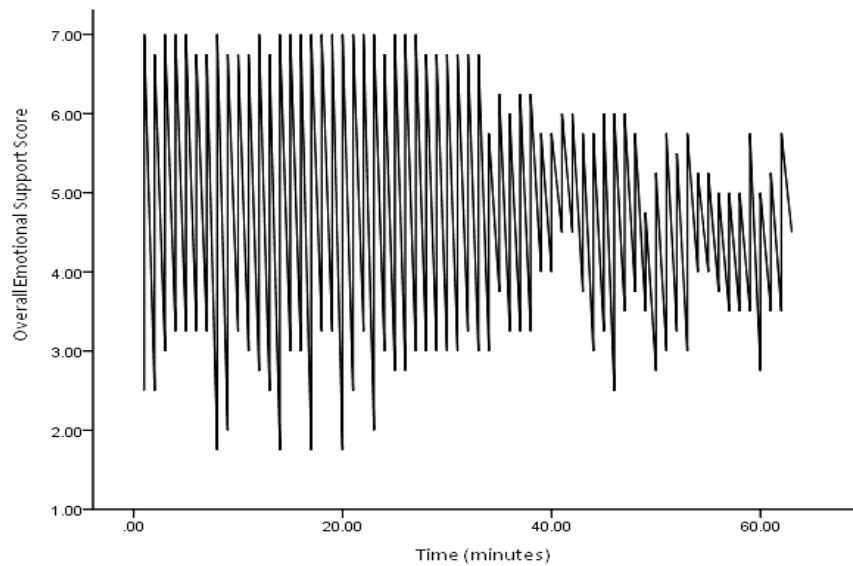
The PACF plot depends on the amount of correlation that is not explained by lower order lags and the amount of spreading of autocorrelation at lag 1, and could be of different variations. For instance, in the case that autocorrelations after lag 1 are due to the spreading of autocorrelation at lag 1, the PACF plot will exhibit a significant spike solely at lag 1. This means that all higher order autocorrelations (at lag 2 and above) are explained by the autocorrelation at lag 1 (Nau, 2005). On the other hand, if the PACF



plot demonstrates a more gradual decay (i.e. shows significant spikes at higher lags), this is an indication that the observed autocorrelation is not due to the spreading of autocorrelation at lag 1, as this plot shows the amount of correlation between a lag and itself that is not explained by previous lags (2005). Therefore, it is necessary to observe both plots together in order to understand the existence and activity of autocorrelation in a time series.

### **Emotional Support Domain**

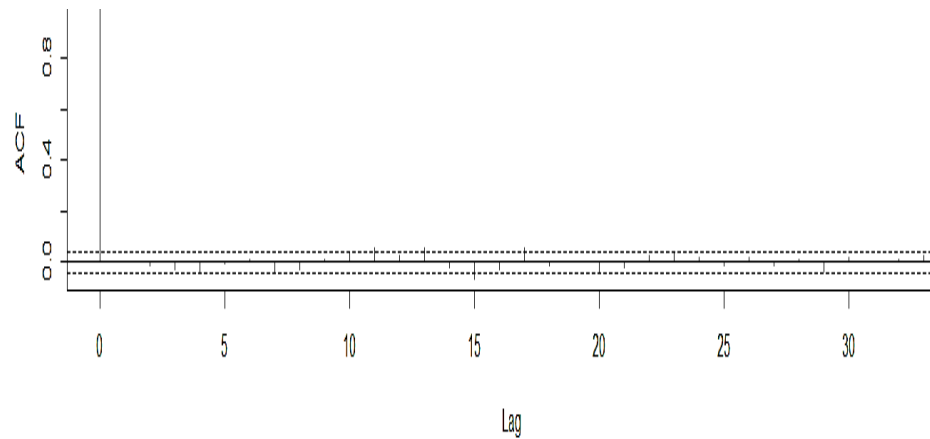
ARMA analysis indicated that overall Emotional Support was not very stable at all with an estimate of 0.087 ( $p < 0.00$ ) (see Table 1 for all ARMA stability and correlation and descriptive statistics). This weak stability estimate is shown in the line graph p in Figure 3, which depicts overall Emotional Support as it changes from minute to minute. The wide band of scores indicates that there was a lot of variability in scores from minute to minute. However, Figure 3 also shows that the band of scores narrows over time suggesting that overall Emotional Support may be becoming more stable over time.



*Figure 3: Overall emotional support score over time by minute*

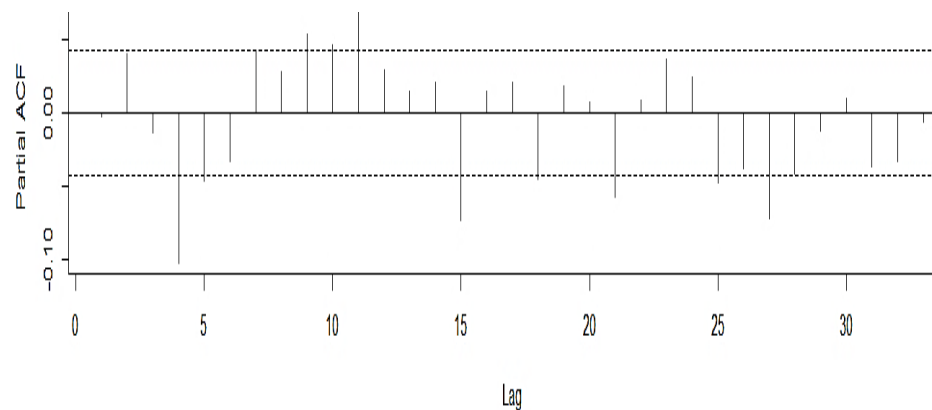
We were also interested in investigating the degree to which raters became more stable in their ratings, through the moving average component, which calculates the average correlations of errors across adjacent time points. Emotional Support demonstrated a strong negative correlation between residuals ( $r = -0.96$ ,  $p < 0.00$ ). The following ACF plot of Emotional Support (Figure 4) shows a sharp cutoff with a negative autocorrelation, where the autocorrelations are significant at the first lag. This negative autocorrelation results in a time series that displays a moving average (MA) signature which is observed as a spike in the first one or more lags of the ACF. This observation is

usually associated with a negative autocorrelation and indicates that the autocorrelation pattern can be explained more easily with an MA term than with an AR term (Gottman, 1981). That is, because there is not a progressive decline in the ACF plot, but rather a sharp cut-off, the MA term best describes the time series, which can be regarded as a moving average of random shocks, as opposed to an autoregressive process where the previous value has a direct effect on current value of the time series. F5



*Figure 4:* Autocorrelation function of residuals of Emotional Support

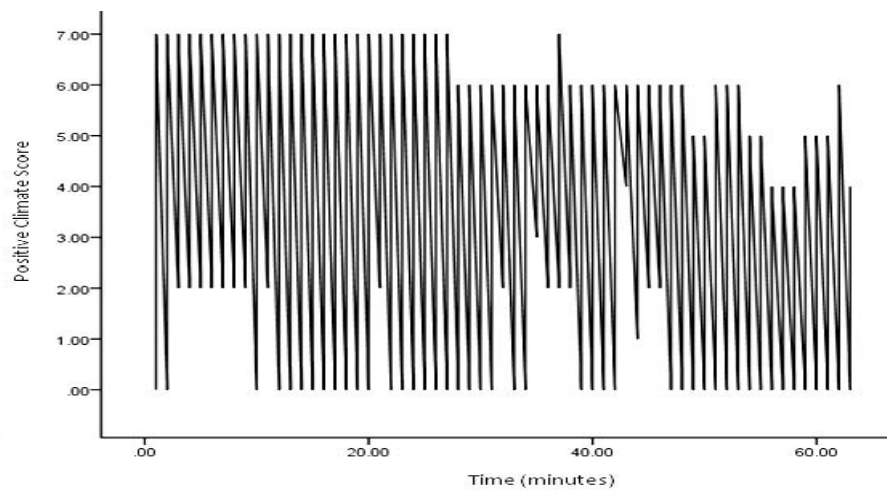
The PACF plot (Figure 4) shows the amount of correlation between teacher Emotional Support and a lag of itself that is not explained by the lower order lags. Thus, the observed partial correlation at lag 2 is the difference between the actual correlation at that lag and the expected correlation due to the spreading of correlation at lag 1 (Nau, 2005). In accordance with the ACF plot (which cut off after lag 1), the PACF has spikes at the higher order lags (lags 2 and above), showing that the partial autocorrelation is not due to the spreading of correlation from the first lag. This observed pattern in the PACF plot indicates an MA signature time series (Nau, 2005).



*Figure 5: Partial autocorrelation function (PACF) of residuals of Emotional Support*

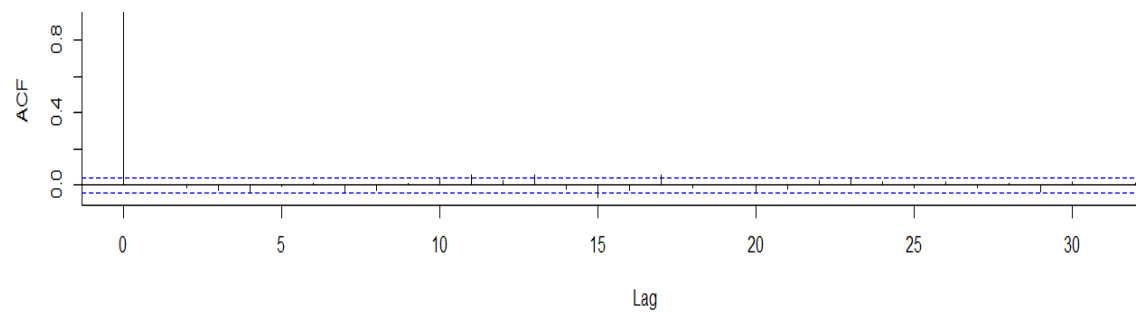
## Positive Climate

Positive Climate had a stability of 0.068 ( $p = 0.00$ ). This autoregressive term answers our first question of how stable the minute-to-minute ratings of Positive Climate are. This low stability can be observed from Figure 6 . Specifically, we can observe overall weak stability of Positive climate from minute to minute, with some pockets of greater stability at different time points. Thus, Positive Climate at every minute does not demonstrate equal low stability.



*Figure 6:* Positive Climate score over time by minute

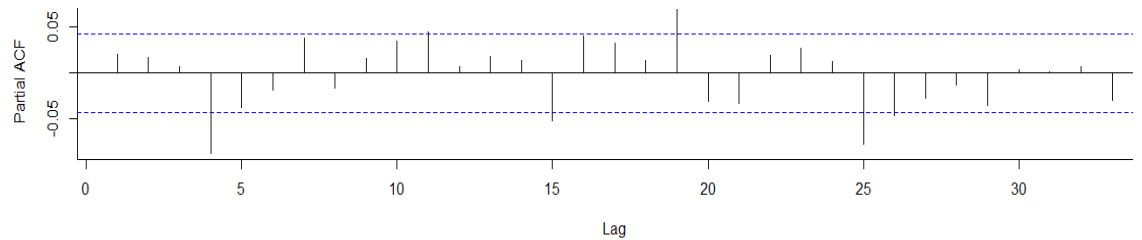
A negative correlation between residuals was observed for Positive Climate ( $r = -0.96$ ,  $p < 0.00$ ), which indicates a strong negative autoregressive correlation. This means that a rater is becoming increasingly consistent in his ratings of Positive Climate over time. Similar to the ACF plot of Emotional Support, the plot (Figure 7) shows a sharp cutoff with a negative autocorrelation, where the autocorrelations are significant at the first lag. The spike observed in the first one or more lags of the ACF indicates that the autocorrelation pattern can be explained more easily with an MA term than with an AR term.



*Figure 7: Autocorrelation function of residuals of Positive Climate*

The PACF plot (Figure 8) shows the amount of correlation between Positive Climate and a lag of itself that is not explained by the lower order lags. In accordance

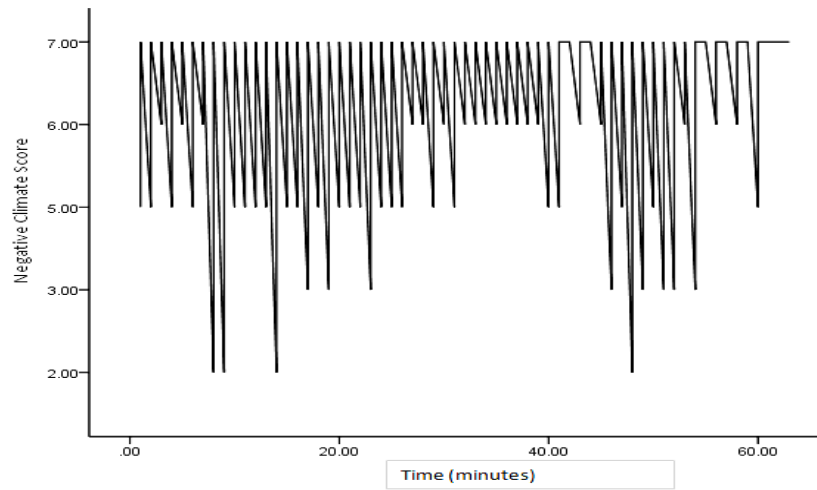
with the ACF plot (which cut off after lag 1), the PACF has spikes at the higher order lags (again indicating an MA signature time series) (Nau, 2005).



*Figure 8: Partial autocorrelation function of residuals of Positive Climate*

### **Negative Climate**

The Negative Climate dimension of teacher Emotional Support showed a complete lack of stability (autoregressive stability estimate = -0.02,  $p = 0.35$ ). This can be seen in Figure 9 . The Negative Climate dimension was reverse coded and plotted by minute over time. The scores begin by varying between 5 and 7, then become more unstable, and begin to show more stability briefly after 40 minutes and 50 minutes of observation, although there was still much instability observed at these time points as well Low stability and varying amounts of that low stability depict lack of stability seen in the stability estimate.



*Figure 9: Negative Climate score over time by minute*

Similar to the other three dimensions of Emotional Support, a strong negative autoregressive correlation was observed for Negative Climate ( $r = -0.98$ ,  $p < 0.00$ ). The ACF plot of Negative Climate (Figure 10) shows autocorrelations at the first lag, and then a sharp cutoff. Again this model displays a moving average (MA) signature (meaning that the autocorrelation pattern can be explained more easily with an MA term than with an AR term).



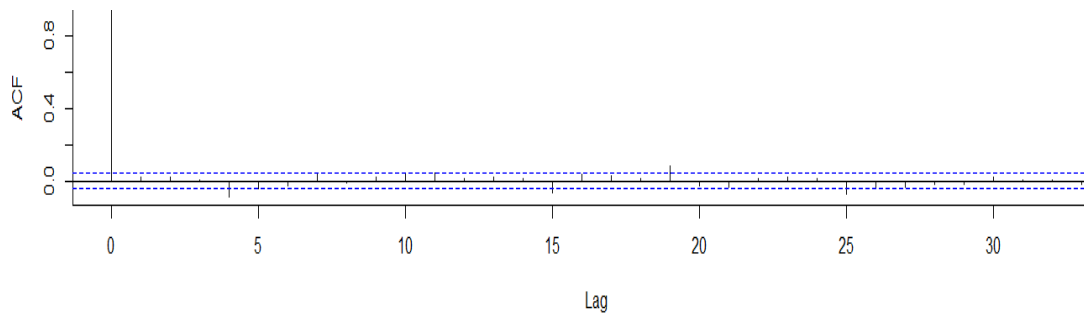


Figure 10: Autocorrelation function of residuals of Negative Climate

The PACF plot (Figure 11) supports an MA signature time series, with spikes at the higher order lags (lags 2 and above). This tells us the difference between the actual correlation at lag 2 for instance and the expected correlation due to the spreading of correlation at lag 1. The spikes indicate that the higher order autocorrelations are explained by the negative autocorrelation at lag 1.

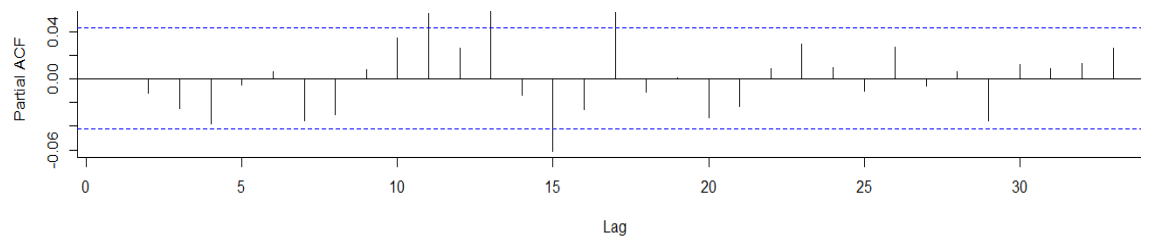
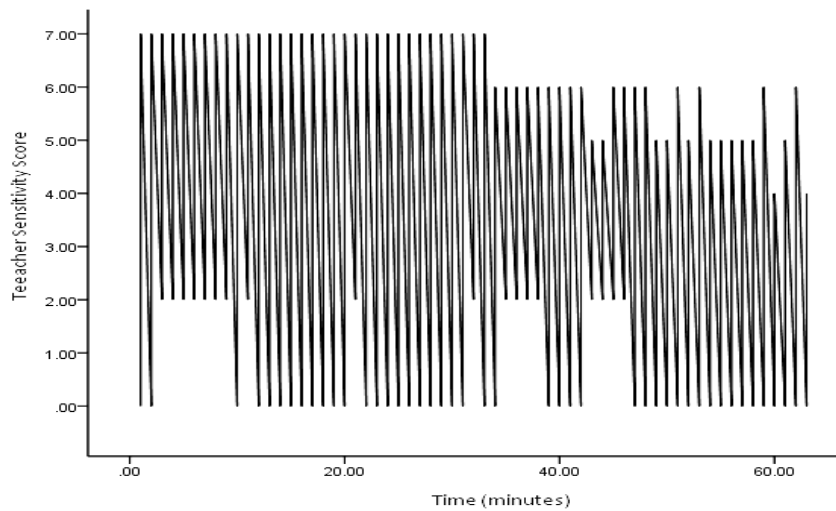


Figure 11: Partial autocorrelation function of residuals of Negative Climate

## Teacher Sensitivity

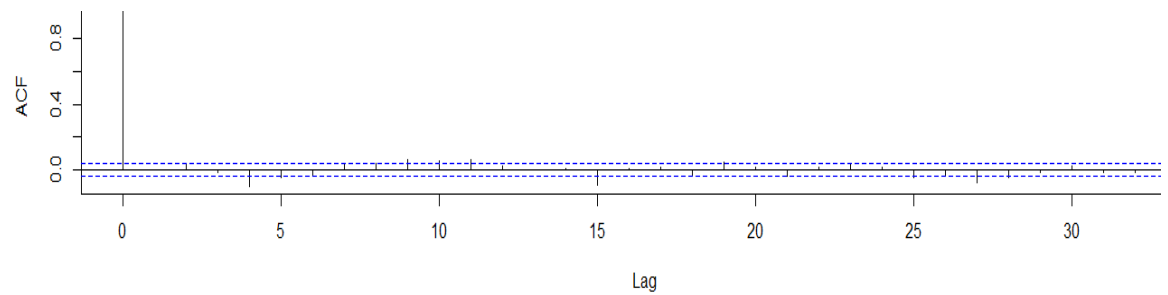
ARMA analysis of Teacher Sensitivity indicated a stability estimate of 0.077 ( $p = 0.00$ ), and was the most stable of the four dimensions over time. This indicates that the minute-to-minute experiences of children in these classrooms, in regards to Teacher Sensitivity, was weakly stable, and more so than children's experiences of Positive Climate. This weak stability can be observed (Figure 12).



*Figure 12: Teacher Sensitivity score over time by minute*

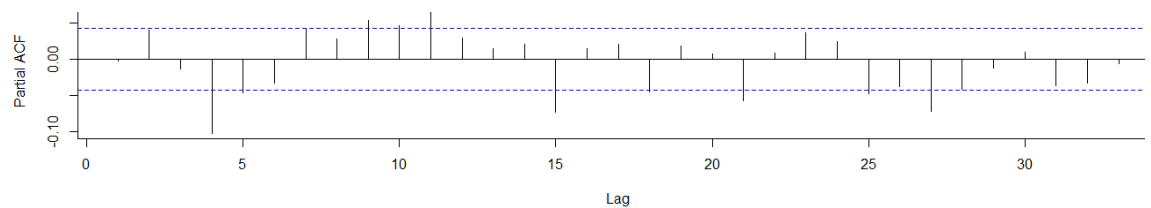
A decrease in correlation residuals was also observed for Teacher Sensitivity ( $r = -0.98$ ,  $p < 0.00$ ), indicating increasingly consistent ratings of Teacher Sensitivity. The

ACF plot of Teacher Sensitivity (Figure 13, ) shows autocorrelations at the first lag, and then a sharp cutoff, displaying a moving average (MA) signature.



*Figure 13:* Autocorrelation function of residuals of Teacher Sensitivity

The PACF plot (Figure 14 ) supports an MA signature time series, with spikes at the higher order lags.

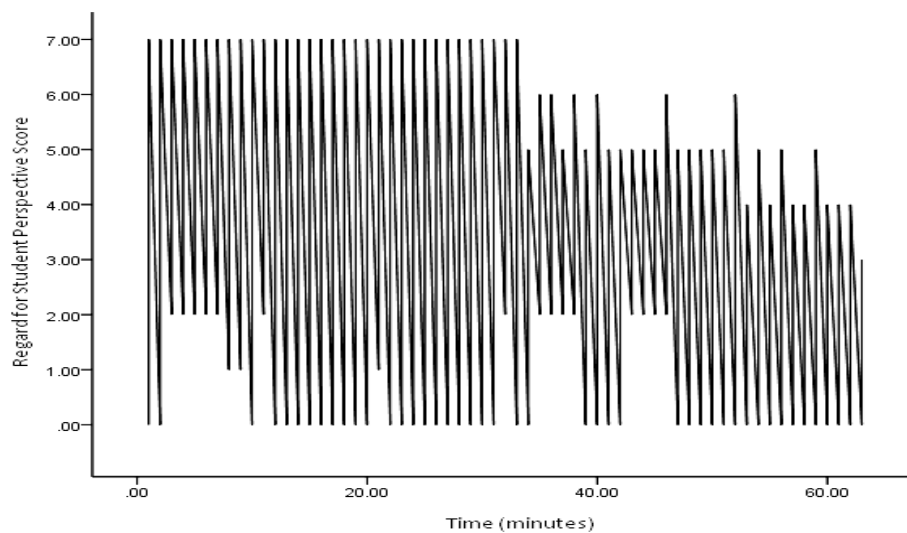


*Figure 14:* Partial autocorrelation function of residuals of Teacher Sensitivity

## Regard for Student Perspectives

Regard for Student Perspective showed a stability estimate of 0.075 ( $p = 0.00$ ).

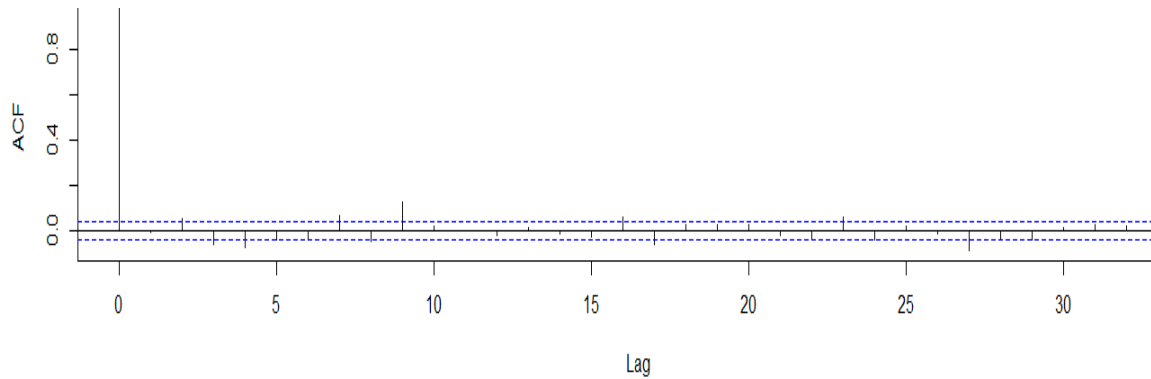
This weak stability of the Regard for Student Perspectives dimension is modeled in Figure 15 .



*Figure 15:* Regard for Student Perspective score by minute over time

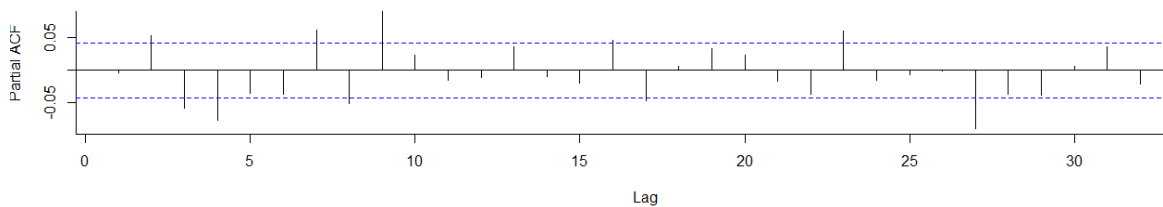
Like for the previous dimensions, there was also a strong negative correlation between error terms ( $r = -0.97$ ,  $p < 0.00$ ). These results indicate that substantial observed stability is due to the raters becoming more consistent in their ratings over time and that the actual experience of a child may not be as stable as even these weak stability

estimates would suggest. The plot below (Figure 16) shows a sharp cutoff with a negative autocorrelation, where the autocorrelations are significant at the first lag



*Figure 16: Autocorrelation function of residuals of Regard for Student Perspective*

The PACF plot (Figure 17) shows the amount of correlation that is not explained by the lower order lags.. Again, along with the ACF plot, the PACF has spikes at the higher order lags, indicating an MA signature to the time series.



*Figure 17: Partial autocorrelation function of residuals of Regard for Student Perspective*

Table 1

*ARMA Stability Estimates of Dimensions of Emotional Support and Overall Emotional Support, With Associated Average Correlation of Errors*

	CLASS		
	Domain/Dimension		
	Mean( $\mu$ )	Autoregressive (AR) Stability Average	Correlation of Errors of Adjacent Time Points (r)
Emotional Support	5.00 (0.903)	0.087***	-0.96***
- Positive Climate	4.58 (1.42)	0.068**	-0.96***
- Negative Climate	1.17 (.62)	-0.02	-0.98***
- Teacher Sensitivity	4.29 (1.52)	0.08***	-0.98***
- Regard for Student	3.97 (1.50)	0.07***	-0.97***
Perspective			

Note: \*\* =  $p \leq 0.01$ , \*\*\* =  $p \leq 0.001$ . Standard Deviations appear in parentheses beside means.

## **10. Discussion**

Results from the present study support the notion that ratings of Emotional Support became more stable as more and more minute ratings were made. All dimensions of Emotional Support, as well as average Emotional Support showed a high negative correlation between residuals. This is an indication that much of the observed stability is due to the raters becoming more consistent in their ratings over time and that the actual experience of a child may not be as stable as the (weak) stability estimates suggests. The ACF plots revealed that the autocorrelation can be more easily explained with an MA term than with an AR term. While the prediction in terms of the autocorrelation direction was confirmed, the potential of this cutoff phenomenon was not anticipated. As a result, the predicted figure of the autocorrelation function of residuals was not confirmed by the observed plots.

Counter to hypotheses about some dimensions being more stable than others, all dimensions showed remarkably little stability. These estimates of stability and their associated residual correlations reveal pertinent information about children's experiences in the classroom. In terms of Emotional Support, students' experiences in the classroom are likely to be less stable than the (weak) stability estimates would indicate, as a result of rater effects, such as non-independent ratings, and raters becoming more consistent in their ratings over time. As we know, pre-kindergarten children who experience more

consistency over a day have been shown to have better academic and social outcomes (Curby et al., 2013). The weak stability which was observed in this study is concerning in regards to those more positive aspects of Emotional Support (Positive Climate, Teacher Sensitivity, and Regard for Student Perspective), because of the worse outcomes of children associated with these less stable experiences. Specifically, as found by Zinsler et al. (in press), worse child outcomes are associated with consistent levels of lower-level support, than are with more variable levels of lower-level support (when looking at the last four months of the school year). Because better child outcomes are associated with more variability in lower-level support, like Negative Climate, the complete lack of stability observed on this dimension of classroom quality is very much encouraging. Indeed, it is further re-assuring if children's experiences of Negative Climate are even less stable than as suggested by the weak stability estimate (because of the negative correlation between residuals).

Aside from the complete lack of stability found for Negative Climate, the results of this study also do not confirm previous findings, which revealed moderate to moderate high stability for the dimensions of Emotional Support over the first few hours of a school day (Curby et al., 2011), although the weak stability observed was significantly different from zero (for all but Negative Climate). Said previous findings show that teachers are stable in their Emotional Support when looking over longer periods of time. This suggests that in looking over longer periods of time, we are missing variability which might exist in teachers' Emotional Support displays. We are able to capture such variability through an investigation which looks within cycles of observation. Therefore,



this study uses a different time scale where raters rated teachers once every minute. The fact that raters rated teachers more frequently introduces greater opportunity for variability to be observed. The observed weak stability might specifically have something to do with Emotional Support being more dependent on and influenced by what is actually happening in the classroom at each minute of observation. That is, the extent to which a teacher demonstrates sensitivity to students in the classroom, takes into account students' needs and preferences, and creates a positive classroom environment, is more a function of what the environmental factors are at a given time and how they might be changing, as opposed to who the teacher is. With traditional CLASS ratings, where raters rate a teacher once every 20 minutes, a rater might decide who a teacher is (in terms of her characteristics), and from then on, not diverge too much from that rating. However, because raters are rating teachers so frequently in this study, their ratings might be highly reflective of what is happening in the classroom at that moment. Thus, they are less stable, and more prone to variation. This study shows that children's experiences of Emotional Support in the classroom are less stable than we might otherwise gather from previous studies, and from even these weak estimates, because of rater effects.

In summary, this study suggests that the variability within a 20 minute observation is being lost, with only one score for each dimension in those 20 minutes. Previous to this study, the variability within a cycle remained uninvestigated, because ratings were solely made overall, at the completion of cycles, and not throughout observations. Further investigation into the variability within rating cycles over a period of time and associated rater effects is necessary.

Why are rater effects so strong in this study? One possible reason is that with greater opportunity to rate teachers, rater effects are more likely to be observed. When a rater rates a teacher once every 20 minutes for three or four cycles of observation, he is required to make an overall judgment about the teacher. That is, he must try to make a general assessment about how to rate a teacher on each of the four dimensions of Emotional Support. The rater cannot record each instance of change (variation) in a teacher's Emotional Support displays. Therefore, less rater effects might be seen in previous studies, where ratings were made overall, at the end of cycles and not throughout observations. In addition to unstable ratings, with the opportunity to rate teachers every minute in this study, we also observed that rater's ratings are not independent of one another, and become more consistent over time.

Other studies have been carried out using the CLASS and also found evidence of rater effects and colleagues demonstrated that domain scores from video observations of the Classroom Assessment Scoring System, Secondary (CLASS-S) followed a downward trend throughout the scoring period, increasing in their scores later in the scoring period (2012). Casabianca and Lockwood carried out a study looking at the nature of rater variation due to overall time trends and variations among raters (2013). Emotional Support and Instructional Support appeared to be more sensitive to time, and Classroom Organization stabilized soon after the start of scoring. The observed much variation in the trends over scoring days in terms of the domain under investigation and the rater.

The most significant contribution of this study is perhaps that it demonstrates the pertinence of using novel statistical techniques in classroom developmental studies which include rater observation. It reveals potential limitations of the majority of previous studies which have investigated the variability/stability of children's classroom experiences, as they have ignored the possibility of autocorrelation in their estimations.

## 11. Implications

Initially, this study was undertaken in order to equally provide the following: the most developmentally appropriate study of person characteristics that influence child development by affecting everyday interactions, and an understanding of the extent of rater effects of those observed interactions. Indeed, both are important lines of inquiry.

Children's developmental competencies emerge in the context of participation in increasingly complex, reciprocal processes over time, and serve as proximal processes, which take place in *microtime* (Bronfenbrenner & Morris, 2006), or in the moment-to-moment experiences of children. Thus, proximal processes (teacher-child Emotional Support interactions in the classroom context) which are actually experienced by children were modeled in a way which allows us to more appropriately align our analytical methods with developmental theory.

Meyer (1993) has noted that the frequency and nature of teacher-child interactions, especially in regards to instructional settings, are potent indicators of the "value added" to children's achievement as a function of attending kindergarten, which further points to the contribution of this study (the data for which were obtained from instructional classroom settings). Furthermore, various studies have shown that variation in teaching quality interactions and behavior is related specifically to student performance on cognitive, achievement, and motivational factors (Bogner et al., 2002; Brophy &

Good, 1986; Dolezal et al., 2003; Meyer, 1993; Stipek, 1988), and these results have been confirmed by NICHD SEECYD as connected to student's learning engagement and literacy skills (NICHD ECCRN, 2003). Important next steps in terms of identifying factors associated with stability in classroom quality should be taken, in order to develop practices and policies which serve to enhance children's experiences in kindergarten and later. Along these lines, the potential influence of raters on the stability/variability of classroom experiences, in previous and future studies needs to be investigated.

In carrying out this study, the role of raters in the obtained stability estimates was observed to be extensive. As a result, what began as an equal investigation of teacher and rater factors, quickly transformed into a rater centered study. This study reveals that over repeated ratings, raters begin to rate teachers more consistently on Emotional Support and its individual dimensions, such that previous ratings influence subsequent ratings. These results indicate that autocorrelation is not only a significant aspect of CLASS ratings, but rather a large part of these ratings. This shows us that rater effects not only exist, but are a large part of ratings of teachers. While there has been some investigation in regards to quantifying the effects which raters might exhibit on ratings of teachers (e.g., Casabianca et al., 2012; Casabianca & Lockwood, 2013) more work is necessary in order to better understand the nature and extent of rater influence on ratings. With this knowledge, we can construct observational tools and measures which combat the potential for rater effects. For instance, perhaps in order to mitigate the occurrence of non-independent ratings, a maximal time threshold should be observed in designing rating systems. Again, further research as to the nature of ratings and how they are made is necessary, as failing

to account for this type of serial dependency could be just as concerning as omitting a relevant variable. In addition, the results and implications of studies which include rater observations of classrooms, yet have not taken into account the potential of autocorrelation, should be reassessed.

## **12. Limitations**

The results of this study might have been confounded by various factors, one being that there was some variability in the amount of data available for each teacher, ranging from as little as 9 minutes of observation to as much as 63 minutes of observation. Because there was more information available about some teacher's Emotional Support than others means that some teachers had more influence on the observed results. In order to minimize the possible effects associated with this inequality in the amount of data available for teachers, analysis could perhaps in the future only include the maximum amount of data points that allows for all teachers to be equally represented. Another solution might be to only conduct an analysis on the first 15 to 20 minutes of teacher Emotional Support ratings. This might be especially appropriate as the CLASS cycles of observation are between 15 and 20 minutes long, and might allow us to speak more to how to limit rater effects associated with real classroom assessment measures. Specifically, this can help us to get an understanding of the maximal time threshold which should be observed in designing rating systems.

Another limitation associated with this study is the fact that the CLASS was not used in the traditional manner, where raters take 10 minutes to rate teachers every 15 to 20 minutes. Rather, raters rated teachers on Emotional Support once every minute. It is thus difficult to assess the reliability associated with these ratings. In order to be able to

declare that these ratings are indeed reliable, a new training procedure should be devised, where raters need to be at least 80% reliable with Master coders, on this minute-to-minute rating.

This study might also be limited in that it only includes investigation of one rater's ratings. This might present issues such as reduced generalizability of the rater effects observed here. Indeed, while having more than one rater observe a teacher at a given time is encouraged in CLASS ratings, it is not a requirement, and one rater often rates one teacher, from the start to the end of an observation. Still, the potential effects of only using one rater's ratings in this study should be investigated.



### **13. Future Directions**

This provides insight into the extent of rater effects which might manifest over time, such as non-independent ratings (a rater might make up their mind about the level of teacher Emotional Support which affects subsequent ratings and results in an autoregressive lag component). Future directions involving possible rater effects in carrying out ratings like those of Emotional Support might be specifically to investigate how long it takes a rater to make up his mind about a rating. This might aid in determining the maximal time threshold which should be observed in designing rating systems, in order to limit rater effects including autocorrelation in ratings, as much as is possible.

This study shows the potential of novel statistical techniques to produce novel findings. Specifically, this research shows the necessity of investigating the existence of autocorrelation in studies which include rater observation ratings. Indeed, previously obtained variability/stability estimates might be a function of autoregressive or moving average terms. Previous studies that included rater observations which failed to account for such serial dependency should be re-evaluated and carried out with the inclusion of this previous omission. Future studies which include rater observation should acknowledge and account for the potential of these rater effects. Indeed, with the recent emphasis by federal agencies on the use of rater observation in classrooms, understanding the extent of rater effects has large policy implications.

## References

- Bailey, C. S.\*, Carlson, A.G.\*, Brock, L.L., Curby, T. W., & Locasale-Crouch, J. (in revision). *Predictors of emotional support consistency among novice, mid-career and late-career teachers.*
- Barbarin, O. (2009). The relations of observed pre-K classroom quality profiles to children's academic achievement and social competence. *Early Education and Development*, 20, 346–372. doi:10.1080/10409280802581284
- Bogner, K., Raphael, L., & Pressley, M. (2002). How grade 1 teachers motivate literate activity by their students. *Scientific Studies of Reading*, 6, 135–165. doi: 10.1207/S1532799XSSR0602\_02
- Brock, L. B., & Curby, T. W. (in revision). Teacher's emotional support consistency and early adaptability: Relations to achievement, social skills, and emotional reactivity in third grade. *Elementary School Journal*.
- Brock, L. B., & Curby, T. W. (in revision). Pre-k teachers' emotional consistency and relationships with children: Relations to concurrent and later social competence and problem behaviors. *Early Education and Development*.
- Brock, L. L., Nishida, T.K., Chiong, C., Grimm, K.J., Rimm-Kaufman, S.E. (2008). Children's perceptions of the classroom environment and social and academic performance: A longitudinal analysis of the contribution of the responsive classroom approach. *Journal of School Psychology*, 46, 129–149.
- Bronfenbrenner, U., & Morris, P. (2006). The bioecological model of human development. In R. M. Lerner (Ed.), *Handbook of child psychology: Vol. 1. Theoretical models of human development* (6th ed., pp. 793–828). Hoboken, NJ: Wiley.
- Brophy, J., & Good, T. (1986). Teacher behavior and student achievement. In M. Wittrock (Ed.), *Handbook of research on teaching* (pp. 328–375). New York: Macmillan.
- Chomat-Mooney, L. I., Pianta, R. C., Hamre, B. K., Mashburn, A. J., Luckner, A. E., Grimm, K. J., Downer, J. T. (2008). *A practical guide for conducting classroom*

- observations: A summary of issues and evidence for researchers*. New York: William T. Grant Foundation.
- Creemers, B. P. M. (1994). *The effective classroom*. London: Cassell.
- Curby, T.W., LoCasale-Crouch, J., Konold, T.R., Pianta, R., Howes, C., Burchinal, M., Bryant, D., Clifford, R., Early, D., & Barbarin, O. (2009). The relations of observed pre-k classrooms quality profiles to children's academic achievement and social competence. *Early Education and Development*, 20, 346-372. doi:10.1080/10409280802581284
- Curby, T. W., Brock, L., & Hamre, B. (2013). Teachers' emotional support consistency predicts children's achievement gains and social skills. *Early Education and Development*, 24, 292–309. doi:10.1080/10409289.2012.665760
- Curby, T. W., Grimm, K. J., & Pianta, R. C. (2010). Stability and change in early childhood classroom interactions during the first two hours of a day. *Early Childhood Research Quarterly*, 25, 373–384. doi:10.1016/j.ecresq.2010.02.004
- Curby, T. W., Stuhlman, M., Grimm, K., Mashburn, A., Chomat-Mooney, L., Downer, J., Hamre, B.K., & Pianta, R.C. (2011). Within-day variability in the quality of classroom interactions during third and fifth grade: Implications for children's experiences and conducting classroom observations. *Elementary School Journal*, 112, 16-37. doi:10.1086/660682
- Darling-Hammond, L. (1997). *Doing what matters most: Investing in quality teaching*. New York: National Committee on Teaching and America's Future.
- Dolezal, S. E., Welsh, L. M., Pressley, M., & Vincent, M. M. (2003). How nine third-grade teachers motivate student academic engagement. *Elementary School Journal*, 103, 239–269. doi:10.1086/499725
- Doll, B., Zucker, S., & Brehm, K. (2004). *Resilient classrooms: Creating healthy environments for learning*. New York : Guilford Press.
- Early, D., Barbarin, O., Bryant, D., Burchinal, M., Chang, F., Clifford, R., . . . Weaver, W. (2005). *Pre-K in eleven states: NCEDL's Multi-State Study of Pre-K and Study of State-Wide Early Education Programs (SWEEP)*.
- Eccles, J., & Gootman, J. A. (2002). *Community programs to promote youth development*. Washington, DC: National Academies Press.
- Feldman, S. (2000, January). 220,000 teachers a year: Putting first-class educators in every classroom. Speech to the Economic Club of Detroit, Detroit, MI.

- Gottman, John. (1981). *Time Series Analysis: A Comprehensive Introduction for Social Scientists*. London: Cambridge University Press.
- Hamre, B. K., & Pianta, R. C. (2005). Can instructional and emotional support in the first grade make a difference for children at risk of school failure? *Child Development*, 76, 949–967.
- Hamre, B. K., & Pianta, R. C. (2007). Learning opportunities in preschool and early elementary classrooms. In R. C. Pianta, M. J. Cox, & K. Snow (Eds.), *School readiness and the transition to kindergarten* (pp. 49–84). Baltimore: Brookes.
- Hamre, B. K., Pianta, R. C., Mashburn, A., & Downer, J. (2007). *Building and validating a theoretical model of classroom effects in over 4,000 early childhood and elementary classrooms*.
- Hannan, E.J., Quinn, B.G. (1980). The determination of the order of autoregression. *J. R. Statist. Society*, 190-195. doi:10.1086/660682
- Hasan, M.A., & Thaut, M.H. (2004). Statistical analysis for finger tapping with a periodic external stimulus. *Perceptual & Motor Skills*, 99, 643-661.
- Korenberg, M.J., Paarmann, L.D. (1989). Applications of fast orthogonal search: Time series and analysis and resolution of signals in noise. *Ann biomed Engineering*, 17, 219-231.
- La Paro, K.M., Pianta, R.C., & Stuhlman, M. (2004). The classroom assessment scoring system: Findings from the prekindergarten year. *The Elementary School Journal*, 104, 409–426. doi:10.1086/499760
- Mashburn, A. J., Pianta, R. C., Hamre, B. K., Downer, J. T., Barbarin, O., Bryant, D., . . . Howes, C. (2008). Measures of classroom quality in pre-kindergarten and children's development of academic, language and social skills. *Child Development*, 79, 732–749.
- Matsumura, L. C., Patthey-Chavez, G. G., Valdes, R., & Garnier, H. (2002). Teacher feedback, writing assignment quality, and third-grade students' revision in higher and lower achieving schools. *Elementary School Journal*, 103, 3–25.
- Meyer, L. A., Linn, R. L., & Hastings, C. N. (1991). Teacher stability from morning to afternoon and from year to year. *American Educational Research Journal*, 28, 825–847.
- Meyer, L. A., Wardrop, J. L., Hastings, C. N., & Linn, R. L. (1993). Effects of ability and settings on kindergarteners' reading performance. *Journal of Educational Research*, 86, 142–160.

- Morrison, F.J., & Connor, C.M. (2002). Understanding schooling effects on early literacy: A working research strategy. *Journal of School Psychology*, 40, 493–500. doi:10.1016/S0022-4405(02)00127-9
- National Center for Education Statistics. (2000). *America's kindergartners*. Washington, DC: U.S. Department of Education.
- Nau, Robert. (2005). Forecasting: Identifying the Number of Ar or MA Terms. Retrieved from <http://people.duke.edu/~rnau/411home.html>
- NICHD Early Child Care Research Network. (2002). The relation of global first grade classroom environment to structural classroom features, teacher, and student behaviors. *Elementary School Journal*, 102, 367–387. doi:10.1086/499709
- NICHD Early Child Care Research Network. (2003). Social functioning in first grade: Associations with earlier home and child care predictors and with current classroom experiences. *Child Development*, 74, 1639–1662. doi:10.1046/j.1467-8624.2003.00629.x
- NICHD Early Child Care Research Network. (2005). A day in third grade: A large-scale study of classroom quality, teacher, and student behaviors. *Elementary School Journal*, 105, 305–323. doi:10.1086/428746
- Nye, B., Konstantopoulos, S., & Hedges, L.V. (2004). How large are teacher effects? *Educational Evaluation and Policy Analysis*, 26, 237–257. doi: 10.3102/01623737026003237
- O'Connor, E., & McCartney, K. (2007). Examining teacher–child relationships and achievement as part of an ecological model of development. *American Educational Research Journal*, 44, 340–369. doi: 10.3102/0002831207302172
- Papoulis, A. *Signal Analysis*. (1984). McGraw-Hill: New York.
- Pianta, R. C. (1999). *Enhancing relationships between children and teachers*. Washington, DC: American Psychological Association. doi: 10.1037/10314-000
- Pianta, R. C., La Paro, K., Payne, C., Cox, M. J., & Bradley, R. (2002). The relation of kindergarten classroom environment to teacher, family, and school characteristics and child outcomes. *Elementary School Journal*, 102, 225–238. doi: 10.1086/499701

- Pianta, R. C., Belsky, J., Houts, R., & Morrison, F. (2007). Teaching: Opportunities to learn in America's elementary classrooms. *Science*, 315(5820), 1795. doi: 10.1126/science.1139719
- Pianta, R. C., La Paro, K. M., & Hamre, B. K. (2008). *Classroom assessment scoring system—K–3*. Baltimore: Brookes.
- Rutter, M., & Maughan, B. (2002). School effectiveness findings, 1979–2002. *Journal of School Psychology*, 40(6), 451–475.
- Stipek, D. (1988). *Motivation to learn: From theory to practice*. Boston: Allyn & Bacon.
- Stipek, D., Feiler R., Daniels, D., & Milburn, S. Effects of different instructional approaches on young children's achievement and motivation. *Child Development*, 66, 209–223.
- Stuhlman, M. W., & Pianta, R. C. (2009). Profiles of educational quality in first grade. *Elementary School Journal*, 109, 323–342. doi: 10.1086/593936
- Thompson, R., & Happold, C. (2002). The roots of school readiness in social and emotional development. *The Kauffman Early Education Exchange*, 1, 8 – 29.
- Weinstein, R. (2002). *Reaching higher: The power of expectations in schooling*. Cambridge, MA: Harvard University Press. doi: 10.1177/0165025409360304
- Zinsser, K., Bailey, C., Curby, T.W., Denham, S.A., Bassett, H.H., & Morris, C. (in press). Exploring the predictable classroom: Preschool teacher stress, emotional supportiveness, and students' social-emotional behavior in private and Head Start classrooms. Revised manuscript submitted for publication to *National Head Start Association Dialog*.
- Zinsser, K., Curby, T.W., & Winsler, A. (in revision). Variability in center-based child care quality and maternal sensitivity: Links with academic and behavior outcomes at 54 months. Manuscript in revision at *Early Childhood Development & Care*.
- Zinsser, K.\*, Shewark, E., Denham, S.A., & Curby, T.W. (under review). A mixed-method examination of preschool teacher beliefs about social emotional learning and relations to observed emotional support. *Infant and Child Development*.

## **Biography**

Noora Hamdan received her Bachelor of Arts in Political Science from James Madison University in 2011. She received her Master of Arts in Psychology from George Mason University in 2013. She will be a PhD student in Developmental Psychology at Temple University starting fall 2013.