INTER-RATER AGREEMENT OF THREE FUNCTIONAL ASSESSMENT INSTRUMENTS DEPENDING ON THE FREQUENCY AND SEVERITY OF THE TARGET BEHAVIOR

by

Shannon Scurlock

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

_____

_____

_____ Department Chairperson

_____ Dean, College of Humanities and Social Sciences

Date: May 1, 2013 Spring Semester 2013
George Mason University
Fairfax, VA

Inter-rater Agreement of Three Functional Assessment Instruments Depending on the Frequency and Severity of the Target Behavior

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

By

Shannon Scurlock
Bachelor of Arts
Arcadia University, 2010

Director: Johannes Rojahn, Professor
Department of Psychology

Spring Semester 2013
George Mason University
Fairfax, VA

Acknowledgements

A thank you goes to Chrestomathy Inc for collecting this data and then allowing it to be analyzed for this project. My committee also needs special appreciation for they each facilitated this project in their own way. Dr. Patrick McKnight, who is brilliant and always willing to help a student, gave me a few nudges in the right direction, both for statistical analyses but also in my career path. Dr. Jeff Stuewig, a wonderful professor to work with, taught me more than I could have hoped for while being the ally I needed. Finally, Dr. Johannes Rojahn, a great advisor to whom I owe my graduate career, has brought together the method with the theory, which is the only reason this could be written.

Table of Contents

List of Tables

List of Figures

Abstract

INTER-RATER RELIABILITY OF THREE FUNCTIONAL ASSESSMENT INSTRUMENTS DEPENDING ON THE FREQUENCY AND SEVERITY OF THE TARGET BEHAVIOR

Shannon Scurlock, M.A.

George Mason University, 2013

Dr. Johannes Rojahn

In applied behavior analysis functional assessment refers to the process of identifying the contingencies (functions) that maintain problem behavior. Successful behavioral intervention to decrease severe behavior problems in individuals with intellectual disability (ID) depends on our ability to accurately identify why a given individual engages in a specific behavior. The Functional Assessment for Multiple Causality (FACT), Functional Analysis Screening Tool (FAST), and Questions about Behavioral Function (QABF) are three commonly used third-party behavior rating instruments designed to evaluate the function of target behaviors, which have previously been shown to have acceptable levels of inter-rater reliability. This study examined whether inter-rater agreement was impacted by types of target behavior problems (S*elf-injurious Behavior, Stereotyped Behavior,* and *Aggressive/Destructive Behavior*), frequency or severity of the target behavior, the assessment instrument, and how these factors interact. The sample

consisted of 115 adults with ID with one or more problem behaviors. Each participant was assessed with all three instruments by two raters. Two separate univariate General Linear Model analyses were used to determine what contributes to the discrepancy in rater scores on each subscale. The predictors included the assessment instrument, type of target behavior, and either frequency or severity of the target behavior (separate models) along with all 2-way interactions. We found that severity of behavior is a significant predictor and that while frequency of behavior is not an individual predictor, it does interact with type of behavior to predict rater discrepancy scores.

Introduction

There are some behaviors which are deemed inappropriate or harmful and need to be decreased and replaced by appropriate behaviors. Examples of problem behaviors can include banging ones' own head onto surfaces, repetitive hand movements, and hitting others. In order to determine the function of these behaviors, or why the problem behavior persists, it is imperative to look at the context in which the behavior occurs. Functional assessment checklists were developed due to the inconvenience of experimental techniques, which relied on the ability of analog sessions to recreate the environment of individuals and their behaviors. These instruments allow raters who are familiar with the subject to determine the maintaining function of a given behavior by answering the items which are categorized into function subscales.

The *Functional Assessment for Multiple Causality* (FACT; Matson et al., 2003), *Functional Analysis Screening Tool* (FAST; Iwata & DeLeon, 1995), *Questions about Behavioral Function* (QABF; Matson & Vollmer, 1995) are three functional assessment instruments that have been used in clinical practice as well as in research. All three behavior rating scales have varying amounts of psychometric testing done in regards to reliability and validity. However, only the QABF has had reliability testing taking into account the frequency of the target behavior (Matson & Wilkins, 2009), which may affect reliability. This study looks to expand on the existing psychometric properties, especially

inter-rater agreement depending on frequency or severity of the target behavior, the type of target behavior, and the assessment instrument.

The FACT and QABF have stronger convergent and discriminant validity than the FAST has with either assessment (Zaja, Moore, van Ingen, & Rojahn, 2011). This may be due to the different format and subscales. The FACT and QABF separates *Attention* and *Tangible* subgroups, which the FAST combines into *Attention*. This can have many effects on the grouping of data, which would affect how well the assessments correlate, and therefore decrease its convergent validity. The FAST seems to have lower discriminant validity due to higher correlations between the subgroups then with comparison subgroups, which also may have to do with the format of the subscales. These differences need to be addressed for benefits and weaknesses of each assessment, and also to identify trends of each assessment instrument.

In addition to having different assessment instruments with varying levels of variability and different subscales to categorize the functions of target behaviors, one behavior often serves multiple functions, which may influence rater agreement (Matson et al., 2003). This means that a person whose main behavior is head banging may use the behavior to escape a task but also to gain attention, and this may be difficult to rate by two different raters who interact with the person in different contexts. Someone who works with the subject on program goals may see head banging only in regards to escaping the tasks, whereas someone who interacts with the subject at home with distractors, such as making food, may see the behavior in regards to seeking attention. Also, subscales themselves have different levels of inter-rater agreement (Zaja et al.,

2011) meaning that some functions of behaviors may be more difficult to agree upon.

Not only can multiple functions of behavior decrease reliability, especially if the raters observe the subject in different contexts, but rater characteristics can also affect how they perceive the function of behavior. Certain staff characteristics included: working hours, gender, internal attribution, and experience working with people with severe or profound intellectual disabilities (Lambrechts & Maes, 2009). This would suggest that even if the functional assessment has good inter-rater reliability, and if the behavior has only one main function, that multiple raters should still be utilized since all raters have their own characteristics which affect their perception.

Matson and Wilkins (2009) studied the reliability of the QABF dependent on the frequency of the target behavior (S*elf-injurious Behavior* and *Aggressive Behavior*). They found that high rate behaviors were reliable, but that the reliability of the QABF depended on the function being analyzed. In general, high rate behaviors were more reliably identified. It was speculated that this was due to those behaviors being more salient to the raters and give a greater opportunity to determine their function. Overall, aggression had higher inter-rater reliability than self-injurious behavior. While this study did not look at severity of the behavior, the results for frequency should theoretically transfer to severity for the same explanations of these findings. The goal of this study is to identify predictors of inter-rater discrepancy across the three main assessment instruments: QABF, FACT, and the FAST. The standard error of measurement (SEM) of each assessment instrument was used to take into account the variance of each measure.

Hypothesis 1: We hypothesize that the factor "frequency of the target behavior"

will be a significant contributor to the inter-rater discrepancy scores on assessment subscales. Specifically that as "frequency of the target behavior" increases the inter-rater agreement will increase. If this hypothesis is correct the univariate GLM analysis should reveal that the relative frequency in which target behaviors occurred will be a significant negative predictor of the "inter-rater discrepancy scores divided by SEM", regardless of the "types of target behavior" and regardless of the "assessment instrument".

Hypothesis 2: We also hypothesize that the factor "severity of the target behavior" will be a significant contributor to inter-rater discrepancy scores. Specifically, that as "severity of the target behavior" increases the inter-rater agreement will increase. In support of this hypothesis, we should expect "severity of the target behavior" to be a significant negative predictor in the univariate GLM analysis.

Hypothesis 3: We anticipate that the factor "types of target behavior" will have different inter-rater discrepancy scores. This will be tested using a priori contrasts between the inter-rater discrepancy scores for *Aggressive/Destructive Behavior* and the inter-rater discrepancy scores for *Stereotyped Behavior*, and between the inter-rater discrepancy scores for *Aggressive/Destructive Behavior* and those for *SIB.* Since this is a planned contrast it will be tested even if "type of target behavior" is not a significant predictor of "rater discrepancy scores divided by SEM" in the model.

Hypothesis 4: It is anticipated that the factor "assessment instrument" will be a significant predictor in each univariate GLM, meaning that it contributes a significant amount of variance in "rater-discrepancy scores divided by SEM" in addition to the predictors "type of target behavior", and target "behavior characteristic" (frequency and

severity of the target behavior).

Secondary hypotheses: All 2-way interactions are included in the model in order to analyze how each predictor interacts with other predictors. Specifically, the interest is in how the target “behavior characteristic” (frequency and severity of behavior) interact with “type of target behavior” and “assessment instrument” in each of their respective models to predict “rater-discrepancy scores divided by SEM”.

Method

Participants

Participants were recruited from a day program for adults with intellectual disabilities in Minnesota. One hundred sixteen adults with varied level of intellectual disability participated (17 had mild, 28 moderate, 41 severe, and 30 profound ID). Age ranged from 23 to 75, with 82 males and 34 females. For each participant one dominant target behavior, based on scores for frequency and severity of behaviors on the BPI-01, was chosen for functional assessment. Thirty-three participants had *Self-Injurious Behavior*, 25 had *Stereotyped Behavior*, and 58 had *Aggressive Behavior*. Demographics (see Table 1) were collected for all participants.

Materials

Functional Analysis Screening Tool (FAST; Iwata & DeLeon, 1995).

The FAST has four subscales (*Social Attention, Social Escape, Automatic Sensory, and Automatic Pain*). It has sixteen items that are geared to categorize the function of the behavior into one of the subscales. The FAST was created to be used as a screening tool and to aid in forming accurate intervention programs for target behaviors.

The FAST was demonstrated to have fair to excellent test-retest reliability (0.69-0.71), poor to good inter-reliability (0.48-0.71), and poor internal consistency (ranging from 0.05 to 0.77 with a mean of 0.39) (Zaja et al., 2011). They also found questionable

convergent and discriminant validity due to lower correlations with the other assessment instruments and the FAST subscales often correlated higher with other FAST subscales then with similar subscales on a different assessment.

Questions about Behavioral Function (QABF; Matson & Vollmer, 1995)

The QABF has five subscales that are rated on a 4-point Likert scale ranging from 0 (never) to 3 (often) The five subscales are *Attention* (draws the attention of others), *Escape* (to escape social and nonsocial demands), *Non-social* (self-stimulation), *Physical* (relieve physical discomfort, and *Tangible* (access preferred items). The rater uses the scales to rate the frequency of the behavior in regards to where, why, and when the behavior occurs. The items are then scored for how many times each subgroup was selected and the frequency of endorsement for each subgroup was tallied which then gives the function of the target behavior and the severity scores.

The psychometric properties of the QABF have been studied by both the authors and by independent researchers due to its popularity in the field. It has been suggested that the QABF has good convergent validity because it has a high correlation with the Motivational Assessment Scale (MAS; Shogren & Rojahn, 2003), which was the first widely used functional assessment instrument. The inter-rater reliability had been found to be rather low, 0.62 (Shogren & Rojahn, 2003) and ranging from 0.63 to 0.68 (Zaja et al., 2010). The test-retest reliability was found to be strong, ranging from 0.81 to 0.82 (Zaja et al., 2010) and 0.62 to 0.93 (Shogren & Rojahn, 2003). The subscales have also been tested for subscale internal consistency, ranging from 0.89 to 0.96 (Zaja et al., 2010) and 0.82 to 0.88 (Shogren & Rojahn, 2003). Interestingly, it was found that the QABF

reliability was higher when the target behavior had a single maintaining function in comparison to when the target behavior was maintained by multiple functions (Matson & Boisjoli, 2007).

Functional Assessment for Multiple Causality (FACT; Matson et al., 2003).

The FACT has five subscales (*Attention, Escape, Nonsocial, Physical, and Tangible).* The FACT consists of forced-choice questions with three options for each question, two options for the function of the behavior, each of which belongs to a particular subscale, or "neither". Each subscale is paired with each other subscale and each subscale is an option fourteen times. The rater must choose which option is the best function of the behavior. This is considered an endorsement of the function. The amount of times a behavior is endorsed creates a frequency hierarchy, which the subscale endorsed the most as the main function of the behavior.

The psychometric properties of the FACT are limited to only a few studies. Matson et al. (2003) found the FACT to have high internal consistency across subscales (0.94-0.95). The FACT was also determined to have acceptable reliability across all three categories of the BPI-01 (Zaja et al., 2011). This included inter-rater reliability (0.65 to 0.78), test-retest reliability (0.86 to 0.87), and strong internal consistency (0.92 to 0.96). It is also thought to have strong concurrent validity since it has the same subscales as the MAS (Matson et al., 2003).

Behavior Problems Inventory (BPI-01; Rojahn et al., 2001)

The BPI-01 assesses forty-nine target behaviors which comprise three separate categories. The *Self-Injurious Behavior (SIB)* subscale has 14 items, the *Stereotyped*

Behavior subscale has 24 items, and the *Aggressive/Destructive Behavior* subscale has 11 items. The rating for each item consists of a five point frequency scale (0 = never, 1 = monthly, 2 = weekly, 3 = daily, 4 = hourly) and a four-point severity scale (0 = no problem, 1 = slight problem, 2 = moderate problem, and 3 = severe problem). Each target behavior, which occurred at least once in the previous two months before assessment, was rated for frequency and severity, and then an overall score for each category of behavior was totaled for a subscale raw score. The frequency and severity of the behavior remain separate scores. Within each behavior category, there is an item labeled "other" for any clinically relevant behaviors that are not listed, these were excluded from this study.

The BPI-01 has been suggested to have acceptable reliability, although some items have lower reliability due to low endorsements of the behavior. It was discussed that perhaps the behaviors with low endorsements should be removed to increase reliability, but this would decrease construct validity (González et al., 2009). When compared with the Inventory for Client and Agency Planning (ICAP), the BPI-01 showed strong convergent validity and concurrent validity within the subscales (Van Ingen, Moore, Zaja, & Rojahn, 2010). They also found high inter-rater reliability for each subscale and stable test-retest reliability, however, internal consistency ranged from poor to excellent. This range in internal consistency may be due to the frequency of items within each subgroup. Lundqvist (2011) studied the generalizability of the BPI-01 and found that the three subscales were highly similar constructs across culture and is applicable with varying living arrangements, diagnoses, ages, and mental functioning.

Procedure

Raters were senior staff members from the adult day program who had worked with the participants for several years. They received training on the administration of the BPI-01 and the three functional assessment instruments. For each participant, two staff members who were most familiar with the individual were selected to independently complete the BPI-01 and all three functional assessments. The frequency scores on the BPI-01 were used to identify the target behavior of each participant. If multiple behaviors had an equivalent frequency score, the raters used the highest severity score. Raters then independently administered each of the three functional assessments for the identified target behavior. A second assessment using the same procedure was repeated approximately eight weeks after the first assessment.

Results

To allow for both categorical and continuous variables univariate General Linear Model (GLM) analyses were conducted separately, one for each "behavior characteristic". The dependent variable (DV) was the inter-rater discrepancy scores divided by the standard error of measurement (SEM) of the assessment instrument. We decided to divide the inter-rater discrepancy scores by the SEM in order to account for the overall reliability of each of the three assessment instruments. From here on we will refer to the inter-rater discrepancy scores divided by the SEM of the assessment instrument as "DV".

The predictors were the (a) three "types of target behavior" (three levels), (b) the functional "assessment instruments" (three levels), and (c) the target behavior characteristics "frequency" vs. "severity" as determined by BPI-01 ratings (two levels). Descriptives for each group are compiled in Table 2. Due to the collinearity of the target behavior characteristics "frequency" and "severity," they were entered into two separate GLM models. First we centered the frequency and the severity scores and then dummy coded the predictors "types of target behavior" and "assessment instrument."

In the next step we computed *a priori* contrasts of the mean DV based on the three "types of target behavior", collapsed across the subscales of the assessment instruments. Since Matson and Wilkins (2009) found that aggressive behavior had better

overall inter-rater agreement on functional property ratings than SIB on the QABF, we compared the mean "DV" of the A*ggressive/Destructive* subscale score with the mean "DV" of *SIB* and the mean DV of *Stereotyped* behavior across all assessment instruments. We found that across all three assessment instruments there were no differences between "types of target behavior" in the "frequency" model (p = .393) or the "severity" model (p = .891; see Table 3).

Frequency of behavior model

The Levene's Test was significant (p = .024) meaning there was heterogeneity of variance across groups. The overall Univariate GLM model was significant (p = .002, R^2= .020), indicating that the predictors in this model explained a significant amount of variance in DV. While there were no significant main effects, there were significant interactions between the "type of target behavior" and the "frequency of the target behavior" (p = .001) as well as between "type of behavior" and "assessment instrument" (p = .013; see Table 4). The parameter estimates for the "type of target behavior x frequency of target behavior" show that the relationship between "frequency of the target behavior" and the "DV" is higher when the "type of behavior" is *SIB* than *aggressive/destructive* (*beta* = 0.05, p = .000; see Table 4). For the interaction between "type of target behavior" and "assessment instrument" the relationship between the "type of target behavior" and the "DV" is higher when the "assessment instrument" is the *QABF* and the "type of target behavior" is *Stereotypical* (*beta* = 0.480, p = .013)

Severity of behavior model

The Levene's test was significant (p = .014) meaning there was heterogeneity of variance across groups. The omnibus Univariate GLM model was significant (p = .012, R2=.017), which means that the predictors in this model explained a significant amount of variance in the "DV". However, the only main effect was "severity of the target behavior" (p = .004), and no significant interaction effects, except that the interaction "type of behavior" by "severity of behavior" was approaching significance (p = .066; see Table 5). The parameter estimates show that when the "type of target behavior" is *SIB* the relationship between "severity of the target behavior" and the "DV" is higher than when it is *aggressive/destructive* behavior (*beta* = 0.021, p = .000; see Table 5). Since no categorical variables were significant predictors, no post hoc analyses were completed to test level differences; however there were interesting trends for "type of target behaviors" across "assessment instruments" (see Figure 1).

Discussion

The first hypothesis which predicted that the "frequency of the target behavior" would be a significant predictor of the "DV" was not supported. This can be interpreted that frequency of behavior does not explain a significant amount of variance in rater agreement.

The second hypothesis which predicted "severity of the target behavior" would be a significant predictor of the "DV" was supported (p = .004), particularly that as "severity of the target behavior" increases, the rater agreement decreases (*beta* = .003, p = .766), which was not the direction of the hypothesis. There are no significant interactions between the predictors, although "type of target behavior" and "severity of the target behavior" are approaching significance (p = .066) with *SIB* being significantly different from *aggressive/destructive.*

The third hypothesis, which predicts that "type of target behavior" will be a significant predictor of the "DV", was not supported in either model. Specifically it was predicted that *Aggressive/Destructive* behavior would have lower score discrepancies, or better rater agreement, than *SIB* and *Stereotyped* behaviors as an extension of the results from Matson and Wilkins (2009). This was found using *a priori* contrasts, with the only significant difference between levels was on the QABF with *Stereotyped* behavior higher than *Aggressive/Destructive* behavior (see Figure 1).

The fourth hypothesis, which predicts "assessment instrument" will be a significant predictor of the "DV", was not supported in either model. There was no specific directional hypothesis to allow an *a priori* contrast and due to nonsignificance no post hoc analysis was computed. However, there are interesting trends to note (see Figure 2), specifically that the *QABF* shows a different pattern for discrepancy scores based on the types of behavior, since *SIB* has the lowest discrepancy scores and *Stereotypical* becomes the highest. This may suggest that on the *QABF*, *SIB* may have greater inter-rater agreement versus being the lowest relative to the other types of behavior on the other instruments.

The secondary hypotheses, which predicted interactions between the individual predictors, were supported in the model including "frequency of the target behavior". The interaction between "type of target behavior" and "frequency of the target behavior" is significant (p = .001) which means that the relationship between the "frequency of the target behavior" and the "DV" depends on the "type of target behavior". The type of target behavior that has a significantly different slope from *Aggressive* is *SIB*, meaning that the relationship between "frequency of the target behavior" and the "DV" is higher when the "type of target behavior" is *SIB* versus *Aggressive* (*beta* = .050; p = .000).

There is also an interaction between "assessment" and "type of target behavior" (p =.013) when "frequency of the target behavior" is included as a predictor, which means that the relationship between "type of target behavior" and the "DV" changes dependent on the level of "assessment instrument". Specifically, when the "assessment" is the *QABF* and the "type of target behavior" is *Stereotypical*, the relationship with the "DV"

is significantly higher than *Aggressive* on the *QABF* as well as *Stereotypical* on the FACT (*beta* = .480, *p* = .013). This shows that *Stereotypical* behaviors have higher discrepancy scores on the *QABF* than on the other assessment instruments and higher than any other type of behavior on the *QABF* (see Figure 1).

The purpose of this study, to replicate and expand upon previous research, was only partially successful. Matson and Wilkins (2009) found that *aggressive* behavior had better overall inter-rater agreement on functional property ratings than *SIB* on the *QABF* and that inter-rater reliability was greater with high frequency behaviors. In this study, the "DV" was changed to be the discrepancy score between raters divided by the SEM of the assessment instrument to account for instrument reliability We attempted to extend the previous results beyond the *QABF,* to include the *FAST* and *FACT* while also including *stereotypical* behavior. We did not replicate the original findings since *aggression* and *SIB* were not significantly different on the *QABF,* although the trend did occur on the other two assessment instruments. In regards to the inclusion of *stereotypical* behavior, it had the lowest discrepancy score on each assessment instrument compared to the other types of behavior. We also hypothesized that the results for frequency of behavior could be applied to severity of behavior. With this "DV", frequency of the target behavior was not a significant predictor, but severity of the target behavior was significant. This means that we failed to replicate the results for frequency of behavior, even though we extended the hypothesis to severity of the target behavior.

The main limitation of this study is that the data is taken from a program which has unequal endorsements for types of behavior. While this is to be expected, the effect

of the unequal group sizes and unequal variances can be reduced with greater overall sample sizes, where a small difference has less impact. The results attempt to replicate and expand upon previous research, to which it is partially successful, but this also should impress the importance of continued research to investigate what factors influence the reliability of these measures. If particular strengths and weaknesses of each assessment can be determined then these functional assessment instruments could become more efficient tools with more accuracy and a reduced need for multiple raters. Since not all of the variance was explained with these models, perhaps rater characteristics (Lambrechts & Maes, 2009) or multiple functions of the target behavior (Matson et al., 2003) can explain more variance in rater scores.

Most importantly this study identifies the relatively small discrepancies in rater scores, but also how these discrepancies can be accounted for. The interactions found in this study show that while the frequency of a target behavior may not directly impact inter-rater agreement, it can interact with other factors which form patterns in discrepancies. More data should be compiled to use analyses with greater power to detect these interactions. Once we understand how these factors, as well as multiple functions and rater characteristics, interact then we may have even more efficient instruments to identify functions of problem behaviors, which will allow for better interventions.

Table 1.
Demographics of Participants

	Groups	Value Label	N
BPI_Categories	1	SIB	32
	2	Stereotypical	25
	3	Aggressive/Destructive	58
Gender	1	Male	82
	2	Female	34
Intellectual Disability	1	Mild	17
	2	Moderate	28
	3	Severe	41
	4	Profound	30

Table 2.

Mean Rater-Discrepancy Scores Divided by SEM on each Assessment Instrument

Instrument	Behavior	Mean	SD	Mean Rater-Discrepancy Score/SEM
FAST	Total	1.91	1.14	1.15
	SIB	2.11	1.16	1.25
	Stereotypical	1.97	1.16	0.95
	Aggressive	1.77	1.10	1.17
QABF	Total	2.89	1.68	1.11
	SIB	3.27	1.64	0.99
	Stereotypical	2.75	1.64	1.30
	Aggressive	2.74	1.69	1.09
FACT	Total	4.43	4.49	1.02
	SIB	4.63	4.43	1.05
	Stereotypical	4.00	4.34	.94
	Aggressive	4.51	4.59	1.03

Table 3.

A Priori Contrast between Types of Behavior on Rater-Discrepancy Score/SEM

Model	Subscale Scores Compared to Aggressive/ Destructive Behavior	Estimate-Hypothesis[1]	*p*
Frequency	SIB	-0.028	.687
	Stereotypic Behavior	0.026	.854
Severity	SIB	-0.038	.593
	Stereotypic Behavior	.112	.276

[1] The hypothesis value = 0 since the null hypothesis is that there is no difference between groups. The Estimate is the difference between the group means. The difference between these values is the alternate hypothesis.

Table 4.
Between-Subjects Results with Frequency as a Predictor of Inter-Rater Discrepancy Score Divided by SEM

Source	*Unstandardized b* Parameter Estimates	*p*	Estimated η^2
Corrected model	--	.002	.020
Intercept	--	.000	.463
	1.060	.000	.123
Assessment Instrument	--	.181	.002
FAST	0.122	.250	.001
QABF	0.011	.914	.000
FACT	--	--	--
Type of Target Behavior	--	.952	.000
SIB	-0.025	.831	.000
Stereotypical	-0.115	.517	.000
Aggressive/Destructive	--	--	--
Centered Frequency of Behavior	--	.352	.001
	-0.013	.077	.002
Type of Behavior x Frequency of Behavior	--	.001	.008
SIB x Frequency of Behavior	0.050	.000	.008
Stereotypical x Frequency of Behavior	0.014	.523	.000
Aggressive/Destructive x Frequency of Behavior	--	--	--
Assessment Instrument x Frequency of Behavior	--	.313	.001
FAST x Frequency of Behavior	0.009	.380	.000
QABF x Frequency of Behavior	0.024	.019	.003
FACT x Frequency of Behavior	--	--	--
Assessment Instrument x Type of Behavior	--	.013	.008
FAST x SIB	0.071	.685	.000
FAST x Stereotypical	-0.056	.786	.000
FAST x Aggressive/Destructive	--	--	--

QABF x SIB	-0.081	.621	.000
QABF x Stereotypical	0.480	.013	.004
QABF x Aggressive/Destructive	--	--	--
FACT x SIB	--	--	--
FACT x Stereotypical	--	--	--
FACT x Aggressive/Destructive	--	--	--

Table 5.
Between-Subjects Results with Severity as a Predictor of Inter-Rater Discrepancy Score Divided by SEM

Source Parameter Estimates	*Unstandardized b*	*p*	Estimated η^2
Corrected model	--	.012	.017
Intercept	--	.000	.462
	1.030	.000	.120
Assessment Instrument	--	.182	.002
FAST	0.137	.189	.001
QABF	0.063	.520	.000
FACT	--	--	--
Type of Behavior	--	.952	.000
SIB	-0.016	.890	.000
Stereotypical	-0.052	.727	.000
Aggressive/Destructive	--	--	--
Centered Severity of Behavior	--	.004	.005
	0.003	.766	.000
Type of Behavior x Severity of Behavior	--	.066	.008
SIB x Severity of Behavior	0.021	.000	.002
Stereotypical x Severity of Behavior	0.034	.078	.002
Aggressive/Destructive x Severity of Behavior	--	--	--
Assessment Instrument x Severity of Behavior	--	.261	.002
FAST x Severity of Behavior	0.005	.727	.000
QABF x Severity of Behavior	-0.004	.753	.000
FACT x Severity of Behavior	--	--	--
Assessment Instrument x Type of Behavior	--	.129	.004
FAST x SIB	0.053	.763	.000
FAST x Stereotypical	-0.103	.605	.000
FAST x Aggressive/Destructive	--	--	--

QABF x SIB	-0.118	.473	.000
QABF x Stereotypical	0.285	.130	.001
QABF x Aggressive/Destructive	--	--	--
FACT x SIB	--	--	--
FACT x Stereotypical	--	--	--
FACT x Aggressive/Destructive	--	--	--

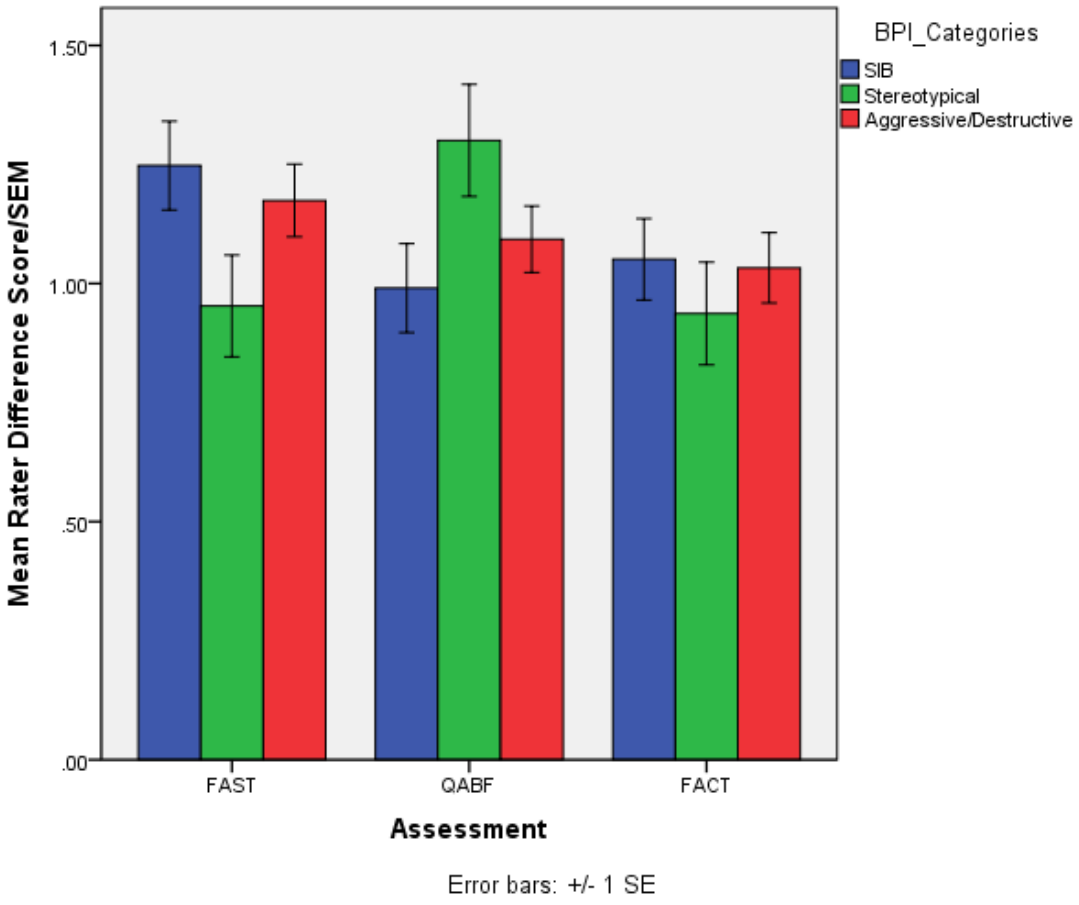

Figure 1.

Mean Rater-discrepancy Scores Divided by SEM for Types of Behavior by Assessment

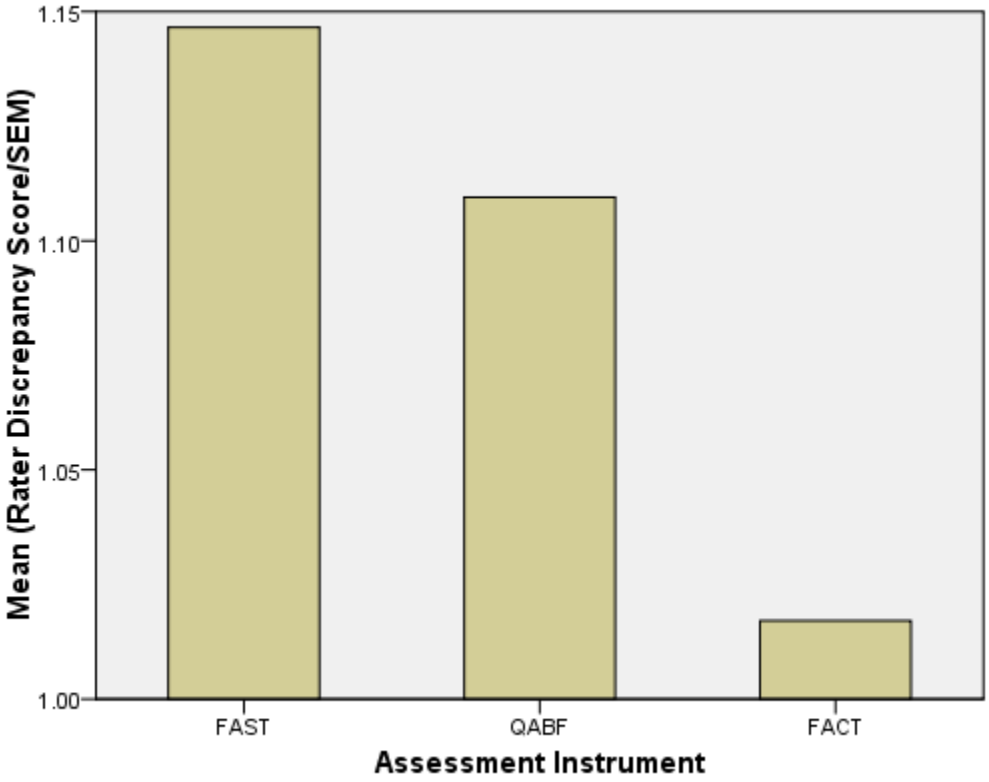

Figure 2.

Mean Rater Discrepancy Score Divided by SEM for each Assessment Instrument

References

References

Boisjoli, J. A. (2007). Multiple versus single maintaining factors of challenging behaviours as assessed by the QABF for adults with intellectual disabilities. *Journal of Intellectual and Developmental Disability*, *32*(1), 39-44. doi:10.1080/13668250601184689

González, M. L., Dixon, D. R., Rojahn, J., Esbensen, A. J., Matson, J. L., Terlonge, C., & Smith, K. R. (2009). The Behavior Problems Inventory: Reliability and factor validity in institutionalized adults with intellectual disabilities. *Journal of Applied Research in Intellectual Disabilities*, *23*(3), 223-235. doi:10.1111/j.1468-3148.2008.00429.x

Iwata, B. A., & DeLeon, I. G. (1995). *The functional analysis screening tool* (*FAST*). Unpublished manuscript, University of Florida.

Lambrechts, G., & Maes, B. (2009). Analysis of staff reports on the frequency of challenging behaviour in people with severe or profound intellectual disabilities. *Research in Developmental Disabilities*, *30*(5), 863-872. doi:10.1016/j.ridd.2008.12.004

Lundqvist, L. (2011). Psychometric properties and factor structure of the Behavior Problems Inventory (BPI-01) in a Swedish community population of adults with intellectual disability. *research in Developmental Disabilities, 32,* 2295-2303. doi: 10.1016/j.ridd.2011.07.037

Matson, J. L., Kuhn, D. E., Dixon, D. R., Mayville, S. B., Laud, R. B., Cooper, C. L., ….Matson, M. L. (2003). The development and factor structure of the Functional Assessment for multiple CausaliTy (FACT). *Research In Developmental Disabilities*, *24*(6), 485-495. doi:10.1016/j.ridd.2003.07.001

Matson, J. L., & Vollmer, T. R. (1995). *The Questions about Behavioral Function (QABF) User's Guide,* Baton Rouge, LA: Scientific Publishers.

Matson, J.L., & Wilkins, J. (2009). Factors associated with the Questions About Behavior Function for functional assessment of low and high rate challenging behaviors in

adults with intellectual disability. *Behavior Modification, 33*(2), 207-219. doi:10.1177/0145445508320342

Paclawskyj, T. R., Matson, J. L., Rush, K. S., Smalls, Y., & Vollmer, T. R. (2000). Questions About Behavioral Function (QABF): A behavioral checklist for functional assessment of aberrant behavior. *Research In Developmental Disabilities, 21*(3), 223-229. doi:10.1016/S0891-4222(00)00036-6

Rojahn J., Matson J. L., Lott D., Esbensen A. J. & Smalls Y. (2001) The Behavior Problems Inventory: An instrument for the assessment of self-injury, stereotyped behavior and aggression/destruction in individuals with developmental disabilities. *Journal of Autism and Developmental Disorders, 31*, 577–588.

Shogren, K. A., & Rojahn, J. (2003). Convergent reliability and validity of the Questions About Behavioral Function and the Motivation Assessment Scale: A replication study. *Journal of Developmental and Physical Disabilities*, *15*(4), 367-375. doi:10.1023/A:1026314316977

Van Ingen, D. J., Moore, L. L., Zaja, R. H., & Rojahn, J. (2010). The Behavior Problems Inventory (BPI-01) in community-based adults with intellectual disabilities: Reliability and concurrent validity vis-à-vis the Inventory for Client and Agency Planning (ICAP). *Research in Developmental Disabilities*, *31*(1), 97-107. doi:10.1016/j.ridd.2009.08.004

Zaja, R. H., Moore, L., van Ingen, D. J., & Rojahn, J. (2011). Psychometric comparison of the functional assessment instruments QABF, FACT and FAST for self injurious, stereotypic and aggressive/destructive behaviour. *Journal of Applied Research in Intellectual Disabilities*, *24*(1), 18-28. doi:10.1111/j.1468-3148.2010.00569.x

Curriculum Vitae

Shannon Scurlock received her Bachelor of Arts in Psychology from Arcadia University in 2010. After finishing her Master of Arts in Psychology with a concentration in Applied Developmental Psychology in 2013, she will take coursework toward her BCBA to become an ABA Behavioral Consultant.