Summer 2019

Dependability of Two Group Observation Methods Across Rater and Time

Kayla E. Bates-Brantley
University of Southern Mississippi

DEPENDABILITY OF TWO GROUP OBSERVATION METHODS ACROSS RATER AND TIME

by

Kayla Elizabeth Bates-Brantley

A Dissertation
Submitted to the Graduate School,
the College of Education and Human Sciences
and the School of Psychology
at The University of Southern Mississippi
in Partial Fulfillment of the Requirements
for the Degree of Doctor of Philosophy

Approved by:

Dr. Evan H. Dart, Committee Chair
Dr. Brad A. Dufrene
Dr. Lauren E. McKinley
Dr. Keith C. Radley

Dr. Evan H. Dart Committee Chair	Dr. D. Joe Olmi Director of School	Dr. Karen S. Coats Dean of the Graduate School

August 2019

ABSTRACT

Collecting efficient and reliable behavior assessment data is often a goal for school districts and school psychologists. Unfortunately, the most accurate methods of behavior observations, systematic direct observations (SDO), can be time-intensive and often requires specific training. This often minimizes the number of trained professional available for observation procedures. Planned activity check (PAC), a variation of momentary time sampling, has the potential to combine the accuracy of SDO with efficiency. However, few studies have evaluated the psychometric principals of PAC. The current study sought to evaluate the reliability and dependability of PAC by comparing PAC to an individual-fixed (I-F) SDO. The current study assessed group-based behaviors using two methods of SDO: individual-fixed and PAC. Observations occurred across 6 classrooms for 10 consecutive days with classroom teachers implementing PAC in conjunction with trained researchers. Results from the current study yield positive outcomes of I-F and PAC being reliable and dependable measures of group-based behaviors with an I-F G-Coefficient of .959 and PAC G-Coefficient of .889. Results also indicated that SDO and PAC can be dependable measures in addition to being efficient with follow-up dependability studies indicating SDO after two days reaches a G coefficient of .826 and PAC after four days reaches a G coefficient of .814. Finally, social validity data taken by teachers at the completion of the study indicated favorable reviews of PAC across acceptability, understanding and feasibility.

ACKNOWLEDGMENTS

This project could not have been possible without the help of so many people. When you decide to run a study that requires 5 observations per day across 10 consecutive days you must make sure you have a good team in place. I did not have a good team; I had a great team! Dr. Evan Dart, my committee chair, was the backbone of this project. His leadership and direction were imperative in every step of this dissertation. There are truly no words to thank him enough. I would also like to thank Dr. Brad Dufrene, Dr. Lauren McKinley and Dr. Keith Radley for serving as committee members. Your advice and direction were what made my project be the best it could be. Finally, I would like to thank Morgan McCargo, Ashely Murphy, Sarah Wright, Rob Derieux and Jennifer Tannehill all who contributed their time and effort to make my vision a reality.

DEDICATION

This page will be simple. I dedicate this project, this degree and this time of my life to my husband Jess Martin Brantley. You are my rock. You have been there since day-one of this journey and there was never a day when I did not feel supported. My prayer for you each day is simple, may you always keep the servant's heart that comes so natural for you. I love you and I thank you- We did it!

TABLE OF CONTENTS

LIST OF TABLES

CHAPTER I - INTRODUCTION

In the United Sates it is estimated that approximately 20% of children and adolescents meet criteria for a psychological diagnosis. Unfortunately, only half of those in need actually receive services (Burns et al. 1995). It became apparent from a public health perspective that the best way to target early intervention efforts was to send psychological services to the place that children spend a majority of their day: school (Strein, 2003). In order to service students in need, methods for monitoring and identification were incorporated into schools (McIntosh, Horner, & Sugai 2009). Assessment of student behavior has long been an interest and an essential component of the work school psychologists perform. Although the topography of how behavior is assessed has changed over time (Demaray et al., 2003; Shapiro & Heick, 2004; Christ, Riley-Tillman & Chafouleas 2009) the goal has remained constant: to provide effective treatment. In order to achieve this goal, accurate and systematic data collection are essential (Deno 2005). A survey of school psychologists practicing in the 1980's suggested that projective test were the most used methods of assessing social-emotional strengths and weaknesses of student behavior (Goh, Teslow & Fuller, 1981); however, this trend decreased during the 1990's, when school psychologist begin using more reliable and valid assessments of student behavior. Rating scales and systematic direct observation (SDO) became the most prominent form of behavior assessment in schools (Hutton, Dubes, & Muir, 1992). It was noted by school psychologist that these methods were more likely to help practitioners link behavior assessment results to appropriate recommendations and interventions (Shapiro & Heick 2004). Beginning in the 2000s, school districts around the county made a push to incorporate positive behavior

interventions and supports (PBIS) into their curriculum and discipline practices (Sugai & Horner 2009). Following the framework of multi-tiered systems of support (MTSS), the PBIS movement was created (Sugai & Horner 2006). PBIS offers schools three levels, or tiers, of intervention, progressing from low levels of support to substantial levels of support in subsequent tiers. A key characteristic of the tiered intervention approach is data-based decision making. For school systems implementing PBIS, Tier I is designed to provide school-wide support in the form of expectations and rules and evidence based instruction for all students. Data collected in Tier I incorporates universal screening assessments, office discipline records, and attendance records. Well established Tier I supports should provide enough services to address the needs of 80% of a school's student body (Bradshaw et. al. 2008). Tier II and tier III are in place to offer further assistance to any student(s) in which Tier I interventions were not enough to promote positive behavior.

Tier II does this by providing supplemental supports to groups of students while Tier III is designed to offer intensive direct interventions to individual students. For example, a common Tier II intervention is class-wide group contingencies (Anderson & Borgmeier, 2010). Class-wide group contingencies are designed to set up a single reinforcement contingency in order to modify behaviors of a group of individuals (Gresham & Gresham, 1982). Data collection for group contingencies typically is conducted by monitoring class-wide behaviors pre-intervention and during intervention implementation. This type of data collection can be conducted in a variety of ways but typically requires more targeted assessments that are often conducted by trained personal such as school psychologists rather than the classroom teachers (Christ, 2008). Although

data collection by trained personnel is common, it is often not sustainable for long periods of time, leaving classroom teachers without the resources to monitor classroom progress effectively (Christ, Riley-Tillman & Chafouleas 2009).

Data should be at the forefront of decisions that are made regarding students' behavioral performance. As part of an MTSS model, PBIS procedures that produce reliable and valid data regarding individual students' behaviors and class-wide behaviors are needed. This is especially important for a school's ability to monitor intervention effects (Christ et al., 2013). Although school districts are adopting system-wide multi-tiered systems of support, change is often slow. School districts still rely heavily on outside trained professionals to observe and assess intervention effects. Valid and reliable assessment measures that can be completed by internal school personnel (i.e., classroom teachers; aides) are critical for schools to become more self-sufficient in evaluating and supporting the effectiveness of Tier I, II and III interventions and supports (Gresham, 2004).

Ideally, school districts would always use objective data to assist with problem identification and description, intervention recommendations, progress monitoring, and outcome evaluations (Deno, 2005). Survey data taken form school psychologists and other clinicians operating in a school setting suggest that the most common methods of assessing student behavior are standardized rating scales, direct behavior rating, and SDOs (Cashel 2002). Strengths and weaknesses for each of these methods of assessing student behavior can be examined in detail.

Standardized Rating Scales

A standardized rating scale is a set of questions that are designed to assess specific constructs or attributes about individuals. They have the ability to offer data about a child's behavior as it is perceived in his or her natural environment. Rating scales operate by having a rater assign a value that closest reflects the presence of a pre-specified behavior or attribute (Pelham, Rabiano & Massetti, 2005). These values are commonly measured using a Likert-type scale in which the rater would respond by selecting a number (e.g., 1 - 5) that best aligns with the person's perceived level of performance. For example, the informant may be asked to rate the frequency (e.g., Never, Sometimes, Often, or Always) with which specific behaviors are exhibited by a child in the past 1 to 6 months (Merrell, 2000). The scale is then scored based on set criteria, which provides examiners scores based on normative samples. Normative samples allow rating scales to produce scores that can be compared to typically developing populations. This norm-referenced score gives clinicians the ability to make quick comparisons as to how the student is functioning in relation to his or her peers (Myers & Winters, 2002). Although normative data can be beneficial, they come with limitations. The term norm-referenced rating scales imply that a representative sample was obtained during the formation of the scale; however, standardizations samples cannot accurately represent all individuals and patterns of behavior (Corcoran & Fischer, 2000). Therefore, if a scale has not been normed with other persons of same race, gender, geographical location, and SES then comparisons to peers might yield inaccurate decisions due to a lack of adequate sampling (Demaray et. al 1995). These factors must be considered when interpreting scores from standardized scales (Myers & Winters 2002).

Standardized rating scales are not without their strengths, when developed with behavior-specific constructs they are often great resources for school personal to use in order to gather more information including possible targets for intervention. Standardized behavior scales can typically be completed by anyone with knowledge of the student. They do not require training prior to use as the instructions are typically printed at the top of the form. Standardized behavior scales are also not as time intensive relative to other behavior assessment methods, which can aid in the efficiency of the problem-solving process (Cashel, 2002). Although rating scales can offer information about possible targets of the referral concern, they typically do not offer situation specific information. For example, rating scales are typically not able to offer information surrounding behaviors such as antecedents or consequences. They are also not able to offer other etiological explanations surrounding a referral concern (McConaughy 1993). In addition, it is often not possible to get objective data about the frequency, duration or magnitude of behaviors from these measures.

Although a generalized summary of a child's behavior may be useful for screening, typically rating scales do not offer enough detailed information to make rapid treatment decisions. Rating scales are in a psychometric balancing act of trying to ensure the assessment is sensitive enough to detect change while also remaining specific enough as to not produce too many false positives, or incorrectly identifying clinically relevant concerns when indeed there are none. (Myers & Winters 2002). Standardized rating scales often are not sensitive enough to detect smaller changes that are due to an intervention rather than other factors such as variability of the scale (Myers & Winters 2002). A final drawback of standardized rating scales is that they rely on information

provided outside of the immediate context of behaviors occurring in the classroom. This property leaves the measures open to subjectivity and possible error. Rating scale data are almost always collected at a time and place outside of where the behaviors of interest actually occur. This relies on raters to recall past events and rate behaviors indirectly. Research has suggested that immediate ratings of behavior yield more accurate results than delayed ratings (Rush et al. 1981). Heneman and Wexley (1983) assessed how much time delay must occur before ratings of behavior became inaccurate. The study was conducted by having college students observe three short videos of an office work place. They observed the same videos one time per week for three weeks. They were then asked to rate aspects of the work place employee behaviors, using a 5-point Likert scale. Results indicated that small inaccuracies began to occur at 48 hours of delayed rating. A significant decrease in accuracy was noted for any ratings occurring after 3-weeks of watching the videos.

Aside from the inaccuracy that can occur from indirect behavior ratings, rating scales require a person to use subjective interpretations of a person's behaviors. Outside variables such as bias and personal preference can also affect the outcomes of ratings. For instance, Myers and Winters (2002) found that mothers on average tend to rate their own children's behavior as higher or more significant than fathers. Some forms of the same bias could likely be found within classroom teachers. Although rating scales can be a useful tool when used appropriately, they should not be the primary form of data being collected to monitor students' behavior change. When assessing a students' response to interventions delivered within the context of a MTSS framework, measurement forms that allow for direct ratings of a students' behavior can be useful for a variety of reasons.

Direct Behavior Rating

Direct behavior rating (DBR) utilizes the efficiency of rating scales with the accuracy of direct observation. This method of assessing behavior was first derived from the work of Chafouleas, Riley-Tillman, and McDougal (2002). DBRs operate by combing principals from standardized behavior scales and SDO (Christ, Riley-Tillman & Chafouleas, 2009). DBRs require that the behavior in question be operationalized, similar to SDO. That is, the target behavior is defined in a way that all informants are clear in understanding what does and does not qualify as an instance of the behavior being measured. This allows for informants to assess behaviors using more systematic and rigorous standards. Once standardized definitions of behaviors are established raters are expected to complete ratings of behaviors observed within a specified window of time. (Christ et al., 2009). Additionally, raters are asked to complete DBRs immediately after the observation period, theoretically reducing the latency between the occurrence of behavior and the assessment.

Chafouleas, McDougal, Riley-Tillman and Hilt (2005) investigated the accuracy of DBRs filled out by general education teachers compared to SDO data that was coded by external observers. Teachers and observers were assessing student's rate of off-task behaviors. Moderate correlations ($r = 0.67$) were found between teacher's perceptions of student behavior according to their DBR score and the direct observations conducted by the researchers. Although this study did not yield results that would indicate DBRs produce an exact replication of data provided by SDO, it did indicate that the two assessment methods produce similar data. An extension of this study was conducted by Riley-Tillman, Panahon, and Hilt (2005) using similar methods. Comparisons were made

between DBR data and direct observations. Significant results were reported with teacher's ratings correlating with direct observations for on-task (r = .81 and disruptive behavior (r = .87). Results from these studies suggest that DBR's equate well with more direct measurement of student behavior.

The previous studies outlined, all assessed the accuracy of DBRs. Although accuracy is important, it is not the only variable of concern for assessment measures, the dependability of DBRs is also of vital concern. Chafouleas and colleagues (2007) conducted a generalizability study to test the dependability of data generated from direct behavior ratings. In this study the authors compared the dependability of direct behavior ratings across raters. The data indicated that a large proportion of variance in DBR scores was attributed to the raters. This means that measurement of behavior changed depending on who was observing and rating the student. A follow-up decision study suggested that the reliability of DBRs would likely increase if 7 ratings were collected across 4-7 days. Furthermore, the decision study indicated that scores would be dependable enough to make high-stakes decision after completion of 10 total DBRs across 4-7 days.

Another generalizability study was conducted by Briesch, Chafouleas and Tillman (2010) on the dependability of DBRs across students, methods (i.e., SDO and DBR), raters, and time. No significant results were found for difference in methods, in that SDO and DBR produced similarly dependable scores; however, their data indicated that the largest contributor to score variance for DBRs again was accounted for by differences in rater while SDO had little to no variation across raters. This is intuitive, since SDO requires rigorous training and concrete operational definitions, likely reducing rater bias. On the other hand, DBR training is very brief and ratings are more subjective. Results of

this study indicate that SDO may be a more dependable method of assessment when the rater is examined. The literature base for DBRs is well established with studies across populations that address not only the question of accuracy but also reliability. Interestingly, DBR research often uses SDO as comparisons for establishing reliability and validity of DBR despite the lack of robust psychometric data supporting SDO.

Systematic Direct Observations

Systematic direct observations (SDO) are conducted by operationally defining a target behavior and then implementing either continuous or discontinuous measurement of that behavior. Continuous measures include keeping counts of the frequency with which the behavior occurs or latency between some stimulus and the behavior of interest. Although, continuous measures of behavior are highly useful, they are difficult to implement due to their rigor and limit observer flexibility. Discontinuous measures such as momentary time sampling, partial interval recording and whole interval recording require a little less observer focus and are more flexible in implementation.

Discontinuous time-sampling techniques provide a representative sample of targeted behaviors without requiring observers to conduct continuous observations (Gardenier, MacDonald, & Green 2004). Discontinuous time sampling measures allow for observers to record the occurrence of behaviors during intervals (e.g., 15sec) with results ideally representing a sample of behaviors comparable to duration recording of behavior. Of the discontinuous measures, momentary time sampling had been deemed the most accurate and feasible method of direct behavior observations (Rapp, Colby, Vollmer, Roane, Lomas & Britton, 2007).

Momentary time sampling assesses student behaviors for short intervals (i.e. 10 seconds) at which behavior is recorded as present or absent at the moment the preset time interval ends (Cooper, Heron, & Heward, 2007). For example, if the observer was assessing behavior every 15 seconds, at the end of the 15 second interval the observer would look up and mark if the targeted behavior was occurring. After the completion of the observation, observers are able to calculate the percentage of intervals during which the target behavior occurred. This direct method of assessing behaviors reduces the likelihood of error that can occur when behavior data are collected via retrospection (Christ, Riley-Tillman & Chafouleas 2009). SDO's have the flexibility to measure the frequency, rate, duration and/or latency of target behaviors (Hintze & Matthews 2004). SDO is also able to measure the behaviors of individuals or group behaviors.

Although schools and teachers are often interested in student behavior at the individual level, assessing class-wide performance and behavior is often the focus of MTSS. Specifically, for evaluating Tier I and Tier II supports, group behaviors are of interest. An early appearance of group SDO occurred in a pilot study of the good behavior game conducted by Barrish, Saunders and Wolf (1969) in which they used SDO to assess the effects the intervention had on class wide behavior. Over the next 48 years, SDO has become the gold standard of measurement for class-wide behavior assessment (Riley-Tillman et al. 2008; Repp et. al 1976; Powell et al. 1977). SDO has the ability to provide precise measurement of specific behaviors that occurs in real-time as the behavior itself is occurring (Cone, 1978). Group SDO time-sampling procedures can assess group behavior by having observers rotate in a fixed order (ie. Individual fixed) through the class or using a random rotation (Individual Random; I-R). Individual-fixed

(I-F) observations operate by having a set order in which students are observed. For observations utilizing I-F students would be observed individually for a pre-set time interval (ie.10 seconds) once behavior was recorded the observer would move to the next student with the progression continuing until each student was assessed. After one full rotation throughout the classroom, the observer would start again with the first student staying with the same rotation throughout the entire observation (McKissick, Hawkins, Lentz, Hailley, and McGuire 2010). I-R observations operate in a similar fashion, but instead of observations remaining on a fixed rotation through the group, once one full classroom observation has been complete, the order in which students are observed is shuffled ensuring students are observed in a randomized order. (Chafouleas, Sanetti, Jaffery, & Fallon, 2012). Both methods of assessing behavior yield results that are consistent with continuous methods of SDO (Briesch et al. 2014; Dart et al. 2016).

Unlike standardized behavior scales and direct behavior rating, SDO requires observers to directly record behaviors as they are occurring in the natural setting. This allows for the most precise measurement of targeted behaviors (Riley-Tillman, Chafouleas, Sassu, Chanese & Glazer 2008).

Validity and Reliability of Systematic Direct Observations

SDO's have traditionally assessed validity and reliability via two approaches: accuracy and interobserver agreement (Johnston & Pennypacker 1993). The first approach, accuracy, has been defined as how close the results from an observation align with the true value of a dimension of a behavior. That is, the actual frequency, duration, or latency with which a behavior is occurring in the natural environment as measured by continuous observation. Harrop and Daniels (1986) launched a psychometric study into

the accuracy of momentary time sampling by comparing it to partial interval recording with regards to duration of continuous observation. The study utilized computer simulated data for the comparison. Momentary time sampling was found to be accurate in measuring absolute durations of behavior. Continuing this line of research, Radley, O'Handley & Labrot (2015) compared MTS with PIR to assess which method measured the duration of social engagements most closely with continuous duration recording results of social engagement. This was completed using recorded data of five school-age children. Authors found that MTS most closely estimated the actual duration of social engagement while PIR overestimated engagement. Additionally, Rapp and colleagues (2007) compared 10-s and 20-s momentary time sampling to continuous duration recording. They did so by comparing the behaviors of four students using continuous duration recording to momentary time sampling. They found that 10 and 20-s MTS were consistent with continuous duration recording by producing almost identical recorded responses. This literature base advocates that MTS is a valid option for recording student behaviors.

Each of the previous studies used SDO to monitor the behaviors of single students or clients. However, it is not always the goal to assess only one students' behavior at a time, group-based SDO allows for observers to assess group behaviors during one observation. Dart, Radley, Briesch, Furlow and Cavell (2016) assessed the validity of group-based SDO by comparing a variety of interval-based observation techniques to a continuous duration recording of behavior. Data from this study suggested momentary time sampling utilizing an I-F or I-R performed most closely with actual behavior. These

data would suggest that momentary time sampling can provide accurate measurement of specific behaviors (Cone, 1978).

Reliability within SDO is almost always calculated using interobserver agreement (IOA). Although IOA might tell us the degree to which two raters agree on a given behavior, it does not ensure that the raters are reliably assessing behavior along other dimensions such as time or setting. IOA is limited in its ability to provide detailed information into the reliability of the observation method itself. Literature further assessing the reliability of SDO for group behaviors is limited and thus need further exploration (Dart et. al 2016).

Implementation of Systematic Direct Observation

Another limitation of many SDO schemes is how time and resource dependent they often are (Riley-Tillman et al. 2008). A standard observation typically lasts between 10 and 40 minutes with multiple observations being essential for monitoring an intervention or classroom progress (Hintze & Matthews, 2004). In 2005, it was estimated that there are 16,000 students per one school psychologist (Charvat, 2005). With such a limited number of school psychologist, it is often not feasible to devote the time needed to conduct SDO's. Logically it would make sense to shift the responsibility of behavior assessment from school psychologist to teachers, who are in the classroom on a daily basis. However, teachers already have an enormous amount of daily responsibilities they are required to perform. Therefore, any additional data collection would have to be time efficient and easy to execute in order for teachers to balance additional data collection with their already hectic classroom requirements.

Although teachers have access to observe their students on a minute by minute basis, rarely are data collected in a systematic manner in order to capture these observations (Shapiro & Heick, 2004). There are a number of reasons that teachers do not typically assess behaviors using direct observation, one of which is how time intensive SDO's can be to complete. Some SDO techniques require advanced training that most other school personnel do not have. In addiotn, asking a teacher to continue to provide instruction while also assessing student behavior using SDO is not always feasible. This logic is why SDO is more often than not conducted by an external observer (Riley-Tillman, Kalberer & Chafouleas, 2005). An alternative reason SDO is not conducted by classroom teachers is preference. Interviews and rating scales are noted in the literature as being teachers' preferred method for data collection procedures (Alberto & Troutman 1999). It is not clear however, if this preference is based on efficiency or lack of knowledge concerning SDO's.

Although teachers could develop the skills to collect direct systematic observations, the tasks are continually assigned to trained interventionist, such as school psychologist or behavior analysts (Christ et al. 2009). With a growing emphasis being placed on schools and teachers to monitor student progress, teachers need a reliable and efficient method to measure student behaviors. Currently, teachers do not have these tools and therefore a shift must be made to give teachers and other support staff the ability to apply valid and feasible tools of direct measurement to the behavior of students within their classrooms. Planned activity check, a variation of SDO, has the potential to fill this need.

Planned Activity Check

Planned Activity Check (PAC; Risley & Cataldo, 1974) is a form of SDO that utilizes intermittent monitoring of behaviors and is more time efficient than I-F or I-R SDO. PAC is a variation of momentary time sampling in which a teacher or other observer uses a head count to measure how many individuals within a group are engaged in a specific behavior at a certain point in time. The observation takes place by assessing the group of students at the end of a preset time interval (e.g., 1, 2 or 3 mins). The observer would count the number of students who are engaged in the targeted task and record that number. A percentage would be derived by dividing the total number of engaged students by total number of students and multiplying by 100. To gain a total percentage of class wide engagement an average of the percentages would be taken across all completed checks.

Doke and Risley (1972) used PAC to assess group participation in AEB by comparing two preschool activity schedules. The study's purpose was to assess which activity schedule would lend to more participation of preschool students. Two observers recorded the number of children physically present in each activity station followed by the number of students who were actively participating in the given activity. PAC occurred at 3-minute intervals. Observers were instructed to count the number of students at a given activity station and then record the number of students who were engaged appropriately with the assigned task. Observers rotated through the stations until each activity station was observed. IOA was calculated for 30% of observations resulting in a mean IOA scores of 91% (range 82-100%). It should be noted that the primary observers in this study were trained researchers and not staff members or teachers in the preschool classroom.

A second appearance of PAC occurred in a study conducted by Dyer, Schwartz and Luce (1984). This study assessed the amount of time students at a residential facility were engaged in age-appropriate functional activities. Observations were conducted by having a staff member observe each resident for "as long as it took" (pg. 252) not exceeding 10 seconds to determine if the activity in which the student was engaged was appropriate. All observers for this study were staff members at the residential facility. IOA was conducted for 25% of observations with a mean IOA of 97% (range = 81-100%). Previous use of PAC in the literature suggest that this observation method has the potential to be used reliably by personnel who are naturally in the classroom such as the classroom teacher or aide.

Despite PAC appearing in the literature during the early 70's, the psychometric properties of PAC have been understudied. In fact, Dart et al. (2016) were the first to assess the psychometric principals of this observation method. The authors utilized two studies to assess the psychometric principals. They found that PAC yielded accurate estimates of group behavior similar to true duration of behaviors. They also found that assessing behaviors in intervals of 1min, 2 min, or 3 min using PAC made little difference in the accuracy, indicating that teachers may be able to assess student behavior every 3 minutes and still achieve results that are strongly accurate when compared with true rates of behavior. It should be noted that Dart and colleagues (2016) utilized simulated data for the first study. A follow-up study was conducted within the Dart et al. (2016) article in which the authors used the exact same observation methods but applied them to a small sample of pre-recorded classroom video footage. Data collected from the

direct observations yielded PAC as an accurate measure of group behaviors. While this study provides initial indications that PAC is an accurate measure of behaviors, more research is needed before assumptions should be made as to the reliability of this observation method. Although this study reports preliminary evidence that PAC can be an accurate method of behavior assessment, no investigation has been conducted to analyze the reliability of PAC within a naturalistic classroom. Furthermore, no studies have investigated the extent to which different variables contribute to variance in behavior estimates produced by PACs.

Generalizability Theory

Traditionally, psychometric evaluations of assessments have operated under principals of classical test theory (CTT). Since SDO's in naturalistic settings rarely produces the same score every time, we assume that there are environmental factors contributing to error in observational data. CCT would attribute any variance in these scores as unspecified error. Although CTT is a common theory for evaluating the psychometric properties of assessments, it is limited in its ability to offer recommendations for how to partition and reduce score variance and eventually strengthen a measure. Generalizability theory (GT) is an alternative to CTT that permits the examination of factors that contribute to variance in assessment scores. This theory is used to assess the dependability of an observation method. Meaning how accurately does an observed sample of behavior measure continuous behavior under a range of possible conditions (Shavelson & Webb 1991). Additionally, analysis of GT models are able to identify where variance lies but also yield suggestions as to how to improve a measure that will minimize error (Briesch, Chafouleas, & Riley-Tillman 2010). To summarize, the

goal of GT is two faceted. First, GT assesses observations in relation to their global perspective in order to identify sources of variances that could contribute measurement error (Briesch, Swaminathan, Welsh and Chafouleas 2014). Second, and perhaps most importantly, the information gained from the GT is used to run decision studies or D studies. D-studies help design a measurement specification that minimizes error for a particular assessment purpose (Shavelson & Webb 1991).

GT starts by isolating sources of score variance to determine which facets of measurement contribute the most to variability in scores. Common variables that are assessed include: time of day during which the measurement occurred, the number of measurements, instrument or method of measurement, and variance associated with raters. GT results are then able to offer guidance as to possible solutions to correct any weaknesses in reliability. For example, if a large portion of variance in an observation is due to raters, GT results would indicate that a change in rater training, the number of raters or observations conducted per rater should be altered to achieve the most dependable observation. Once sources of variance are identified, follow-up decision studies can be run to assess what measurement models might yield the most dependable estimates of student behavior.

Multiple studies have been conducted utilizing GT for other behavior assessment methods, such as DBRs (Chafouleas, Christ, Riley-Tillman, Briesch, & Chanese, 2007; Chafouleas et al., 2010; Briesch, Chafouleas and Tillman, 2010); however, only a few have evaluated SDO using this model. Hintze and Matthews (2004) used GT to assess SDO across the facets of setting and time. The study used momentary time sampling to record instances of on-task and off-task behaviors across fourteen 5th grade classrooms.

They completed observation twice per day for 10 days. Analysis using GT suggested that 62% of variance was attributed to individual differences among participants, 14% attributed to the person conducting the observation and setting in which it's conduct and 24% of the score variance remaining unexplained. A follow-up d-study concluded for adequate levels of reliability, four observations per day across 20 days would be required. Other GT studies evaluations have yielded similar results with individual student differences attributing to much of the variance among SDO. For example, Briesch, Chafouleas and Riley-Tillman (2010) conducted a GT analysis and again found that a large portion of variance in SDO was attributed to variations in student behaviors. A follow-up decision study was also ran and found that only one observation per day for five days was necessary to achieve adequate levels of reliability.

Purpose of Present Study

Limited research has assessed SDO using GT. Furthermore, no published studies have assessed the dependability of class-wide observations utilizing variations of SDO. Despite literature suggesting the accuracy of these methods for assessing classroom behavior, additional research is needed before analysis can be made on the dependability of these measures.

Limited psychometric information on SDO is available in the literature with even less published literature regarding SDO of group-behavior. Additionally, we know very little about the psychometric properties of PAC other than one study suggesting it is an accurate assessment of group behavior (Dart et al., 2016). The purpose of the current study was to examine the dependability of PACs and individual-fixed SDO in regard to class-wide SDO over time, and rater. Two GT models were constructed in order to assess

the variability of scores across two facets, rater and day. Thus, a measurement model examining variance of PAC across classroom, rater, and day was completed concurrently with a second measurement mode model examining variance of individual-fixed SDO across rater, and day. Follow-up dependably studies were assessed for both measurement techniques. As stated above, teachers often rely on other measurement tools due to preference. Teachers perceptions and acceptance of the assessment tool is vital for the fidelity and future use of PAC (Horner et al., 2005). Therefore, an additional goal of this study is to assess how teachers perceive the acceptability, understanding, and feasibility of PAC. The following questions were addressed:

1. Which facet (i.e., rater or time) or combination of facets accounted for the largest proportion of variance in both GT models?
2. What was the dependability of Individual-Fixed SDO and PAC in the observed measurement model?
3. What measurement specifications resulted in dependable (i.e., $\Phi \geq .80$) estimates using Individual-Fixed SDO and PAC?
4. Was PAC rated as socially valid by classroom teachers?

CHAPTER II METHOD

2.1.1 Participants and Settings

Participants included six teachers from general education classrooms in one public elementary school located in the Southeastern United States. Subject area nor years of experience were assessed for inclusion for the study; however, each teacher selected was required to have a portion of class time in which task demands in the form of direct instruction were presented. This was a requirement since no previous studies have looked at the feasibility of teachers implementing PAC while also providing classroom instruction. Selection of teachers was based on willingness to participate. Teachers were informed that the nature of the study was observational and therefore their classroom routine would not be altered. Basic demographic information was obtained from each teacher. Additional classroom demographic information can be found in *Table 2.1*.

Table 2.1 *Classroom Demographics*

Classroom	Number of Students	Male	Female	African American	Asian	Caucasian	Hispanic
1	21	45%	55%	45%	0%	55%	0%
2	24	54%	46%	50%	0%	50%	0%
3	29	45%	55%	38%	0%	59%	3%
4	17	47%	53%	41%	0%	59%	0%
5	18	61%	39%	39%	0%	61%	0%
6	19	58%	42%	32%	15%	63%	0%

Teacher 1 was a Caucasian female kindergarten teacher in her first year of teaching. She held a bachelor's degree in education. Teacher 1's class contained 20

students, two of her students also received special education services for speech. Teacher 2 was a Caucasian female 2nd grade teacher who held a bachelor's degree of education and one year of classroom experience. Classroom 2 had 24 students with five receiving special education services through rulings of speech, developmental delay and emotional disturbance. Teacher 3, a Caucasian female, was a 4th grade teacher with four years of teaching experience. Teacher 3 held a bachelor's degree in education and was working on her master's degree in higher education. Classroom 3 had 29 students with no students receiving special education services. Teacher 4 was a Caucasian female 5th grade teacher who had 8 years of experience. She held a master's degree in education. Classroom 4 had 17 students with no students receiving services for special education. Teacher 5 was a Caucasian female 5th grade teacher in her first year of teaching. She held a bachelor's degree in education. Classroom 5 had 18 students with four students receiving special education services under rulings of autism and specific learning disabilities in reading and math. Finally, teacher 6 was an African American female 5th grade teacher with three years of classroom experience. She held a bachelor's degree in education. Classroom 6 had 19 students with no students receiving special education services.

2.1.2 Materials

MotivAider. A MotivAider® is a small device that is often used to aid in the implementation of behavioral interventions (Behavioral Dynamics, 2000). The device is designed to provide tactile prompts in the form of vibrations on a pre-set interval time schedule. This device was worn by teachers during the course of the study and was used to provide them with a prompt to conduct PACs throughout the assessment period.

*Record Form***.** In order to track student behaviors, PAC record forms (Appendix A) were used by the lead teacher. This record form contained empty boxes where the teacher filled in the number of students who were deemed academically engaged for a given interval.

Social Validity. At the completion of the study, each teacher was asked to complete the Usage Rating Profile - Assessment (URP-A; Chafouleas, Miller, Briesch, Neugebauer, & Riley-Tillman 2012)(Appendix B). The URP-A is a self-report measure that was used to assess the teachers' perceptions of the usability of PAC. Teachers were asked to respond to 28 items using a six-point Likert scale ranging from strongly disagree (1) to strongly agree (6) with total scores ranging between 28 and 168. A score of 4.0 or above may indicate that the assessment was perceived as useful. The URP-A assesses social validity along 6 factors (i.e., acceptability, understanding, home school collaboration, feasibility, system climate and system support) with Cronbach's alpha for each factor ranging from .63 - .90, suggesting adequate internal consistency and providing preliminary evidence for the measure's construct validity. (Miller et al. 2014).

2.1.3 Measures

Systematic Direct Observations. Class-wide, academically engaged behavior (AEB) was used as the primary variable for the current study. AEB included both passive and active forms of academic engagement. The definitions used for the current study are ones adapted from the Behavioral Observation of Students in School (BOSS; Shapiro, 2013). AEB was defined as any verbal or physical behavior related to engagement in academic task demands such as: writing, raising a hand, reading aloud, orientating to the teacher with eye contact, talking to the teacher or peer about assigned task demands, and

orientation to a book with eye contact directed to books content. Class-wide AEB were collected via I-F and PAC.

A 10-second I-F method was used to assess student AEB, as it has been found to be a valid method for assessing group behavior within a classroom setting (Dart et al. 2016). Coding sheets were used (Appendix C) to mark observed behaviors. At the end of each 10 second interval, observers looked up and coded the individual student as academically engaged or left the interval blank. The observers coded a different student at the end of each interval, cycling through the classroom in a fixed rotation. Once all the students in the classroom had been observed, the observers started over beginning with the first student and worked their way through the class again in the same order each time.

In addition to individual-fixed sampling, class-wide AEB was also observed by teachers and trained observers using PAC. Every three minutes, the teacher and trained observers looked up and counted the total number of students who were academically engaged. Meaning, every three minutes, the teacher and trained observers would observe each student long enough to determine if the student was academically engaged (e.g., 1-2 seconds) if the student was academically engaged the observers would count them toward a total number of students in the class who were academically engaged. If the student was not engaged, the observers would not count that student in the total for class academic engagement. This method continued until each student in the class has been assessed and a number could be written with the number of academically engaged students in the class (i.e. 12/20).

2.1.4 Procedures

Observation Training. Observers for this study consisted of six trained graduate students in a school psychology doctoral program and the participating classroom teachers. Before observations took place, each observer underwent didactic instruction followed by direct practice for coding student behaviors using video footage. Graduate students were trained on I-F procedures and PAC, whereas classroom teachers were only trained in conducting PAC. The videos used for this training were provided by the primary researcher from previously recorded classroom footage in an elementary school classroom. Video coding occurred for 10 minutes. Direct feedback was provided to individuals whose practice codes were below mastery criteria. Feedback occurred immediately upon completion of each practice observation. Mastery level was determined by comparing observers coded behaviors against the training videos. Both graduate students and teachers were trained until they reached mastery level for observation techniques of 90% agreement or higher. Percent agreement was calculated by the first author. If mastery level was not met, additional feedback was provided followed by additional practice coding with a different 10-minute video until mastery is met. This only occurred for one teacher. Teacher 1 required retraining after not meeting mastery with the first video training. However, Teacher 1 met mastery after a 2nd training.

Following training, observations took place across the six classrooms once per day for ten consecutive school days. Observations occurred during a 15-minute period while a classroom activity of either direct instruction or testing was ongoing. Observation times were scheduled with individual teachers to ensure observations were conducted while direct instruction or testing was occurring. Observations occurred at consistent times each day for all classrooms. Observations involved of a primary observer

conducting I-F assessment and PACs, a secondary observer also conducting I-F and PACs and finally, the classroom teacher conducting PACs. All procedures were approved by the University of Southern Mississippi institutional review board before being conducted (Appendix D).

Interobserver Agreement. Interobserver agreement (IOA) was calculated for the primary observer and secondary observers following individual-fixed observations and PAC for 100% of observations. Due to the nature of the study, two external observers were present for all observations therefore, IOA was calculated everyday. IOA was also calculated between the primary observer and teacher for PAC for 100% of all observations. IOA was calculated using an exact agreement method. That is, each interval of the PAC was compared between observers to assess agreements to disagreements. A percentage was derived by diving the number of agreements by the total numbers of intervals. Table 2.2 displays IOA percentages across classrooms.

Table 2.2 *Interobserver agreement across classrooms.*

	IOA		
Classroom	Researcher I-F	Researcher PAC	Teacher PAC
1	95.5%	89.3%	74.9%
2	94.7%	93.7%	89.1%
3	95.5%	94.5%	84.1%
4	94.8%	95%	91.9%
5	94.4%	94.4%	84.1%
6	97.8%	98.3%	89.5%

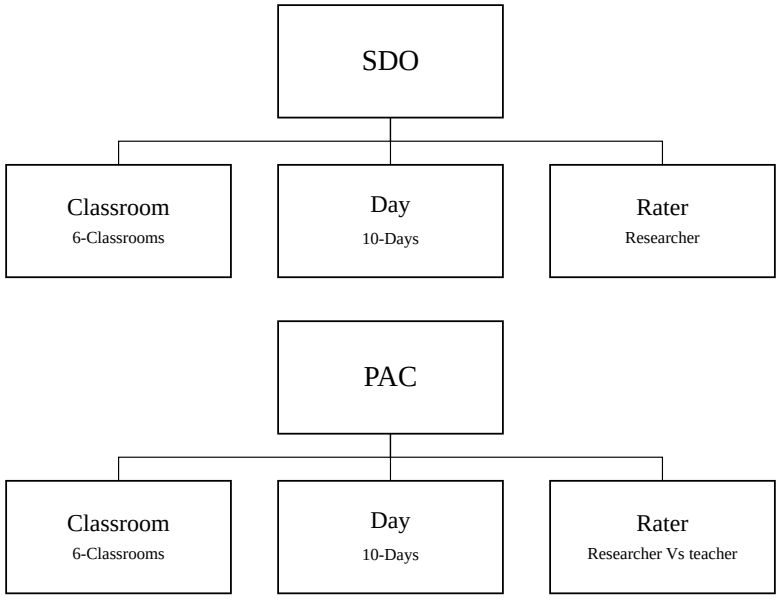
Overall Mean	95.3%	94.2%	85.5%

Total IOA for SDO across classrooms was 95.3%. Total IOA for PAC across classrooms was 94.2%. Finally, total IOA across teachers was 85.5% (Range = 74.9%-91.9%).

2.1.5 Design and Analysis

The generalizability theory was used to analyze data. GT was used to assess if PAC and momentary time sampling are dependable measures of class-wide AEB across day and rater (Figure 1.)

Figure 2.1 *Assessment Specifications for Generalizability Theory Analysis*



This was completed using two GT measurement models both using a two-facet design. Specifically, each model fully-crossed days of observation (*d*) with raters (*r*) so that the same raters are conducting observations during all assessment days. Generalizability studies were conducted to determine the dependability of each

assessment method (i.e., I-F and PAC). Following this analysis, follow-up decision studies were run to determine which assessment specifications are ideal for producing dependable estimates of class-wide AEB. A criterion of $\Phi = .80$ was used to determine dependable measurement models. This criterion has been suggested as appropriate for low-stakes decisions (Briesch 2014), which is likely sufficient for decisions based on class-wide behavior. SPSS was used to assess variance components and conduct the GT analysis using syntax developed specifically for this purpose (Mushquash & O'Conner, 2006).

CHAPTER III RESULTS

Table 3.1 represents the average AEB across classrooms for all ten observation days by observational method. On average, PAC yielded higher percentages of AEB (81.6%) than I-F (76.9%) across classrooms with a standard deviation (SD) of 11.2 for PAC and 11.8 for I-F.

Table 3.1 *Combined AEB by Classroom and Observation Method across 10 days*

Classroom	SDO	PAC
1	54.2%	59.6%
2	79.2%	81.9%
3	80.3%	86.6%
4	83.4%	88.7%
5	76.4%	82.7%
6	87.8%	89.9%
Total	76.9%	81.6%
SD	11.8	11.2

The full model G-Study for I-F (*c* x *d* x *r)* is presented in Table 3.2. From this model the differences in AEB across classrooms (*c*) accounted for the most variance (54.4%). Differences across classrooms (*c*) and day (*d*) in which they were observed accounted for 22.7% of the model's variance. Meaning that variations in student behaviors across different days explained almost a quarter of total model variance. The day (*d*) in which the observation was conducted accounted for the next largest percentage (20.3%). The rater (*r*) accounted for a small portion of variance (0.6%). Finally, unspecified error across factors (*c* x *d* x *r)* made up only 2% of variance. No other facets

of the model accounted for unique score variance. Using the proportions of variance explained in the model, G coefficients and Phi coefficients were calculated. For I-F, the full model G coefficient was .959 and the Phi coefficient or index of dependability was .925. These coefficients signify that the assessment model provides a dependable estimate of class-wide AEB.

Table 3.2 *Proportion of Variance and Dependability Coefficients for SDO*

Systematic Direct Observation	
Component	%
Classroom (*c*)	54.4
Day (*d*)	20.3
Rater (*r*)	0.6
Classroom x Day	22.7
Classroom x Rater	0.0
Rater x Day	0.0
Error (*c* x *d* x *r)*	2.0
G Coefficient	.959
Phi Coefficient	.925

The full model G-Study for PAC (*c* x *d* x *r)* is presented in Table 3.3. From this model similar to I-F, differences in classrooms (*c*) account for the most variance (41.9%). The second highest percentage of variance of 24.3% can be accounted for by differences across classrooms (*c*) and day (*d).* Rater (*r*) and day (*d*) accounted for the next largest percentage of variance (7.2%). The day (*d*) in which the observation was conducted accounted for (6.1%) of variance. The rater (*r*) accounted for a 5.8% of the variance. The interaction between rater (*r*) and classroom (*c*) accounted for 4.7% of variance in the model. Using the proportions of variance explained in the model, G coefficients and Phi coefficients can then be calculated. For PAC, the full model G coefficient was .889 and

the Phi coefficient was .881. These coefficients signify that the full model is also a reliable estimate of class-wide AEB.

Table 3.3 *Proportion of Variance and Dependability Coefficients for PAC*

Planned Activity Check	
Component	%
Classroom (*c*)	41.9
Day (*d*)	6.1
Rater (*r*)	5.8
Classroom x Day	24.3
Classroom x Rater	4.7
Rater x Day	7.2
Error (*c* x *d* x *r)*	9.6
G Coefficient	.889
Phi Coefficient	.811

Table 3.4 *G Coefficients as a function of days across I-F*

Days									
1	2	3	4	5	6	7	8	9	10
.704	.826	.877	.905	.922	.934	.943	.950	.955	.959

A follow-up decision study was run to determine what changes can be made in the measurement model to increase the dependability of I-F. Table 3.5 depicts results of the decision study based on I-F. As mentioned previously, a criterion of .80 as set as a determinant of dependable measurement models (Briesch, Swaminathan, Welsh, & Chafouleas, 2014). This criterion was selected based on previous literature that suggested $\Phi = .80$ as appropriate criterion for low-stakes decisions. (Briesch 2014). Results indicate that it would take 2 days of completing I-F to reach a G-coefficient of .826 which is deemed acceptable for low stakes decision making. After 4 days of

completing I-F in a classroom G-coefficient of .905 can be reached indicting a high stakes decision could be made based on the data gathered from observations.

Finally, Table 3.5 indicates results of the decision study based on PAC with criterion also set at .80. Results indicate that after four days of completing PAC a G coefficient of .814 would be reached. It would take over 10 days of completing PAC to reach a G coefficient of greater than .900.

Table 3.5 *G Coefficients as a function of days across PAC*

Days									
1	2	3	4	5	6	7	8	9	10
.572	.713	.777	.814	.837	.854	.866	.875	.883	.889

3.1.1.2 Social Validity

The URP-A (Table 3.6) was given to each teacher at the completion of the ten observation days to rate their overall perspective of using PAC. The URP-A assess 6-factors of social validity: acceptability, understanding, home/school collaboration, feasibility, system climate and system support. Scores closest to 6 indicate positive ratings of the assessment method. PAC did not require for classroom teachers to have home/school collaboration, assess for system climate changes or require system support. Therefore, the 3-factors of most concern are acceptability, understanding, and feasibility. Teacher 1 endorsed a score of 5 across acceptability, understanding and feasibility with an overall mean of 5 yielding positive overall ratings for PAC. Teacher 2 reported an overall mean of 5.2 indicating she agreed with the overall acceptability, understanding, and feasibility of PAC. With a rating of 5.7 for acceptability, 5 for understanding and 5 for feasibility. Teacher 3 scored an overall mean of 5.7 with positive scores across

acceptability (5.7), understanding (6) and feasibility (5.5). Teacher 4 recorded the highest overall outcomes with a mean of 5.8. Similar positive outcomes were endorsed across domains with a score of 5.8 for acceptability, 6 for understanding and 5.8 for feasibility. Teacher 5 has an overall mean of 5.5, rating acceptability at 4.8, understanding at 5.3 and feasibility at 5. Finally, Teacher 6 yielded the lowest social validity scores at 4.7. While lower than other teachers, teacher 6's scores endorsed positive perception of acceptability (5) and feasibility (5.3) with a rating of 4 for understanding.

Table 3.6 *Teacher Social Validity Data*

Classroom	Acceptability	Understanding	Feasibility
1	5	5	5
2	5.7	5	5
3	5.7	6	5.5
4	5.8	6	5.8
5	4.8	5.3	5
6	5	4	5.3
Total	5.4	5.2	5.3
Mean for Overall Acceptance: 5.3			

CHAPTER IV – DISCUSSION

Psychometric literature utilizing generalizability studies for SDO are limited (Briesch, Chafouleas & Riley-Tillman 2010;Hintze & Matthews, 2004). Furthermore, psychometric literature of PAC is almost non-existent. The current study found that PAC resulted in higher overall estimates of classroom behaviors (M= 81.6, SD 11.2) in comparison to I-F (M= 76.9, SD 11.8). Despite the slight overestimation of AEB across students, generalizability studies indicated that PAC and I-F are both dependable measures of group-wide student behaviors with PAC yielding a G coefficient of .889 and SDO a G coefficient of .959. It should be noted that criterion of .80 is set for determining dependable assessment meaning both PAC and I-F met this criterion (Briesch, Swaminathan, Welsh, & Chafouleas, 2014).

For I-F, similar to Briesch and colleagues (2010), the current study found that the largest proportion of variance was attributed to differences in classroom behavior (54.4%). This is to be expected because each of the classrooms was comprised of different students, each of whom engage in different levels of AEB. This large proportion of variance is consistent with previous literature in that individual differences in student behavior are to be expected across classrooms. This percentage of variance indicates that variations in observer scores for I-F is most likely attributed to classroom differences and not observer error.

The specific day observations occurred accounted for 20.3% of variance. Similar to how differences can be expected in student behaviors, we do not expect for students to display the exact same behaviors across days. This finding is also consistent with previous literature. Variation in students' behaviors across days are to be expected and

was accounted for by the current model. When classroom differences were crossed with days a large proportion of variance (22.7%) was explained by the model. Again, this is to be expected, when accounting for students' behaviors varying across classrooms and days. When considering a majority of the model's variance (97.4%) is attributed to classroom specifics and days in which the observation occurred. From this decision study data can be utilized to inform future observations. For I-F one observation across two days yielded a G coefficient of .826 meaning with only two total observations schools can have a dependable picture of classroom behaviors.

Finally, rater (0.6%) attributed to almost no variance in data. The current model attempted to account for variations in classroom activities (direct instruction vs testing) for I-F no variation was found. This indicates that the classroom task did not affect observer outcomes. Additionally, for rater variance, little variation in scores were present across observers. For I-F all observers were trained researchers or advanced level graduate assistants, these individuals were privy to more extensive training outside of this isolated research study. Variance from the current study would indicate that interrater reliability across raters was high. This is consistent with findings of IOA taken during the current study. It should be noted that 98% of variance could be accounted for with only 2% of variance being attributed to error. This indicates that facets included in the model accounted for nearly all score variance suggesting classroom, day, and rater are import factors to consider when utilizing I-F.

PAC generalizability data indicated similar positive results with 41.9% of variance attributed to differences between classrooms and with classroom x days accounting for 24.3% of variance in the model. Similar to I-F, observational differences

between classrooms and the particular day the observation occurred accounted for the most model variance. Based on prior SDO studies this finding aligns with the literature base and was to be expect. When combined individual classroom differences and individual days accounted for over 66% of total variance in the model. Utilizing decision studies, a G coefficient of .829 was determined after one observation per day for three days. Meaning differences across classrooms and days can be programed for and thus a reduction in variance is expected by completing one observation per day across three days. This information yields that dependable data can be recorded efficiently and dependably utilizing PAC.

Unlike I-F, more variance was attributed to raters for PAC with rater accounting for (5.8%) of variance, classroom x rater (4.7%) and rater x day (7.2%). It should be noted that PAC observations included both researchers and teachers. More variance is to be expected for these factors because unlike researchers who could devote all of their attention to the observation, teachers were required to continue teaching in addition to managing other classroom duties. Due to teachers divided attention, more difference is to be expected, thus high variance was attributed across raters. In total, less than 10% of the model (9.3%) was attributed to error meaning the PAC model was able to cover a majority of variant factors.

Finally, and perhaps most importantly, across all 6 teachers social validity data indicates that PAC was an acceptable, easy to understand and a feasible measurement tool to use in their classrooms. Previous literature suggests that lower social validity scores often lead to poor implementation of an intervention or assessment measure if it is attempted at all (McDuffie & Scruggs, 2008). In addition, research suggests that when

teachers consider an intervention or assessment tool as useful to their work, feasible and acceptable they are more likely to use it in the future (Greenwood & Abbott, 2001). Teachers rated PAC with an overall mean of 5.3 out of 6 indicating high acceptance of the measurement tool. PAC across all teacher received favorable reviews providing at least preliminary data that this assessment tool might be one that teachers are willing to adopt and use over time.

4.1.1.1.1 Limitation and Future Research

The current study only presents emerging evidence for the dependability of PAC. Future studies should seek to replicate findings. Although the current study did establish through GT's that I-F and PAC are dependable measures for group behaviors, we still have limited data for PAC to support it is an accurate measure of actual behavior (Dart et. al, 2016) . In addition, reliability of I-F and PAC were calculated using interobserver agreement. Although IOA for the study indicated that raters agreed with each other, that does not ensure that PAC is an accurate measure of classroom behaviors. The current study did not compare behaviors to duration recording so it remains unknown how well PAC data corresponds to continuous measurements. However, previous studies have supported the accuracy of PAC (Dart et. al 2016). Future studies should continue to assess the accuracy of PAC.

Finally, IOA between teacher's PACs with the trained primary observer was lower across classrooms than IOA between the primary observer and other trained researchers. Primary observer's IOA with other trained researchers averaged 94% (Range = 89%-98%) while teachers IOA average was 85% (Range = 74%-92%). Since continuous duration recordings were not taken, no concrete statements can be made on

whether teacher or trained observers were more consistent with actual classroom behaviors. However, due to the extensive nature of training that graduate students receive in SDO, it is possible that teachers were less reliable than trained observers when tracking PAC. IOA data for teachers was lower across all 6 teachers in comparison to trained researchers. This suggests that although the training provided to teachers during this study was adequate to achieve mastery during practice trails, with videos, additional training might be needed to ensure teachers are accurately recording in-vivo behaviors. Despite lower levels of IOA it should be noted that five out of six teachers were able to maintain acceptable levels of IOA with 80% or higher. Teacher one was the only participate who's IOA averaged below 80%. Future studies should consider incorporating live practice sessions to their training before teachers engage in PAC independently.

One question the current study was seeking to answer was could classroom teachers maintain their classroom duties including classroom management and teaching while also keeping dependable data. Results from the current study indicate PAC is feasible for teachers to track while also completing the duties of their classroom. This observation method lends value to push for schools be self-sufficient in collecting accurate and reliable assessment measures across tiered interventions (Gresham, 2004). The primary focus of this study was to assess the psychometric principals of PAC, future studies should assess teachers ability to use PAC in the context of monitoring the progress of tiered interventions.

CHAPTER V CONCLUSION

SDO continues to be the gold-standard for measuring classroom behaviors. However, data from this study suggest similar dependable data can be obtained with a less time-intensive observation method that classroom teachers can feasibly conduct while also managing their classrooms. The current study's results indicate that PAC is a dependable measure of classroom behavior. With one observation conducted across four days results exceed the standard of .80 with PAC yielding a G coefficient of .814. This means that PAC when conducted by teachers or trained researchers was a dependable measurement tool. Although PAC should not be used to make high-stakes decisions such as changing a child's placement or as the sole criteria for incentive pay for teachers, it does offer an alternative to traditional SDO (I-F or I-R) observation techniques that require trained personal. The utility for the use of PAC in schools offers a feasible and dependable measurement tool for measuring group behaviors that does not require graduate level training. Further evidence that PAC might be a useful tool for school personal is the efficiency of the measurement tool. Five out of six teachers were trained to use PAC to mastery in a 25-minute training session that was conducted during the teachers standard planning period. Both of the current measurement models indicate that dependable data can be recorded efficiently. However, this is only the first study to evaluate the psychometric proponents of PAC, future research should be conducted.

APPENDIX A – Teacher Observation Sheet

Figure A.1 *Observation Sheets Used by Teachers*

Teacher: _____ Date: _____

Instructions: When prompted by the motivator, look up and count the number of kids who are academically engaged. Write that number in the box with the number of total kids below.

Ex:

3-mintues
15/20

3-Mintues	6-mintues	9-minutes	12-minutes	15-minutes

Teacher: _____ Date: _____

Instructions: When prompted by the motivator, look up and count the number of kids who are academically engaged. Write that number in the box with the number of total kids below.

Ex:

3-mintues
15/20

3-Mintues	6-mintues	9-minutes	12-minutes	15-minutes

APPENDIX B URP Assessment

Figure B.1 *Social Validity Scale Used with Teachers*

USAGE RATING PROFILE

URP-Assessment

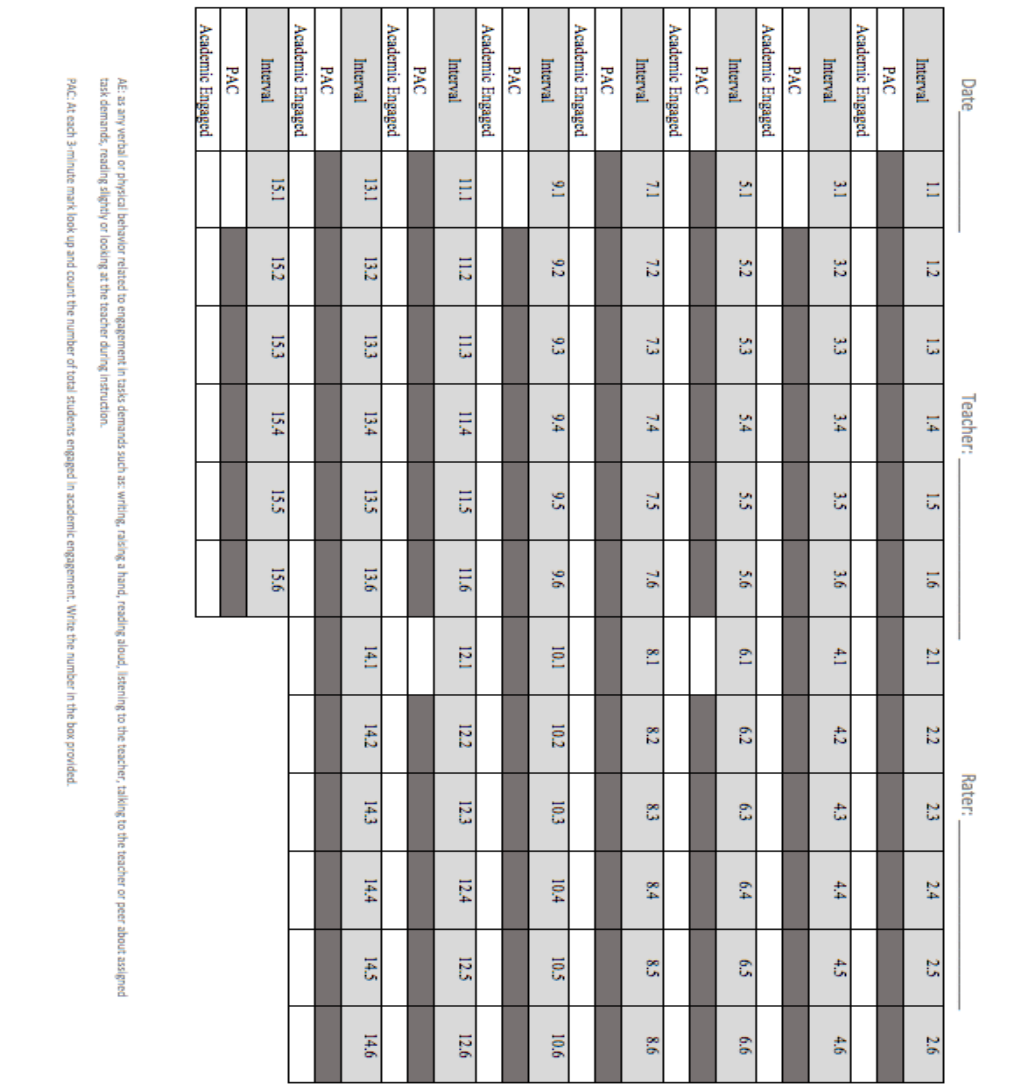
Directions: Consider the described assessment when answering each of the following statements. Circle the number that best reflects your agreement with the statement, using the scale provided below.

		Strongly Disagree	Disagree	Slightly Disagree	Slightly Agree	Agree	Strongly Agree
1.	This assessment is an effective choice for understanding a variety of problems.	1	2	3	4	5	6
2.	I would need additional resources to carry out this assessment.	1	2	3	4	5	6
3.	I would be able to allocate my time to implement this assessment.	1	2	3	4	5	6
4.	I understand how to use this assessment.	1	2	3	4	5	6
5.	A positive home-school relationship is needed to use this assessment.	1	2	3	4	5	6
6.	I am knowledgeable about the assessment procedures.	1	2	3	4	5	6
7.	The assessment is a fair way to evaluate the child's behavior problem.	1	2	3	4	5	6
8.	The total time required to implement the assessment procedures would be manageable.	1	2	3	4	5	6
9.	I would not be interested in implementing this assessment.	1	2	3	4	5	6
10.	My administrator would be supportive of my use of this assessment.	1	2	3	4	5	6
11.	I would have positive attitudes about implementing this assessment.	1	2	3	4	5	6
12.	This is a good way to assess the child's behavior problem.	1	2	3	4	5	6
13.	Preparation of materials needed for this assessment would be minimal.	1	2	3	4	5	6
14.	Use of this assessment would be consistent with the mission of my school.	1	2	3	4	5	6

		Strongly Disagree	Disagree	Slightly Disagree	Slightly Agree	Agree	Strongly Agree
15.	Parental collaboration is required in order to use this assessment.	1	2	3	4	5	6
16.	Material resources needed for this assessment are reasonable.	1	2	3	4	5	6
17.	I would implement this assessment with a good deal of enthusiasm.	1	2	3	4	5	6
18.	This assessment is too complex to carry out accurately.	1	2	3	4	5	6
19.	These assessment procedures are consistent with the way things are done in my system.	1	2	3	4	5	6
20.	Use of this assessment would not be disruptive to students.	1	2	3	4	5	6
21.	I would be committed to carrying out this assessment.	1	2	3	4	5	6
22.	The assessment procedures easily fit in with my current practices.	1	2	3	4	5	6
23.	I would need consultative support to implement this assessment.	1	2	3	4	5	6
24.	I understand the procedures of this assessment.	1	2	3	4	5	6
25.	My work environment is conducive to implementation of an assessment like this one.	1	2	3	4	5	6
26.	The amount of time required for record keeping would be reasonable.	1	2	3	4	5	6
27.	Regular home-school communication is needed to implement these assessment procedures.	1	2	3	4	5	6
28.	I would require additional professional development in order to implement this assessment.	1	2	3	4	5	6

APPENDIX C Observation Sheet

Figure C.1 *Observation Form Used by Trained Observers*



Date_____ Teacher:_____ Rater:_____

Interval	1.1	1.2	1.3	1.4	1.5	1.6	2.1	2.2	2.3	2.4	2.5	2.6
PAC												
Academic Engaged												
Interval	3.1	3.2	3.3	3.4	3.5	3.6	4.1	4.2	4.3	4.4	4.5	4.6
PAC												
Academic Engaged												
Interval	5.1	5.2	5.3	5.4	5.5	5.6	6.1	6.2	6.3	6.4	6.5	6.6
PAC												
Academic Engaged												
Interval	7.1	7.2	7.3	7.4	7.5	7.6	8.1	8.2	8.3	8.4	8.5	8.6
PAC												
Academic Engaged												
Interval	9.1	9.2	9.3	9.4	9.5	9.6	10.1	10.2	10.3	10.4	10.5	10.6
PAC												
Academic Engaged												
Interval	11.1	11.2	11.3	11.4	11.5	11.6	12.1	12.2	12.3	12.4	12.5	12.6
PAC												
Academic Engaged												
Interval	13.1	13.2	13.3	13.4	13.5	13.6	14.1	14.2	14.3	14.4	14.5	14.6
PAC												
Academic Engaged												
Interval	15.1	15.2	15.3	15.4	15.5	15.6						
PAC												
Academic Engaged												

AE: as any verbal or physical behavior related to engagement in tasks demands such as: writing, raising a hand, reading aloud, listening to the teacher, talking to the teacher or peer about assigned task demands, reading slightly or looking at the teacher during instruction.

PAC: At each 3-minute mark look up and count the number of total students engaged in academic engagement. Write the number in the box provided.

APPENDIX D IRB Form

Figure D.1 *IRB Approval Form*

INSTITUTIONAL REVIEW BOARD
118 College Drive #5147 | Hattiesburg, MS 39406-0001
Phone: 601.266.5997 | Fax: 601.266.4377 | www.usm.edu/research/institutional.review.board

NOTICE OF COMMITTEE ACTION

The project has been reviewed by The University of Southern Mississippi Institutional Review Board in accordance with Federal Drug Administration regulations (21 CFR 26, 111), Department of Health and Human Services (45 CFR Part 46), and university guidelines to ensure adherence to the following criteria:

- The risks to subjects are minimized.
- The risks to subjects are reasonable in relation to the anticipated benefits.
- The selection of subjects is equitable.
- Informed consent is adequate and appropriately documented.
- Where appropriate, the research plan makes adequate provisions for monitoring the data collected to ensure the safety of the subjects.
- Where appropriate, there are adequate provisions to protect the privacy of subjects and to maintain the confidentiality of all data.
- Appropriate additional safeguards have been included to protect vulnerable subjects.
- Any unanticipated, serious, or continuing problems encountered regarding risks to subjects must be reported immediately, but not later than 10 days following the event. This should be reported to the IRB Office via the "Adverse Effect Report Form".
- If approved, the maximum period of approval is limited to twelve months.
  Projects that exceed this period must submit an application for renewal or continuation.

PROTOCOL NUMBER: 18013007
PROJECT TITLE: Dependability of Two Group Observation Methods across Rater and Time
PROJECT TYPE: Doctoral Dissertation
RESEARCHER(S): Kayla Bates-Brantley
COLLEGE/DIVISION: College of Education and Psychology
DEPARTMENT: Psychology
FUNDING AGENCY/SPONSOR: N/A
IRB COMMITTEE ACTION: Expedited Review Approval
PERIOD OF APPROVAL: 02/14/2018 to 02/13/2019
Lawrence A. Hosman, Ph.D.
Institutional Review Board

REFERENCES

Anderson, C. M., & Borgmeier, C. (2010). Tier II interventions within the framework of school-wide positive behavior support: Essential features for design, implementation, and maintenance. *Behavior analysis in practice, 3*(1), 33-45.

Barrish, H. H., Saunders, M., & Wolf, M. M. (1969). Good Behavior Game: Effects of individual contingencies for group consequences on disruptive behavior in a classroom. *Journal of Applied Behavior Analysis, 2*, 119–124.

Briesch, A. M., Hemphill, E. M., Volpe, R. J., & Daniels, B. (2014). An evaluation of observational methods for measuring response to classwide intervention. School Psychology Quarterly, 30, 37–49.

Burns, B. J., Costello, E. J., Angold, A., Tweed, D., Stangl, D., Farmer, E. M., & Erkanli, A. (1995). Children's mental health service use across service sectors. *Health affairs*, *14*(3), 147-159.

Cashel, M. L. (2002). Child and adolescent psychological assessment: Current clinical practices and the impact of managed care. *Professional Psychology: Research and Practice*, *33*(5), 446.

Chafouleas, S., Hagermoser Sanetti, L. M., Jaffery, R., & Fallon, L. (2012). An evaluation of a class-wide intervention package involving self-management and a group contingency on classroom behavior of middle school students. Journal of Behavioral Education, 21, 34–57.

Chafouleas, S. M., Riley-Tillman, T. C., & McDougal, J. (2002). Good, bad, or in-between: How does the daily behavior report card rate? *Psychology in the Schools, 39,* 157–169.

Chafouleas, S. M., Christ, T. J., & Riley-Tillman, T. C. (2007). Generalizability and Dependability of Direct Behavior Ratings to Assess Social Behavior of Preschoolers. School Psychology Review, 36(1), 63–79

Chafouleas, S. M., Riley-Tillman, T. C., Sassu, K. A., LaFrance, M. J., & Patwa, S. S. (2007). Daily behavior report cards: An investigation of the consistency of on-task data across raters and methods. *Journal of Positive Behavior Interventions, 9,* 30 –37.

Chafouleas, S. M., McDougal, J. L., Riley-Tillman, T. C., Panahon, C. J., & Hilt, A. M. (2005). What do daily behavior report cards (DBRCs) measure? An initial comparison of DBRCs with direct observation for off-task behavior. *Psychology in the Schools, 42*(6), 669-676.

Christ, T. J. (2008). Best practices in problem analysis. In A. Thomas & J. Grimes (Eds.), *Best practices in school psychology*

Christ, T. J., Riley-Tillman, T. C., & Chafouleas, S. M. (2009). Foundation for the development and use of direct behavior rating (DBR) to assess and evaluate student behavior. *Assessment for Effective Intervention, 34*(4), 201-213.

Cone, J. D. (1978), The behavioral assessment grid (BAG): A conceptual framework

and taxonomy. *Behavior Therapy. 9.* 882-888,

Cooper, J. O., Heron, T. E., & Heward, W. L. (2007). Applied behavior analysis (2nd ed.). Upper Saddle River, NY: Pearson Prentice.

Corcoran K, Fischer J (2000c), *Measures for Clinical Practice: A Sourcebook,* 3rd ed, Vol I. New York: Free Press, pp 43–48

Dart, E. H., Radley, K. C., Briesch, A. M., Furlow, C. M., & Cavell, H. (2016). Comparing the accuracy of group observation methods: Two analyses utilizing simulated data. *Behavioral Disorders, 41*(3), 148 - 160.

Deno, S. L. (2005). Problem solving assessment. In R. Brown- Chidsey (Ed.), *Assessment for intervention: A problem-solving approach* (pp. 10–42). New York: Guilford.

Demaray, M. K., Schaefer, K., & Delong, L. K. (2003). Attention- deficit/hyperactivity disorder (ADHD): A national survey of train- ing and current assessment practices in the schools. *Psychology in the Schools*, *40*, 583–597.

Demaray, M. K., Ruffalo, S. L., Carlson, J., Busse, B. T., Olson, A. E., McManus, S. M., & Leventhai, A. (1995). Social skills assessment: A comparative evaluation of six published rating scales. *School Psychology Review*, *24*, 648-671.

Deno, S. L. (2005). Problem solving assessment. In R. Brown-Chidsey (Ed.), *Assessment for intervention: A problem-solving approach* (pp. 10–42). New York: Guilford.

Doke, L. A., & Risley, T. R. (1972). The organization of day-care environments: Required vs. optional activities. *Journal of Applied Behavior Analysis*, *5*(4), 405-420.

Dyer, K., Schwartz, I. S., & Luce, S. C. (1984). A supervision program for increasing functional activities for severely handicapped students in a residential setting. *Journal of Applied Behavior Analysis*, *17*(2), 249-259.

Gardenier, N. C., MacDonald, R., & Green, G. (2004). Comparison of direct observational methods for measuring stereotypic behavior in children with autism spectrum disorders. *Research in Developmental Disabilities*, *25*(2), 99-118.

Goh, D. S., Teslow, J., & Fuller, G. B. (1981). The practice of psycho- logical assessment among school psychologists. *Professional Psychology*, *12*, 696–706.

Greenwood, C., & Abbott, M. (2001). The research to practice gap in special education. Teacher Education and Special Education, 24, 276–289.

Gresham, F. M. (2004). Current status and future directions of school-based behavioral interventions. School Psychology Review, 33, 326– 343.

Gresham, F. M., & Gresham, G. N. (1982). Interdependent, dependent, and independent group contingencies for controlling disruptive behavior. *The Journal of Special Education*, *16*(1), 101-110.

Harrop, A., & Daniels, M. (1986). Methods of time sampling: A reappraisal of

momentary time sampling and partial interval recording. Journal of Applied Behavior Analysis, 19(1), 73 77.

Hintze, J. M., & Matthews, W. J. (2004). The generalizability of systematic direct observations across time and setting: A preliminary investigation of the psychometrics of behavioral observation. *School Psychology Review*, *33*, 258 - 270.

Horner, R. H., Carr, E. G., Halle, J., McGee, G., Odom, S., & Wolery, M. (2005). The use of single-subject research to identify evidence-based practice in special education. Exceptional Children, 71, 165–179.

Horner, R. H., Sugai, G., Smolkowski, K., Eber, L., Nakasato, J., Todd, A. W., & Esperanza, J. (2009). A randomized, wait-list controlled effectiveness trial assessing school-wide positive behavior support in elementary schools. *Journal of Positive Behavior Interventions*, *11*, 133-144.

Hutton, J. B., Dubes, R., & Muir, S. (1992). Assessment practices of school psychologists: Ten years later. *School Psychology Review*, *21*, 271–284.

McDuffie, K. A., & Scruggs, T. E. (2008). The contributions of qualitative research to discussion of evidence-based practice in special education. Intervention in School and Clinic, 44, 91–97.

McIntosh, K., Horner, R. H., & Sugai, G. (2009). Sustainability of systems-level evidence-based practices in schools: Current knowledge and future directions. In

R. H. Horner & G. Sugai (Eds.), *Handbook of positive behavior support* (pp. 327–352). New York, NY: Springer.

McKissick, C., Hawkins, R. O., Lentz, F. E., Hailley, J., & McGuire, S. (2010). Randomizing multiple contingency components to decrease disruptive behaviors and increase student engagement in an urban second-grade classroom. Psychology in the Schools, 47, 944–959.

Merrell, K. W. (2000). Informant reports: Theory and research in using child behavior rating scales in school settings. In E. S. Shapiro & T. R. Kratochwill (Eds.), *Behavioral assessment in schools: Conceptual foundations and practical applications* (pp. 233–256). New York: Guilford.

Minguet, J. L. C., & Fernandez, I. L. (2010). Effects of class content on practice time in the physical education of elementary and high school students. Studia Sportiva, 4, 77–84.

Myers, K., & Winters, N. C. (2002). Ten-year review of rating scales. I: overview of scale functioning, psychometric properties, and selection. *Journal of the American Academy of Child & Adolescent Psychiatry*, *41*(2), 114-122.

Powell, J., Martindale, B., Kulp, S., Martindale, A.,& Bauman, R. (1977). Taking a closer look: Time sampling and measurement error. Journal of Applied Behavior Analysis, 10, 325 332.

Radley, K. C., O'Handley, R. D., & Labrot, Z. C. (2015). A Comparison of Momentary

Time Sampling and Partial-Interval Recording for Assessment of Effects of Social Skills Training. *Psychology in the Schools*, *52*(4), 363-378.

Rapp, J. T., Colby, A. M., Vollmer, T. R., Roane, H. S., Lomas, J., & Britton, L. N. (2007). Interval recording for duration events: a re-evaluation. *Behavioral Interventions*, *22*, 319-345.

Raspa, M. J., McWilliam, R. A., & Maher Ridley, S. (2001). Child care quality and children's engagement. Early Education and Development, 12, 209–224.

Repp, A. C., Roberts, D. M., Slack, D. J., Repp, C. F., & Berkler, M. S. (1976). A comparison of frequency, interval and time-sampling methods of data collection. Journal of Applied Behavior Analysis, 9, 501-508.

Riley-Tillman, T. C., Chafouleas, S. M., Sassu, K. A., Chanese, J. A., & Glazer, A. D. (2008). Examining the agreement of direct behavior ratings and systematic direct observation data for on-task and disruptive behavior. *Journal of Positive Behavior Interventions*, *10*, 136-143.

Riley-Tillman, T. C., Kalberer, S. M., & Chafouleas, S. M. (2005). Selecting the Right Tool for the Job: A Review of Behavior Monitoring Tools Used to Assess Student Response-to-Intervention. *California School Psychologist*, *10*.

Risley, T., & Cataldo, M. (1973). Evaluation of planned activities: The PLA-check measure of classroom participation. Lawrence, KS: Center for Applied Behavior Analysis.

Shapiro, E. S., & Heick, P. F. (2004). School psychologist assessment practices in the evaluation of students referred for social/ behavioral/emotional problems. Psychology in the Schools, 41, 551–561.

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer*. Sage.

.