Using Generalizability Theory to Investigate Sources of Variance of the Autism Diagnostic Observation Schedule-2 with Trainees

Dorothy Rita Parriott
Univeristy of Nevada, Las Vegas, schoen11@unlv.nevada.edu

Follow this and additional works at: https://digitalscholarship.unlv.edu/thesesdissertations
Part of the Educational Psychology Commons

Repository Citation
https://digitalscholarship.unlv.edu/thesesdissertations/2721
USING GENERALIZABILITY THEORY TO INVESTIGATE SOURCES OF VARIANCE OF
THE AUTISM DIAGNOSTIC OBSERVATION SCHEDULE-2 WITH TRAINEES

By

Dorothy Parriott

Bachelor of Arts - Psychology
University of Rhode Island
1992

Masters of Arts - Educational Psychology
Kean University
2001

A dissertation submitted in partial fulfillment
of the requirements for the

Doctor of Philosophy – Educational Psychology

Department of Educational Psychology and Higher Education
College of Education
The Graduate College

University of Nevada, Las Vegas
May 2016
This dissertation prepared by

Dorothy Parriott

entitled

Using Generalizability Theory to Investigate Sources of Variance of the Autism Diagnostic Observation Schedule-2 with Trainees

is approved in partial fulfillment of the requirements for the degree of

Doctor of Philosophy – Educational Psychology
Department of Educational Psychology and Higher Education

Scott Loe, Ph.D  
*Examination Committee Chair*

Kathryn Hausbeck Korgan, Ph.D.  
*Graduate College Interim Dean*

Tara Raines, Ph.D.  
*Examination Committee Member*

Joe Crank, Ph.D.  
*Examination Committee Member*

Shannon Smith, Ph.D.  
*Graduate College Faculty Representative*
Abstract

Autism Spectrum Disorder (ASD) can be considered a serious developmental concern, which, complicated by its rising rate, creates a challenge for psychologists in properly and consistently diagnosing the disorder. Various types of assessments such as reporting measures, observation systems, and standardized assessments are currently used in the identification of ASD. Any one instrument typically examines multiple domains of functioning such as intellectual, neuropsychological (including adaptive, attention, sensory, motor, language, memory, executive functioning, academic, and social/emotional), and behavior (social and repetitive/restricted). Often, evaluators combine assessments that were not originally meant to detect ASD with those that were intended for that purpose. The most respected method of diagnosis of ASD at this time includes direct assessments as well as indirect in an attempt to maximize accuracy of clinical judgment. However, this method is not reflected in common identification practice for ASD; instead, many clinicians are relying on rating scales. Unfortunately, there are not a sufficient amount of studies examining the reliability of some of the available measures, especially with trainees. In particular, studies examining the sources of variance on the Autism Diagnostic Observation Schedule-Second Edition (ADOS-2) are limited in quantity and depth. The conceptual framework for this study will be Generalizability (G) Theory. This study is designed to expand upon currently available information regarding the interrater reliability of the ADOS-2 by using techniques available in G Theory in order to understand multiple sources of variance associated with the instrument, in particular with relation to coding by trainees.

Keywords: Autism, G Theory, ADOS, Interrater Reliability
Acknowledgements

I would like to thank my husband, Carey Parriott, and my children, Lilly, Dexter, and Crosby, for the patience and support they showed me during this process. I want to thank my parents, who always encouraged me to take chances and reach for the stars. I thank my friends for supporting and encouraging me when I needed it most and for helping me relax throughout it all. I would like to thank my Committee Chair, Dr. Scott Loe, for his encouragement and vote of confidence. He always knew how to make it seem simple when I tended to make it complex. I would like to thank Dr. Tara Raines for the time and effort she put into providing me with feedback and serving on my committee. I would like to thank Dr. Joe Crank for being part of my committee and supporting me. Also, I would like to thank Dr. Shannon Smith for joining my committee late in the game when I needed him. Thank you all for taking this journey with me. It has been an insightful process.
Dedication

This dissertation is dedicated to Dr. Paul Jones, who leaves behind him a legacy of students who have most definitely been privileged with an amazing mentor and role model. Dr. Jones was a responsive and caring mentor who always made time for his students, even if it meant responding to emails in the middle of the night. He made everyone feel worthy and confident by providing assurance and interest in their ideas. He always knew just the right questions to ask to help us self-reflect and choose the correct path. His wisdom and smile will live forever in our hearts and minds. May I follow in his path by offering the same support to those who follow.
Table of Contents

Abstract ............................................................................................................................ iii
Acknowledgements........................................................................................................ iv
Dedication .......................................................................................................................... v
Table of Contents ........................................................................................................... vi
List of Tables .................................................................................................................. ix
Chapter 1. Introduction .................................................................................................. 1
Diagnostic Practices: Clinical and Research ................................................................. 2
Statement of the Problem ............................................................................................ 10
Significance of the Problem ......................................................................................... 11
Current Study ............................................................................................................... 13
Configurations ............................................................................................................. 13
Chapter 1 Summary .................................................................................................... 15
Chapter 2. Review of Literature .................................................................................. 17
Diagnosis and Identification of Autism Spectrum Disorder ......................................... 17
Diagnostic assessment ............................................................................................... 18
  Cognitive assessments ................................................................................................. 19
  Neuropsychological assessments ............................................................................... 20
Diagnosis and development ....................................................................................... 27
  Infants and toddlers .................................................................................................... 27
  Early childhood .......................................................................................................... 28
  Middle childhood ....................................................................................................... 29
  Adulthood .................................................................................................................. 29
Autism Diagnostic Observation Schedule (ADOS-2) ................................................................. 29
Reliability of the ADOS ................................................................................................................ 34
Classical Test Theory vs. Generalizability Theory ................................................................. 40
Chapter 2 Summary ..................................................................................................................... 43
Chapter 3. Methods ..................................................................................................................... 45
Introduction .................................................................................................................................. 45
G Theory study design ................................................................................................................. 48
Reliability ..................................................................................................................................... 49
Sources of Variance ..................................................................................................................... 50
Research Questions ...................................................................................................................... 51
Data .............................................................................................................................................. 52
Trainees/Raters ............................................................................................................................. 54
Clients .......................................................................................................................................... 54
Sampling Plan ............................................................................................................................... 55
ADOS 2 Training .......................................................................................................................... 55
Data Analysis ............................................................................................................................... 56
Summary ....................................................................................................................................... 57
Chapter 4. Results ....................................................................................................................... 59
Purpose of Study ............................................................................................................................ 59
Data Analysis ................................................................................................................................ 59
Reliability ...................................................................................................................................... 59
Research question #1 .................................................................................................................... 60
Variance Components .................................................................................................................. 61
List of Tables

Table 1: 29 items-Variance Components before collapsing item scores ............................... 103
Table 2: 29 items-Variance Components after collapsing item scores ................................. 104
Table 3: 14 items-Variance components with collapsed item scores ................................. 105
Table 4: Comparison of facet variance across configurations ........................................... 106
Table 5: D-Study G Coefficients ......................................................................................... 107
Chapter 1

Introduction

The Centers for Disease Control and Prevention (CDC; 2016a, Para. 1) website defines Autism Spectrum Disorder (ASD) as: “A group of developmental disabilities that can cause significant social, communication, and behavioral challenges.” The CDC (2016c) further reports that “individuals with ASD might repeat certain behaviors and might not want change in their daily activities. Many people with ASD also have different ways of learning, paying attention, or reacting to things.” The CDC describes Autism as an important public health concern (CDC, 2016c).

Concerns regarding Autism Spectrum Disorder are intensified by the fact that the rate of Autism appears to be rising. There are multiple studies (Boyle et al. 2011; Baio, 2012) that have reported increasing prevalence rates of Autism. The CDC (2016d) has deemed Autism a large enough concern to fund a group of programs, collectively known as the Autism and Developmental Disabilities Monitoring (ADDM) Network. The ADDM monitors the prevalence of ASD. The ADDM (2016) reports that the latest Autism prevalence rate in 2012 is 1 in 68 children, which is up from 1 in 88 in 2008 (ADDM, 2012). There was a 78% increase from 2002 to 2008 and a 23% increase from 2006 to 2008, but while the rate in 2012 continues to be higher, the increase from 2008 to 2012 was not significant (ADDM, 2016).

One of the major theories for increasing rates of Autism is the inconsistent criteria for the diagnosis or definition of Autism used in research (Fombonne, 2003; Williams, 2006). The changes in the Diagnostic and Statistical Manual of Mental Disorders (DSM) diagnostic criteria for ASD have created much turmoil, with many fearing that follow-up studies would be near impossible since the criteria used in the original study would no longer be in practice.
Fortunately, the creators of the DSM-5 took much care in attempting to create a definition of Autism that would maintain similar rates of Autism diagnosis (APA, 2012a). In addition, the ADDM has gone to great lengths to maintain consistent criteria for its prevalence rate studies, which are in part based on the DSM criteria (ADDM, 2012). Despite all of these efforts to maintain consistent criteria when identifying ASD, clinicians and school psychologists in the field may not be using consistent criteria or methods for the identification of ASD. One study by Rasmussen (2009) determined that school psychologists participating in a national sample were mostly using brief rating instruments due to a lack of training with more extensive assessment measures. The exact rating instruments were not indicated.

**Diagnostic Practices: Clinical and Research**

The ADDM (2012) identifies problems with the diagnosis of Autism as a lack of biological markers and changing clinical diagnostic criteria. These factors contribute to the complicated process of tracking Autism prevalence rates. To minimize the effect of inconsistencies in clinical diagnosis, the ADDM practitioners do not rely on reporting from parents or professionals but rather have a team of trained individuals review evaluation records for consistent criteria (ADDM, 2012).

Similarly, various methodology and instruments for clinical diagnosis of ASD are also being used. In an attempt to standardize data on which they base their clinical judgments, some clinicians will use a measure that has been standardized on either individuals with Autism or typically functioning individuals. Many different instruments exist, but most seem to fall into either a report or rating system that the patient or family of the patient complete or an observation system that the clinician uses as a guide for scoring behaviors related to Autism.
There have been many studies attempting to identify the *gold standard*, or most effective way of identifying ASD (Falkmer, 2013; Flamer & Horlin, 2013). Zander et al. (2015) even attempt to investigate a measure for ASD that is reliable when used by less experienced professional raters. Kleinman et al. (2008a) referred to the gold standard in diagnosis of Autism for educational professionals as *clinical judgment*. In that study, clinical judgment was based on the DSM-IV. Clinical judgment, even though based on the well-established diagnostic criteria of the DSM, is still subject to personal biases, which is one reason test developers created instruments that would help standardize the diagnostic process.

Kleinman et al. (2008a) found that the stability from age 2 to age 4 of his clinical judgment diagnosis was similar (80%) to diagnostic stability based on the combined use of two commonly used instruments in identifying ASD: the Autism Diagnostic Observation Schedule-Second Edition (ADOS-2; Lord et al., 2012a) and the Autism Diagnostic Interview (ADI; Lord et al., 1994). Studies such as these are encouraging when considering instruments to use for making a diagnosis of ASD. However, we must also question the original source of clinical judgment that we use as our standard of comparison when determining stability levels of diagnostic instruments such as the ADOS and ADI.

Others such as Falkmer, Anderson, Falkmer, and Horlin (2013) more recently describe the gold standard as an extensive and time consuming process requiring highly qualified multi-disciplinary teams who can assess behavioral, historical, and parent report information in order to make a diagnosis. Diagnosis made by multidisciplinary teams requires consensus of professionals who may or may not be viewing the same behaviors in the same way. Falkmer et al.’s study (2013) is intended to identify the best instruments to be used for diagnosis of ASD and does identify the combination of the ADOS and the ADI as being comparable to their
definition of the gold standard in terms of diagnostic accuracy (80%). This study adds additional data regarding the psychometric properties of instruments being used to diagnose ASD to Kleinman et al.’s (2008a) study, which only addresses diagnostic stability. Again, we must question the original source of comparison, since multidisciplinary teams are still impacted by personal differences in how behaviors are viewed. The ADOS and ADI have been shown to contribute to the accurate diagnosis of ASD in multiple studies.

Sappok et al. (2013) provided information confirming the sensitivity of the Autism Diagnostic Observation Schedule-Generic (ADOS-G) and ADI in identifying ASD in adults who also have an intellectual deficiency. They indicate that the measures can be overinclusive or have low specificity of ASD in this population. No studies regarding the ADOS-2 with this population were found.

A study by Wiggins (2006) indicates that only 7% of clinicians use the ADOS when identifying students with ASD. Reis (2012) claims that only 50% of the 88 respondents in her study of school psychologists reported using the ADOS when identifying ASD. Eighty-one percent of those 44 respondents had attended a clinical training on the ADOS.

It seems questionable why instruments with such high diagnostic stability and sensitivity would not be used more often in practice when diagnosing ASD. Perhaps the training or cost of the measure is too expensive. Alternatively, maybe clinicians recognize a need for investigation of additional psychometric properties beyond sensitivity and specificity. Because it is difficult to find a reliable source of comparison for such diagnostic instruments, it is important to determine the reliability of the measures themselves without considering outside comparisons.

Since some of the available measures such as the ADOS-2 are relied upon by practitioners of all experience levels when diagnosing Autism, it is vital to examine the
reliability of such measures to ensure effective diagnostic decisions, usefulness of information, and better services to follow (Naglieri & Chambers, 2009). In addition, it would be useful to know if a less costly and shorter training would yield adequate interrater reliability. In order to determine the amount of error in the measurement of ASD, one needs to understand the reliability of the measure being used. Higher reliability means lower error and a smaller confidence interval, meaning that evaluators are closer to the true score. As a result, practitioners could be more confident in their interpretation of the results and, in the end, their diagnosis.

Psychometric properties of diagnostic instruments such as the ADOS and ADI have been historically investigated using methodology based in Classical Test Theory (CTT) (Allen & Yen, 2002), which is the traditional way of viewing reliability. The major premise of CTT states that the observed score equals the true score plus error. This concept of the observed score equaling the sum of both the true score and the error is typical of the reasoning associated with the CTT. This particular concept has shortcomings when investigating interrater reliability for observational rating scales. It does not allow one to examine specific sources of variation that could be a factor when adding the element of a rater into the equation. CTT places all error together and does not allow one to investigate the specific possible causes of variation. For example, a rater might code certain items more stringently than others or a rater might rate one child more stringently on all items than he or she rates other children.

Interrater reliability is often used to gauge the consistency of observational instruments when used by different individuals. According to Shavelson and Webb (1991), Generalizability (G) Theory was presented by Cronbach in 1972 as an alternative explanation to Classical Test Theory when explaining reliability. G Theory provides a unique conceptual framework and
statistical procedures for quantifying and providing explanations for consistencies and
inconsistencies in observed scores for objects of measurement or persons (Brennan, 2003).
Brennan suggests that G Theory is not merely an extension of CTT using Analysis of Variance
(ANOVA) procedures. Rather, he believes this is an oversimplification.

G Theory and CTT have some commonalities in concepts but use different terms. In G
Theory, true scores, as we know them in CTT, are termed universe scores, and G Theory facets
are equivalent to factors as defined in CTT. Despite the similar concepts associated with both
theories, generalizability methods allow a researcher to differentiate between various sources of
error that would simply be considered undifferentiated error in CTT. In addition, a D study used
in Generalizability Theory allows for the estimation of reliability if factors were changed, such
as if the number of raters or items were changed. This estimation is similar to the Spearman–
Brown formula but allows for a variation in more than one source of variation.

Common statistics grounded in Classical Test Theory that are used to evaluate interrater
reliability are: (a) joint probability of agreement (percent agreement, percent within 1), (b)
Cronbach’s alpha, (c) Pearson Product moment correlation, (d) Kappa, and (e) Intraclass
Correlations (ICC; Gisev, Pharm, Bell, & Chen, 2013). Another approach used to investigate
reliability of measures is based in G Theory (Shavelson & Webb, 1991). Similar to these CTT
approaches at examining reliability, G Theory techniques also produce a reliability quotient that
represents the generalizability or dependability of scores over the facets used in the study.
However, G Theory techniques use one analysis to further investigate the individual sources of
variance and the interaction of those sources, as opposed to CTT approaches that investigate
individual sources of variance, such as raters or items, separately.
Suen and Lei (2007) point out that CTT and G Theory have much in common. They explain that, like CTT, G Theory is focused on the observed scores from the test. They also remind us that the goal of both theories is to understand the quality of the test instrument by estimating reliability coefficient and standard errors. They further point out that although both theories are considered true score models by some, there are differences as well.

When conducting an interrater reliability study, there are some distinct benefits of using G Theory over CTT. One, it allows for the inclusion of numerous sources of measurement error in one study as opposed to multiple studies. Two, it supplies information regarding additional sources of error by way of examining error caused by the interaction of two sources. Examples of CTT approaches and their criticism will be examined here.

One approach used in Classical Test Theory to examine interrater reliability is to investigate agreement among raters. These approaches provide limited information regarding reliability of a measure and, depending on the method of agreement, (e.g., exact agreement, agreement within 1), can provide varying results (Goodwin, 2009). One problem with using rater agreement is that it tends to rely on only two raters for each occasion (Gisev et al., 2013). Another criticism of methods using rater agreement is that they do not account for agreement obtained due to chance (Gisev et al., 2013). Critics argue that just because raters agree on the number of behavioral occurrences, it does not mean they are observing the same exact behavioral incidents (Gisev et al., 2013). For example, if one rater observes one set of behaviors and another rater sees an entirely different set of behaviors, but they both see the same number of behaviors, it appears as if they have seen the exact same thing. Yet, in reality, they have not. Although statistics such as Kappa and weighted Kappa allow for ratings by more than one rater (Fleiss’s Kappa), correct for chance, and provide options for treating certain disagreement by
raters as more important, these methods are still restricted to using nominal or categorical data that do not consider ordering of score pairs (Gisev et al., 2013).

To correct the concerns found with reliability measures of agreement, correlational coefficients such as Pearson’s or Spearman’s assume ordinal or continuous ordering but consider only relative positioning, so that ratings of 1,2,1,2 would be considered the same as 3,4,3,4 (Gisev et al., 2013). These types of coefficients would not indicate if the raters’ means were significantly different, and they do not reflect any differences in how the raters used the provided scoring scale. Also, it must be considered that the amount of variation in the data will affect the size of the coefficients when using correlation techniques for reliability (Goodwin, 2009).

The Intraclass Correlations Coefficient (ICC) examines the proportion of variance caused by between person variability (Gisev et al., 2013). The ICC provides a ratio of true variance among persons to the sum of true variance plus random error variance. When used in interrater reliability studies, the error includes variance caused by rater differences (Goodwin, 2009). Limits of agreement can also be determined with regards to interrater reliability, which allows the researcher to not only consider differences among raters but also give an indication of whether one rater tends to consistently rate higher than another rater (Gisev et al., 2013).

Although the ICC is the reliability estimate most comparable to G Theory techniques, Generalizability Theory improves on the ICC method by allowing for the separation of multiple sources of error rather than just one, as in ICC. Both the ICC and G coefficients are used to evaluate reliability of instruments that use ratio or interval data and when more than two raters are considered (Gisev et al., 2013). ICC is often used when data are grouped rather than paired and describes the degree of similarity of units in the same group. According to Barch and Mathalon (2011), the coefficients used in G Theory are actually a type of ICC. The distinctions
between coefficients based in G Theory and ICCs are that ICCs consider only one source of error. ICCs provide an indication of the similarity of data in different groups but do not consider variation accountability to intrarater versus interrater differences. G Theory allows for more complex ANOVA designs and consideration of multiple sources of error. In addition, ICCs based in Classical Test Theory do not consider the type of decision that is to be made based on the data. G Theory accounts for whether the decision will be a relative decision or absolute decision (Shavelson & Webb, 1991). Relative decisions are based on how the person’s score compares to others, and absolute decisions are based on a set criteria. The ADOS-2 would be considered an instrument used to make an absolute decision, since there is a specific cut-off to determine diagnosis, and it is not based on how that person scores or ranks compared to others. Classical Test Theory only considers relative decisions. G Theory allows for the consideration of absolute decisions by including the variance components associated with the main effects of the various measurement facets (Barch & Mathalon, 2011).

A D Study, often used after a G Study, also extends the advantages of the Spearman–Brown formula, which allows for a prediction of what the reliability might be for a measure if more or fewer raters were used. G Theory allows the researcher to change multiple factors when predicting reliability. Even if a researcher were able to acquire all of the needed information from multiple Classical Test Theory analyses, he or she would do so at the expense of increasing the alpha, or chance of Type I error (finding a significant difference when there is not one). Criticisms and explanations of some of the Classical Test Theory approaches allow us to see the value in G Theory approaches, particularly for a study that involves an observational instrument.
**Statement of the Problem**

It is important for psychological measures to be reliable, because diagnostic instruments are heavily relied upon for making clinical decisions by professionals with varying levels of experience. Their reliability is also important for the purpose of research (ADDM, 2012). A measure is only as good as its reliability. If one cannot rely on a measure to provide comparable results for each individual in every situation, then it truly is not useful in the clinical or research process. Goodwin and Goodwin (1991) state that if a measure is not reliable or does not produce consistent scores, then it cannot be valid or measure what it is supposed to measure. Consistency as related to interrater reliability means the uniformity of different raters in their use of the measure (Goodwin & Goodwin, 1991).

Falkmer et al. (2013) report that the Autism Diagnostic Interview-Revised (ADI-R) and the Autism Diagnostic Observation Schedule (ADOS) have the largest evidence base and highest sensitivity and specificity. Sensitivity is used as an indication of the percentage of people correctly identified as not having ASD. Specificity is the percentage of people correctly identified as having ASD. When the ADI-R and ADOS were used in combination, they revealed levels of accuracy very similar to the correct classification rates for Falkmer et al.’s definition of the gold standard diagnostic procedure for ASD at 80.8%. The ADI-R alone was found to be .85 but dropped to .78 for children under three years old as compared to Falkmer et al.’s gold standard method of diagnosis or clinical judgment. The ADOS alone was also found to be .80 as compared to Falkmer et al.’s standard.

This is certainly valuable information regarding the validity of these measures, but when carefully reviewing all of the studies—which Falkmer et al. reference in relation to the ADOS—there is very limited information revealing any indication of reliability scores for this measure.
In particular, when searching for studies that focus on interrater reliability of the ADOS, the majority of studies were authored by the test creators themselves (Lord et al., 2012) and used highly experienced raters from their labs. Other studies also used experienced raters but continued to use Classical Test Theory. A more recent study, which attempted to use less experienced raters and reliability methods associated with Generalizability Theory, still did not closely examine the various sources of error of the ADOS-2 and failed to utilize beginning-level clinicians without an extensive and costly training. These studies utilize only well trained and experienced clinicians, which may not be reflective of the interrater reliability associated with new professionals or trainees.

Herein lays the basis of our problem: if clinicians, regardless of experience level, and researchers need to rely on such measures to maintain identification across raters and individuals being identified, then the measure being used must be reliable with varying rater populations. Also, if beginning-level clinicians do not have the means to obtain the extensive, recommended amount of training on the ADOS-2, then it would be useful to investigate whether a less intensive training has been effective as well. Available studies examining the reliability of the ADOS-2 measure are currently limited in that they do not investigate the possibility of using raters who are briefly trained. In addition, investigation into the proportion of variance caused by various error sources has not been reported.

**Significance of the Problem**

The limited focus of interrater reliability studies for the ADOS-2 has great implications for practice and research. Specifically within the field of school psychology, only three percent of children with Autism are identified by nonschool professionals. Fifty-seven percent are identified by a combination of school and nonschool resources and 40% by school resources
alone (Yeargin-Allsopp et al., 2003). This emphasizes the importance of approaches used by school psychologists for identifying Autism. School psychologists are a primary resource in the ASD identification process. As such, it is important that school psychology trainees develop the skills necessary for accurate diagnostic practice.

In the field of school psychology research, in order to make any valid claims about individuals with ASD, researchers must be able to obtain consistent identification of the individuals to be used in their study. Any inconsistency caused by factors within the measure being used for identification itself would invalidate any claims attributed to the true factors being investigated in the study.

Similarly, the importance of a reliable measure withstands in the clinical setting. As Klin (2007) states, one of the major goals of the DSM is to improve agreement on specific diagnosis. Having consistent diagnostic criteria is vital, but it is also important to find measures that help consistently identify these behaviors associated with the diagnostic criteria. When presenting guidelines for choosing psychological measures, Cicchetti (1994) lists reliability as one criterion for test selection. He further reports that most biostatisticians view interrater reliability as the most important measure of reliability. Choosing reliable measures is important in the identification of individuals with Autism and supports the effective planning of treatment and educational services. The existence of reliable measures is the first step in ensuring consistent diagnostic or identification practice.

In addition to the importance of choosing a generally dependable measure and having confidence in its ability to provide consistent outcomes, professionals also need to be able to understand the sources of variability when considering the results of measures such as the ADOS-2 (Shavelson, 1991). If it is determined that the main source of variability is actually due
to certain items, then the examiner using the ADOS-2 may need to not only consider the total score of the ADOS-2 but also reflect on the responses given on particular items. In addition, researchers may want to consider this information when designing future versions of the ADOS-2 or new tests (Shavelson, 1991). Another benefit of analysis associated with Generalizability Theory is that a D study will allow for prediction of reliable scores when using a certain number of items and/or raters.

**Current Study**

The current study is a G study that investigates not only the overall reliability among trainees who have received a brief training on ADOS-2 rating as part of their master’s level course but also analyzes the proportion of variance caused by these raters, ADOS-2 items, and the interaction of both. It also includes a D study, which estimates how changing the number of trainee raters or items might impact the reliability of the ADOS-2. Data obtained from a previous training were used to conduct analyses.

**Configurations**

Scoring procedures recommended in the ADOS-2 manual (Lord et al., 2012) require that, prior to adding item scores to obtain totals, scores must collapse item scores of 3 to a score of 2 and scores of 7 must be changed to a zero to reflect the true nature of the rating. Scores of 7 reflect the absence of the behavior and are therefore recoded as 0. Scores of 3 are collapsed to avoid disproportionality when weighing single items in the algorithm classification (Lord et al., 2012a).

- Configuration 1: The first study was conducted after simply converting any raw scores of 7 on a particular item to 0.
- Configuration 2: The second study not only converted scores of 7 to 0 but also collapsed scores of 3 to a score of 2.

- Configuration 3: Gotham (2007) proposed that certain items of the ADOS best differentiated between diagnoses. He adjusted scoring procedures to include only these items and divided them between two domains: Social Affect and Restricted, Repetitive Behaviors. A G study and D study were done on the 14 items that are currently used to obtain scores for determining level of Autism using the ADOS-2: Module 3.

The following eight research questions guided the proposed G study and examined for multiple G study configurations:

1. What is the overall reliability of the ADOS-2 when used by trainees with a simple and inexpensive training program?
2. What percentage of variance is caused by raters?
3. What percentage of variance is caused by persons?
4. What percentage of variance is caused by items?
5. What percentage of variance is caused by the interaction of items and raters?
6. What percentage of variance is caused by the interaction of items and persons?
7. What percentage of variance is caused by the interaction of persons and raters?
8. What is the remaining percentage of variance that is caused by an interaction of person, rater, and item plus error?

In addition, one question was posed when considering comparisons across configurations:

9. Do the variance proportions differ across various study configurations?
An additional question was posed regarding the D study that was performed solely on configuration 3, which examined only the 14 items currently used by professionals for scoring of the ADOS-2:

10. At what number of trainee raters and items is it estimated that acceptable reliability can be achieved?

Brennan (2000) explains that G Theory does not test hypotheses but rather estimates and uses variance components. Although generalizability analysis techniques do not allow for the statistical testing of a hypothesis, the expectation is that the items will account for the most significant source of variance in the ratings. In particular, because the raters are trainees, their knowledge of particular constructs measured on certain items may be more limited than that of more experienced raters.

The major assumption of this study is that persons are in a steady state, which means they are not showing effects of maturation or learning within one test administration period. Given the varying nature of tasks on the ADOS-2, persons’ performance was not expected to be influenced by learning. Also, since ADOS-2 administration occurs in one single test session, effects of maturation are eliminated.

Chapter 1 Summary

This chapter provided an overview of the proposed study involving the interrater reliability of the ADOS-2 measure. The background of the problem was covered, as was the nature of the study and its significance. Research questions were outlined. Chapter 2 will provide a more extensive review of the literature, including factors important to the diagnosis of Autism in general and more specifically to the analysis of assessment measures used in diagnosis. Chapter 2 will also define and introduce the theoretical framework of G Theory and
how it compares to Classical Test Theory in assessing interrater reliability of these measures. Chapter 3 will describe the methodology in significantly more depth and outline the applicability of G Theory to the current study.
Chapter 2

Review of Literature

Diagnosis and Identification of Autism Spectrum Disorders

Autism Spectrum Disorders are diagnosed based on the criteria set forth in the Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition (DSM-5; APA, 2012) and/or in the educational setting—the Individuals with Disabilities Education Improvement Act (IDEIA). There are similarities in criteria established by both the DSM-IV (APA, 2000) and the DSM-5 (APA, 2012). Both versions include general elements of both social communication and social interaction as areas of deficit and further outline specific examples of deficit that may be seen in ASD. The DSM-5 (APA, 2012) eliminates the criteria for Autism that requires a delay in or lack of spoken language. This is reflective of the inclusion of previously diagnosed Asperger Syndrome, which could be diagnosed without a spoken language delay. Restrictive and Repetitive Behaviors are similarly addressed in both the DSM-IV (APA, 2000) and the DSM-5 (APA, 2012).

Individuals can only be diagnosed under the DSM-5 criteria if there is evidence of symptoms starting in early childhood, even if the symptoms are not recognized until later in life (APA, 2013). This indicates that early diagnosis is encouraged in the DSM-5 but flexibility is allowed in cases where the symptoms were not recognized until later. According to the American Psychiatric Association (APA, 2013), the DSM-IV allowed patients to be diagnosed under four separate disorders: Autistic Disorder, Asperger’s Disorder, Childhood Disintegrative Disorder, or the catch-all diagnosis of Pervasive Developmental Disorder not otherwise specified. In DSM-5 (APA, 2012), these four categories were combined in an attempt to
maximize consistency across clinical and research settings (APA, 2013). The work group that created these new ASD criteria reports that this change should improve the diagnosis of ASD without affecting the sensitivity or changing the number of children diagnosed (APA, 2013). The DSM-5 (APA, 2012) allows all individuals diagnosed with ASD to still fall under the same disorder despite varying symptoms and behaviors. This should be helpful in maintaining consistency for follow-up studies in situations where researchers included all four categories from DSM-5 in their sample but not for those who only included some.

In the field of School Psychology, professionals are recommended to follow guidelines set forth in the Individuals with Disabilities Education Improvement Act (IDEIA) rather than the DSM criteria for eligibility for special education. Fogt, Miller, and Zirkel (2003) point out that, in educational law cases, the DSM is rarely used as the primary determinant of eligibility but rather the IDEA is seen as the “controlling authority.” The latest IDEIA major amendments, which affected eligibility criteria, were those of IDEA Regulations Part 300 A 300.8 C. The definition of Autism in the IDEIA addresses similar components as the DSM-5, including deficits in social communication, social interaction, and repetitive behaviors. However, the IDEIA regulations further outline considerations for school psychologists and require ruling out an emotional disturbance as the primary cause for difficulty and also require certain methods of assessment, such as using multiple assessment tools, as well as “technically sound instruments” that are used for the purposes for which the assessments are reliable.

**Diagnostic assessment.** There are many assessment instruments available to detect traits associated with ASD, and these instruments can be categorized in multiple ways. One method is to categorize instruments by content area, such as language, social functioning, and cognitive ability. Some instruments attempt to assess just one content area but others actually cross
multiple content areas. The reliability of instruments that are designed to assess specific elements (speech/language, behavior, cognitive) of ASD provides useful information but have limited predictive and diagnostic utility (Brock et al., 2006).

Another way to categorize assessment instruments is by the method of data collection, such as indirect assessment (Brock, 2006) in the form of self-report (ratings or interviews), third-party report (ratings and interviews), or direct assessment (Brock, 2006) in the form of observation (continuum of unstructured to semi-structured) and standardized assessment procedures. In general, each type of measurement instrument has its strengths and weaknesses. Reporting measures, especially rating instruments, are subject to raters’ misunderstanding items as well as attempts at deception by the rater, but they allow the examiner to obtain information that he or she might not have the opportunity to observe in a short period. Unstructured observations or play-based assessments may or may not by chance provide the opportunity to observe desired skills needed to determine diagnosis. Structured observations, on the other hand, allow the examiner to guide the play in a direction that may or may not allow the examiner to observe the skill needed to make an accurate assessment.

**Cognitive assessments.** Naglieri and Chambers (2009) report that attempts at discovering a consistent pattern of cognitive strengths and weaknesses in individuals with ASD have not been successful. They suggest carefully considering patterns of uneven cognitive skills rather than simply reporting a single cognitive or IQ score. Studies finding profiles of cognitive strengths and weaknesses among individuals with ASD are often unsupported by additional studies. For example, a review in 2007 by Lincoln, Hansel, and Quirmbach reports that individuals with ASD have an intellectual profile of higher nonverbal IQ than verbal IQ. However, previous studies have not supported this profile (Siegel, Minshew, & Goldstein, 1996;
Ehlers et al., 1997). Although the 2007 study (Lincoln et al.) is more recent and could reflect more current testing instruments, replications of this finding would be necessary before one could confidently use verbal versus nonverbal IQ as a determining factor in ASD diagnosis. In addition, there are no studies demonstrating that other disorders do not show this same cognitive profile, allowing clinicians to safely rule out competing disorders.

Overall, the usefulness of cognitive measures is found more in their ability to predict success rather than to identify ASD. Considering this, it is not surprising that the DSM-5 (APA, 2012) has not adopted intellectual capability areas as part of its diagnostic criteria. The criteria set forth by the DSM-5 (APA, 2012), such as social communication and functioning and specific patterns of behavior, do not correlate with typical intellectual assessment areas.

Klinger, Kelley, and Mussey (2009) recommend intellectual assessment of students with ASD as part of a battery of assessments for diagnosis of ASD or in designing educational plans or when looking at progress or prognosis. Unlike the DSM-5 (APA, 2012), they suggest that intellectual measures be used to determine if social and communication skills are falling substantially below the developmental level, which should be required for a diagnosis of ASD. The authors also point out the importance of cautiously considering tasks on intellectual assessments that require skills such as pointing and imitation.

**Neuropsychological assessments.** Another closely related category of assessments that are considered in diagnosis of ASD are neuropsychological assessments. In addition to the typical cognitive assessments, neuropsychological evaluations look at other areas of functioning such as adaptive skills, attention, sensory functioning, motor functioning, language, memory, executive functioning, academic functioning, social-emotional/social perceptual skills, and visual spatial skills. Barron-Linnankosk et al.(2014) used the NEPSY-II (Korkman, Kirk, & Kemp,
2007) to identify strengths in children with higher functioning ASD on tasks involving verbal reasoning skills and weaknesses in set shifting, verbal fluency, narrative memory, facial memory, and fine and visuomotor skills.

Attention. Corbett et al. (2009) report that attention is often an area of weakness for individuals with ASD. Corbett and Constantine (2006) compared results on the Integrated Visual and Auditory Continuous Performance Test for individuals with Attention Deficit Hyperactivity Disorder (ADHD) and ASD. Their results indicated that children with ASD show significant deficits in visual and auditory attention and greater deficits in impulsivity than children with ADHD or typical development. MacDonald et al. (2006) report deficits in individuals with ASD in the initiation of joint attention using a self-created structured observation system based on questions from the Early Social Communication Scales (ESCS; Mundy, Hogan, & Doehring, 1996).

Sensory. Commonly reported differences in sensory functioning of individuals with ASD include overreactions and over-focusing. Visual processing patterns are different among children with ASD. For example, Corbett et al. (2009) report that visual and/or auditory discrimination is a strength in ASD. Korkman, Kirk, and Kemps (2007) use the Arrows subtest of the Developmental Neuropsychological Assessment-second edition (NEPSY-II) to detect weak central coherence in individuals with ASD. Tomcheck and Dunn (2007) report that 95% of their sample of children with ASD demonstrated some degree of sensory dysfunction as measured by the Short Sensory Profile (McIntosh, Miller, & Shyu, 1999). The greatest differences were found on Under Responsive/Seeks Sensation, Auditory Filtering, and Tactile Sensitivity sections.
**Motor.** Motor deficits have also been found to be associated with ASD. One study by Whyatt and Craig (2011) found that children with ASD demonstrated weaknesses in catching a ball and static balance when compared to typically developing peers. The researchers suggest that motor skill deficits associated with Autism may not be global but more obvious in complex actions or core balance tasks. Vanvuchelen et al. (2007) used comparisons of performance on the Movement Assessment Battery for Children (M-ABC2; Henderson & Sugden, 1992) and the Peabody Developmental Motor Scales (PDMS; Folio & Fewell, 1983) and found that individuals with low-functioning Autism performed significantly worse than those with mental retardation on the motor test and on all imitation tasks. On the other hand, they found that the group with high-functioning Autism performed significantly worse than typically functioning individuals on the motor test but not on imitation tasks, with the exception of nonmeaningful gestures. This study supports the idea that perceptual-motor impairment is the primary cause of imitation problems as opposed to a cognitive weakness of symbolic representation.

**Communication.** Language assessments have shown impairments in comprehension, reduced attention to language, nonverbal communication and gestures, echolalia, and prosody differences. Commonly, prosody and pragmatics and certain areas of semantics such as connotations are impaired in ASD (Corbett et al., 2009). Grossi et al. (2013) used a self-created questionnaire to evaluate echolalic behavior and found that individuals with Autism tend to echo language more in situations that require them to participate in a direct verbal interaction as opposed to standing by and witnessing the verbal interaction of others. Volden and Phillips (2010) used the Children’s Communication Checklist-2 (CCC-2; Bishop, 2006), a parent report instrument, and the Test of Pragmatic Language (TOPL; Phelps-Terasaki & Phelps-Gunn, 2007)
and discovered that the CCC-2 detected pragmatic language difficulties better than the TOPL in individuals with Autism.

Memory. In the area of memory, deficits in ASD have been found in verbal memory, memory for faces, and social situations. Working memory findings are inconsistent dependent on the type of task. Southwick et al. (2011) used the Test of Memory and Learning (TOMAL; Reynolds & Bigler, 1994) and determined that episodic memory is broadly reduced in Autism. In one study (Bennetto, Pennington & Rogers, 1996) comparing individuals with ASD and individuals with a clinical sample, differences in certain areas of memory were found but the clinical sample contained various diagnoses (ADHD n=2, dyslexia n=13, etc.), making it difficult to use these results when discussing discriminative value in memory assessments.

Executive functioning. Individuals with ASD show impairment in certain areas of executive functioning. Set shifting, planning (Ozonoff & Jensen, 1999), switching, and strategy (Kleinhans, Akshoomoff, & Delis, 2005) are all examples of deficits found in ASD. It has been reported that children with ASD typically demonstrate broader and more impaired executive functioning profiles than other disorders (Geurts, Verte, Oosterlaan, Roeyers, & Sergeant, 2004; Goldberg et al., 2005), but there is no known cut-off score by which to make this determination; therefore, we are not able to use this alone as a criterion for diagnosis. Corbett et al. (2009b) used the Delis-Kaplin Executive Function System (D-KEFS; Delis, Kaplin, & Kramer, 2001) and found that children with ASD showed significant deficits in vigilance compared to the typical group and significant differences in response inhibition, cognitive flexibility/switching, and working memory compared to both groups.

Academic. In academics, children with ASD tend to do better on word reading (regular and pseudo), spelling, and computation than they do on comprehension at anything.
above the sentence level. In general, writing is a weakness associated with children with ASD (Corbett et al, 2005). Mayes and Calhoun (2008) found that children with high functioning ASD scored low on the Wechsler Individual Achievement Test (WIAT; Wechsler, 1991) written expression score. In a study (Wei, Christiano, Yu, Wagner, & Spikerhich, 2015) that used the Woodcock Johnson Test of Achievement (WJ; Woodcock, McGrew, & Mather, 2001), four distinct achievement profiles were discovered: higher-achieving (39%), hyperlexia (9%), hypercalculia (20%), and lower-achieving (32%) among individuals with ASD.

**Social Emotional and Behavioral Skills.** Social emotional and social perceptual skills are considered primary deficit areas of children with ASD. According to Klinger et al. (2009), there are many studies showing the greatest deficit areas in children with ASD to be in socialization. Although, Kenworthy, Case, Harms, Martin, and Wallace (2010) found delays in all adaptive areas, socialization was found to be the lowest subtest even relative to other low adaptive areas.

Other impairments in the socialization domain include processing of emotions, processing of facial expression, judgment of mental state, and theory of mind (Corbett et al., 2009). There are studies (Hass et al., 2012; Volker, 2010) of indirect social-emotional reporting instruments such as the Behavior Assessment System for Children (BASC; Reynolds & Kamphaus, 1992) that do show a significant difference in scores on certain subscales between those diagnosed with ASD and control groups with no disorders, but, again, there is a lack of studies that demonstrate an ability to differentiate between ASD and other disorders strictly based on assessments intended for examining social-emotional behaviors. Demopoulos, Hopkins, and Davis (2013), however, found that both children with ADHD and ASD exhibited delays in social, facial, and vocal affect recognition, social judgment and problem-solving, and
parent- and teacher-report of social functioning, making it difficult, once again, to use measures of this type in differentiating ASD from another disorder.

Another study demonstrated that the Social Skills subscales of the BASC (Reynolds and Kamphaus, 1992) could not be used to differentiate those with High Functioning Autism from those with comorbid OCD and ADHD-combined type (Lindner, 2005). In addition, it is reported that the atypicality subscale cannot differentiate High Functioning Autism from ADHD even though it is higher in High Functioning Autism than typically developing peers. They do present a significant difference in scores on the withdrawal subscale between the group with High Functioning Autism and all ADHD groups, including ADHD with comorbid OCD as well as the group of typically developing children. Caution would still have to be exercised when differentiating among emotional disabilities such as anxiety and depression and High Functioning Autism.

Another study (Giovingo, 2009) looking at differences in ratings on an indirect measure called the Child Behavior Checklist (CBCL; Achenbach & Rescorla, 2001) between a group diagnosed with ASD and a group referred for behavioral difficulties indicates that those with ASD scored higher on the Attention and Thought problems subscales. This difference was also noted when comparing the ASD group with a control group. However, this study did not examine a group with an ADHD or anxiety diagnosis. A study by Juechter (2012) reported a significant difference between the Social Disorders Content scale of the preschool BASC-2 of a group with ASD and other groups including those with other developmental disorders and typical peers. This study did not show a difference on any specific subscales nor did it include children with other diagnoses such as ADHD. There is a need for studies that compare assessment scores of various different disorder groups with ASD before instruments not intended
for assessing ASD can be used solely for diagnosing ASD. One study actually reports that certain behavioral measurements not intended for assessing ASD, including the BASC and the Child Behavior Checklist, are not able to adequately distinguish between ASD and other clinical diagnoses in preschool aged children (Myers et al., 2014).

Adaptive. Adaptive measures must be discussed, as well, when assessing cognitive ability. Klinger et al. (2009) report that individuals with Autism tend to possess relative strengths in daily living skills, yet other studies show different results including deficits in daily living skills. Kenworthy, Case, Harms, Martin, and Wallace (2010) used the Adaptive Behavior Assessment System-Second Edition (ABAS-II; Harrison & Oakland, 2003) to compare scores of individuals with High Functioning Autism to typical peers and found that the overall domain and subskills scores were lower than that of the typical peers. Perry, Flanagan, Dunn, and Freeman (2009) analyzed Vineland score profiles of children with Autism as a function of their cognitive level and found that adaptive profiles varied depending on their cognitive levels. Individuals diagnosed with Autism who had higher cognitive levels tended to have adaptive skills lower than their cognitive level, but individuals with lower cognitive levels demonstrated adaptive scores that were higher than their cognitive scores. This provides evidence that using adaptive scores alone to diagnose ASD would not be effective since adaptive profiles vary greatly within the spectrum of Autism. Just like cognitive measures, adaptive measures may be one piece in differentiating ASD from other developmental disorders, intellectual deficiency, or other clinical disorders but cannot be used as the sole factor in diagnosing ASD.

In summary, despite the extensive nature of neuropsychological measures and the associations found between specific subdomains within neuropsychology and ASD, neuropsychological assessments are not recommended to be used as the sole determinant of an
ASD diagnosis, mainly due to the inability to rule out another disabling condition as the cause for deficits in any particular area of neuropsychological functioning. Their usefulness lies in the ability to identify target behaviors that can be addressed in educational programming. Although Corbett et al. (2009) present theories and studies surrounding these areas of functioning, findings are still not reflective of any neuropsychological profile or particular strength or weakness that can be used as a determinant in identifying ASD. Neuropsychological assessments begin to get closer to the criteria set forth in the DSM-5 (APA, 2012) than do the traditional intellectual assessments, but there is still the major problem of ruling out another disabling condition as the cause of the strength or deficit.

**Diagnosis and development.** Diagnosis of ASD is complicated by developmental levels and/or age. Some instruments used in diagnosis are specific to certain ages or developmental levels and others can be used across the span of development. In younger years, assessments are most often used to diagnose individuals, but as individuals become older, measurement instruments to guide instruction or skill development become more important (Shea & Mesibov, 1991).

**Infants and toddlers.** Studies have shown that social, emotional, and social communication delays that are present in preschool aged children are also present by the second year of life (Gamliel & Yirmiya, 2009; Ibanez et al., 2014). Early diagnosis and intervention is believed to be important in predicting the success of individuals. Vacanti-Shova (2012) discovered that the age of diagnosis and intensity of services in early childhood predicted the diagnostic category (severe/moderate ASD, mild ASD, or no diagnosis) at follow-up between the ages of 7 to 15 years old. Being diagnosed at a younger age and receiving a high intensity of early behavioral interventions was associated with being in the no diagnosis group rather than
mild ASD. However, it was also associated with being in the severe/moderate group rather than the mild ASD group.

Early diagnosis of ASD allows early interventions to be delivered that can capitalize upon the increased brain plasticity and rapid development of infant and toddlers (Rogers & Vismara, 2014). Although complete remediation of ASD is not typical, improvements in language and intellectual functioning have been found (Rogers & Vismara, 2014). Intensive early behavioral interventions have been found to improve behaviors associated with Autism across multiple studies (Reichow, 2012), and a few recent studies have demonstrated long-term maintenance of such improvements (Matson & Konst, 2013; Akshoomoff et al., 2010). As a result, screening instruments have been designed to help medical professionals recognize and refer children who might have ASD (Rogers & Vismara, 2014). Screeners such as the Modified Checklist for Autism in Toddlers (M-CHAT; Robins et al., 1999) have been found effective for referring individuals with suspected ASD for a more thorough evaluation (Kleinman et al., 2008b). The latest version of the ADOS-2 (Lord et al., 2012) includes a toddler module that can be used in the more thorough diagnosis of ASD.

**Early childhood.** For children who are at least three years old, the diagnosis of ASD becomes more stable (Charman & Baird, 2002) and the diagnostic battery typically consists, in part, of assessments specific to ASD (Shea & Mesibov, 2009). The CDC (2016d) reports that the average age of diagnosis of ASD is 4.5 years old. A common concern when attempting to diagnose ASD in small children is the importance of differentiating developmental delay from ASD. Two instruments commonly used in the differential diagnosis for ASD are the Autism Diagnostic Interview–Revised (ADI-R; Lord et al., 1994) and the Autism Diagnostic Observation Schedule (ADOS-2; Lord et al., 2012a).
**Middle childhood.** School age diagnosis poses a set of unique issues. Mandell, Novak, and Zubritsky (2005) found that the average age of diagnosis of Autism is related to the severity of symptoms. Higher functioning individuals with Autism tend, on average, to receive a diagnosis at 7.2 years of age, whereas the more severely affected were, on average, diagnosed at 3 years old. Therefore, instruments used at this age need to be sensitive enough to detect Autism for high-functioning individuals. In addition, Hiller, Young, and Weber (2015) discovered that females are diagnosed later than males. One characteristic they found to be different in boys and girls, which may have contributed to the later diagnosis, was the fact that girls navigated social situations differently than boys, with boys tending to withdraw more than the girls. Differences in gender in the types of restricted interests displayed were also found.

**Adulthood.** As the population of individuals diagnosed as having ASD becomes older, there has been an increasing need for assessments to identify ASD in adults. Challenges unique to the assessment of adults include differentiating between psychopathology and ASD. The ADOS-2 adult module has been found to be effective in differentiating psychopathy and ASD but not as effective in differentiating ASD from schizophrenia (Hein et al., 2011). Another difficulty encountered when evaluating adults is differentiating between intellectual disability (ID) alone and the comorbid diagnosis of intellectual disability and ASD. The ADOS-2 adult module was found to be highly sensitive to ASD in intellectually disabled adults but tended to be overinclusive, meaning it was found to have a low specificity rate of ASD from ID.

**Autism Diagnostic Observation Schedule (ADOS-2)**

As previously reviewed, there are many instruments utilized for the purposes of providing ASD diagnoses. It is believed that the strongest instruments combine information gathered from interviews with that gathered in structured observations. One such instrument
which is considered by some as the gold standard of Autism assessment is the Autism Diagnostic Observation Schedule-Second Edition (ADOS-2; Falkmer et al., 2013; Flamer and Horlin, 2013).

The original Autism Diagnostic Observation Schedule was created in 1989 by Catherine Lord, Ph.D., Michael Rutter, M.D., FRS, Pamela C. DiLavore, Ph.D., and Susan Risi, Ph.D. Its intention as described in the ADOS-2 manual (Lord et al., 2012a) was to provide a method of standardizing direct observations of social behavior, communication, and play in children ages 5 to 12 who were thought to have Autism. The children observed needed to have a minimum of expressive language skills at a 3-year-old level. It was proposed as a complementary measurement tool to the Autism Diagnostic Interview (ADI), a semi-structured interview for the parent or caregiver that provides current and historical functioning in the developmental areas related to Autism. Both the ADOS and the ADI were originally designed for research. Ideas for activities included in the ADOS were obtained from empirical research in Autism and development according to Lord et al. (2012).

Lord et al. (2012) described the need for the first revision of the ADOS as a response to growing interest in using the instrument in clinical settings. This resulted in a need to revise the test for use with younger children and less verbal individuals. Further, the need for more items involving reciprocal social interaction and additional contexts for social behavior and communication, which would be relevant for younger children, naturally precipitated the revisions (Lord, 2012). In addition, the authors of the ADOS were part of a longitudinal study of children referred for Autism at the age of 2, which also influenced the revision of the ADOS. Because of the extended age range on the revised ADOS, the authors made the format more flexible, offering briefer activities and greater use of play materials. Algorithms for codes to be converted to scores were also added. The new revision was called the Prelinguistic Autism
Diagnostic Observation Scale (PL-ADOS, 1995) and it was intended for those aged 2 to 5. The PL-ADOS is reported by Lord et al. (2012) to have been effective for discriminating among individuals with Autism and those simply with developmental delays. However, there seemed to be a gap between the two versions, which consisted of the fact that neither seemed to be sensitive enough to individuals with Autism who had some expressive language and were between the ages of 2 and 5. Also, still missing was an assessment module for older adolescents and adults. The authors also believed that the instrument could be more efficient and reliable (Lord et al., 2012).

The ADOS did not become available commercially until 2001 when Western Psychological Services started producing it (Lord et al., 2012). This was yet another new version called the ADOS-G. This version incorporated the modules structure that we see today, which allows examiners to choose the module that best matches the language level of the client. Another change in this version was that scoring was saved until the end of all tasks. Also, some tasks were eliminated if they did not add additional information. Algorithms provided cut-offs for ASD as well as the narrower concept of Autism. Cut-offs were solely based on communication and social items. In a 2000 validity study, Lord et al. showed that the ADOS-G was effective in differentiating Autism from Autism Spectrum. In this study, predictive values were maintained for children with milder ASD but sensitivity and specificity were a bit lower for milder ASD (Lord et al., 2012).

For the ADOS-2 (Lord et al., 2012), the algorithms for modules 1-3 were revised in response to concerns about the effect of age and language on domain totals. In particular, associations between Reciprocal Social Interaction domain totals and level of cognitive impairment in preschool aged children became a concern. Children with specific language
impairment were being overidentified. Validity concerns in adequately identifying small children with mild intellectual deficiencies occurred. Also, nonverbal individuals with ASD were not being identified well because of a restricted range on the Module 1 algorithm. The revised algorithms reduced limitations on raw score distributions for young children without language and reduced associations between raw scores and participant characteristics such as intelligence (Lord et al., 2012).

A benefit of the new ADOS-2 was an increased ability to compare algorithms across modules. A comparison score was created so examiners could compare the overall score with children the same age and language level who have ASD. This score can also be used to show progress over time, from one module to another. The ADOS-2 reflects a change in protocols, which provide more explicit administration and coding instructions as well as more extensive coding and interpretation guidance. Also available with the ADOS-2 is a new scoring program (Lord et al., 2012).

The ADOS-2 process generally takes from 30 to 60 minutes to administer. This assessment is described by Lord et al. (1999) as a semi-structured, standardized assessment of communication, social interactions and relatedness, play, imagination, and stereotyped or repetitive behaviors. The assessment includes planned social interactions to encourage social initiations and responses. There are also tasks designed to encourage imaginative play.

Each person is administered activities from just one of the four modules. The selection of an appropriate module is based on the developmental and language level of the referred individual. The ADOS-2 added a toddler module for children ages 12-30 months who do not consistently use phrased speech in an attempt to assist with diagnosis of children at a younger age. The previous version of the ADOS used Module 1 to assess these children, but it was found
to be overinclusive of nonverbal children with mental age of 15 months or below (Gotham et al., 2007) with specificity under 20%. Module 1 is used with children who use little or no phrased speech. Module 1 of the ADOS-2 has two new algorithms: one for children who use fewer than five words during the ADOS-2 administration and the other for those who use more than five words or occasional phrases. The new algorithm for module 1 improved specificity for those recommended to use the toddler module but only to about 50%. The new toddler module matched activities better to evaluating children 12 to 30 months old.

Persons that do use phrased speech but do not speak fluently are administered Module 2. Module 2 also has two algorithms. One is for children below 5 and one for those above 5 years old. Since these modules both require the person to move around the room, the ability to walk is generally taken as a minimum developmental requirement for use of the instrument as a whole. Module 3 is for younger persons who are verbally fluent and Module 4 is used with adolescents and adults who are verbally fluent. The ADOS-2 added an algorithm for Module 3 and 4, which includes all verbally fluent children and adolescents.

The new algorithms used with the ADOS-2 combine two domains: Social Affect and Restricted and Repetitive Behavior to give an overall total score. This was done based on previous studies by Lord that support a unitary social-communication factor (Lord et al., 1999) and the diagnostic relevance of including restricted and repetitive behaviors. The new algorithms increase specificity in low-functioning individuals, retain the desired predictive validity, and create uniformity (Lord et al., 2012). Module 4 was not revised because the authors (Lord et al., 2012) thought that the adolescent and adult population warranted individual study, and they did not have access to sufficient samples. Some examples of Modules 1 or 2 include response to name, social smile, and free or bubble play. Modules 3 or 4 can include reciprocal
play and communication, displays of empathy, or comments on others’ emotions. This instrument yields scores in the three domains: social, communication, and a combined score. Diagnostic classification is made by exceeding cut-off scores in these three areas (social, communication, and combined). A child can be classified as having Autistic Disorder, or PDD-NOS, or as nonautistic.

**Reliability of the ADOS.** In general, the ADOS is considered a reliable measure of ASD and, along with the ADI, is often called the gold standard of ASD assessment. There are multiple studies verifying reliability and validity of the ADOS (an earlier version) and a few that address interrater reliability of the current ADOS-2. Studies addressing *interrater reliability* of the ADOS are mostly limited in their theoretical analysis to techniques associated with Classical Test Theory. The experience level of raters in available interrater reliability studies on the ADOS and ADOS-2 is limited to more experienced and professionally trained raters. These studies do not address interrater reliability of trainees using the ADOS-2. Studies addressing reliability of the ADOS-2 also do not offer information about the proportion of variance that can be attributed to various sources of variance, such as items and raters.

There are multiple studies investigating the reliability and validity of the ADOS-G, an earlier version of the ADOS. One study (Lord et al., 2000) found that the ADOS-G effectively differentiated Autism from nonspectrum disorders (Specificities: .93–1.0) but was less effective at distinguishing Autism from PDD-NOS (Specificities: .68–.79). Interrater reliability was reported as intraclass correlations for each domain as follows: Social .93, Communication .84, Social Communication .92, and Restricted and Repetitive Behavior as .82. Exact item agreement was found to be strong, as was weighted Kappa’s, for all modules. In particular, social reciprocity agreement was higher than restricted repetitive behavior agreement. Use of the
Kappa statistic limits researchers to the comparison of only two raters at a time. All modules showed *interrater agreement* of diagnostic classification above 90% for Autism versus nonAutism differentiation. When PDD-NOS individuals were included, the rate fell to as low as .81 for Module 2. A Fisher exact test showed significant disagreement between Autism and PDD-NOS. Lord et al. (2000) did not report who the raters were. They did specify that the individuals who administered the test to the persons were also involved with test creation. Their study did report that when certain cases revealed disagreement among raters who had viewed the administration on video, these were thrown out and raters were assigned new cases to watch live.

Naglieri and Chambers (2009) discuss the reliability of the ADOS-G (Lord, 2000) as reported by the test creators. They report the interrater reliability as presented by the test authors. They presented mean agreement for all items in each module as well as the least percentage of agreement for individual items. In addition, minimum Kappa’s were presented for each module. Module reliability was reported as follows:

- **Module 1:** mean exact agreement for 91.5%, items exceeded 80% agreement across raters, mean weighted Kappa was .78 with most above .60. Repetitive behaviors and sensory abnormalities were below .60.
- **Module 2:** mean exact agreement .89, all items exceeded 80% agreement. Fifteen of 26 items exceeded .60 with Kappa's and the mean Kappa was .70. Some items had Kappa's as low as .38 to .49 and agreements from 78% to 93%.
- **Module 3:** mean exact agreement 88.2%, all but two items had 80% or more agreement, mean Kappa was .65, 17 items had Kappa's higher than .60.
- **Module 4:** all items above 80% exact agreement, mean Kappa .66, 22 items exceeded .60 Kappa’s, and the remainder had over .50.
Interclass correlations for algorithm subtotals and totals for each module and the combined modules were reported. The range for the individual modules was .88–.97 for the social interaction domain, .74–.90 for the communication domain, .84–.98 for the communication-social interaction total, and .75–.90 for stereotyped behaviors and restricted interests.

Interrater agreement for diagnosis classification of Autism versus nonspectrum classification was 90% or higher for all modules. Agreement dropped when the classification for PDD-NOS was also considered. The lowest module was module 3 at 81%, then 84% for module 4, 87% for module 2, and 93.5% for Module 1. A Fisher’s exact test indicated significant results at $p < .01$, particularly between PDD-NOS and Autism. The reliability as presented by Naglieri and Chambers (2009) are consistent with those obtained directly from the study of the test creators.

One study by Olson et al. (2012) attempted to determine the level of consistency of diagnoses across multiple sites when ASD measures were used reliably. It found that despite the consistency of the measure administrations, inconsistencies in diagnosis were found across sites. This study included interrater reliability checks but it was not the main focus of the study. Regardless, the interrater reliability information provided does provide hope for a replication of the psychometrics provided by the original test creators. The acceptable standard set for this study was 75% for the total module score on the ADOS. This standard is below the acceptable standard put forth by Bracken in 1987 as .90 for an overall assessment. In addition, this study utilized the ADOS and not the most current version of this assessment, the ADOS-2.

The ADOS-2 manual (Lord et al., 2012a) explains the validity and reliability studies for the newest version of the ADOS. They do include the results of their interrater reliability
analysis, which focuses on domain, overall total scores, and classification. Analysis methods used were percent agreement, weighted Kappa’s, and intraclass correlations. Some results were the same as those of the ADOS-G study (Lord et al., 2000) and others were different. Kappa results appear to be different from the ADOS-G study, but interrater agreement percentages for items appear to be the same. Lord et al. (2012) attribute these distinctions to the following: the ADOS-2 items remained functionally identical to ADOS items but there were changes in protocols that provided more explicit administration and coding instructions as well as more extensive coding and interpretation guidance for the ADOS-2. These factors, thus, could impact interrater reliability. In addition, Naglieri and Chambers (2009) also report that some items were thrown out after the ADOS-G reliability studies. In addition, items were changed or discarded as a result of findings from these studies and would therefore create a need for a new reliability study. In general, a replication study would be useful in contributing to examiner confidence in the use of the ADOS-2 measurement instrument for ASD identification purposes.

Reliability data for the toddler module is presented in the ADOS-2 manual (Lord et al., 2012a) also. Most Kappa’s for diagnostic agreement were greater than .60 (substantial) and the rest were greater than .45 (moderate). Percent exact agreement across all items and rater pairs was 87%. Most items had agreement above 80% and the remainder ranged from .71-.78. The combined interclass correlation was .96, .95 for Overall Total, .95 for the Social Affect (SA) domain, and .90 for Restricted and Repetitive Behavior (RRB) domain. The lowest correlation was on the RRB domain for older children with some words at .74. The highest correlation was .99 for the SA domain for the older children with some words and the overall total for toddlers with some words.
Falkmer et al. (2013) did a literature review on the ADOS. They found 24 articles addressing its accuracy, reliability, validity, and utility. However, in reviewing the articles individually, there is limited information regarding the interrater reliability of the ADOS and in particular the ADOS-2. Only one of the studies (Kim & Lord, 2012) is dated after the release of the ADOS-2 in 2012 but nevertheless evaluates the reliability of the ADOS-G. This study mentions that interrater reliability was periodically assessed for the purpose of maintaining research standards of administration, but results are not published and the main focus of the study was on specificity and sensitivity of the ADOS and ADI for diagnosis.

Another study by Ventola et al. (2006) found validity of the ADOS as compared to other instruments and the DSM-IV-based clinical judgment to be very high but does not truly address the topic of interrater reliability. There are other studies (Rissi et al., 2006; Oosterling, 2010; Lord et al., 2012) that address the diagnostic ability or validity of the ADOS as well.

The review of literature by Falkmer et al. (2013), as well as other studies mentioned above, examine existing research on the psychometric properties of instruments used in the diagnosis of Autism and provide robust evidence of the sensitivity and specificity of the ADOS and the ADI. Unfortunately, they still fail to provide existing evidence of the depth and breadth of information needed on the interrater reliability of these measures, which is required to establish clinical confidence in using these assessments.

The original studies by the test creators are among the few studies found which do address the interrater reliability of the ADOS (Lord et al., 1989; Luyster et al., 2009). Lord et al. (1989) report that adequate interrater reliability had been established for prior observation scales, but validity of discriminating between different diagnostic groups was not. The interrater reliability study (Lord et al., 1989) for the original ADOS consisted of 20 individuals with
Autism and 20 with intellectual deficiency, matched for chronological age and verbal IQ (as an estimate of language ability), all between ages 6-18 and IQ of 50-80. A balanced incomplete block design was used, and weighted Kappa's were determined in order to take into account degree of disagreement. All persons met DSM III-R criteria for Autism and Childhood Autism Rating Scale (CARS) scores over 30. Lord notes (1989, p. 192) raters in this study were individuals who had previously worked on creating the test. Although the test manual says that one does not need extensive experience with the ADOS-2 in order to rate behavior, the interrater reliability study does not match this description nor does it match the typical raters that one would expect to use this instrument. A further study is needed utilizing raters who have minimal experience using the ADOS-2.

Most recently, Zander et al. (2015) did a G study that looked at all four modules of the ADOS-2, and the results revealed adequate reliability. This study was an attempt to use raters in more typical clinical settings as opposed to the research setting employed by the original Lord (2012) study, which had used well experienced researchers. However, the Zander et al. study (2015) still used some clinicians who were research reliable and certified ADOS trainers. The raters also had undergone the two-and-a-half-day ADOS training recommended by the test creators. Results revealed acceptable reliability even with raters with more variation in ADOS-2 training and experience. For example, G coefficient for the Module 3 was .85. This study, however, did not discuss the variance caused by multiple sources, which is one of the primary benefits of using G Theory.

The bulk of the studies that do address interrater reliability on the ADOS do so on older versions of the instrument. They typically approach the analysis via the Classical Test Theory, which does not allow for the identification of specific sources of variation among raters. For
example, Lord et al. (2000, 2012) utilized the intraclass correlation method, which is based in Classical Test Theory. Although Zander et al. (2015) did use G Theory to examine interrater reliability of the current ADOS-2 version, they did not further explore the multiple sources of variance of the ADOS-2.

**Classical Test Theory vs. Generalizability Theory**

Classical Test Theory (CTT; Allen & Yen, 2002) is the traditional way of viewing reliability, which is represented by the equation stating that the observed score equals the true score plus error. According to Shavelson and Webb (1991), Generalizability Theory (G Theory) was presented by Cronbach in 1972 as an alternative explanation to CTT when explaining reliability. G Theory provides a unique conceptual framework and statistical procedures for quantifying and providing explanations for consistencies and inconsistencies in observed scores for objects of measurement or persons (Brennan, 2003). Some view G Theory as an extension of CTT using ANOVA procedures, but Brennan (2003) suggests that this is an oversimplification. Although the universe score of G Theory can be compared to the true score of CTT, generalizability methods allow a researcher to differentiate between various sources of error that would simply be considered undifferentiated error in CTT.

Shavelson and Webb (1991) wrote a primer on G Theory in an attempt to provide an easily understandable introduction to the topic. They describe G Theory as a statistical theory that examines the dependability of behavioral measures. They define dependability of measures as the accuracy by which a particular measure or score generalizes from the average of all possible scores; for example, the accuracy by which the score of a particular rater would generalize to all possible raters.
Shavelson and Webb (1991) point out that an assumption made when examining dependability of a measure is that the individual being measured is in a steady state so that the score is not affected by maturation or learning, for example. Therefore, they explain, the differences in measurement can be assumed to be from one or more sources of error, and the researchers further note it is unlikely that any one score would be completely dependable or the same on different occasions, alternate test forms, or with a different test administrators.

A frequently used method for the estimation of interrater reliability is based on CTT and is the calculation of the correlation between two raters (Goodwin & Goodwin, 1991). G Theory techniques represent a combination of correlation and a separate comparison of raters’ means. Unlike CTT, G Theory can estimate various sources of error in one single analysis that deflates the Type I error rate, keeping it at an alpha of .05, for example. CTT can estimate the same sources of error but would only be able to do so with one at a time (Shavelson and Webb, 1991). Brennan (1992) argues that G Theory frees CTT by providing methods for disentangling multiple sources of error. The coefficient of G Theory provides a summary of dependability (Shavelson and Webb, 1991). G Theory allows the researcher to determine how many test occasions, forms, or administrators are necessary to obtain dependable scores.

Suen and Lei (2007) point out that CTT and G Theory have much in common. They explain that, like CTT, G Theory is focused on the observed scores from the test, and they also remind us that the goal of both CTT and G Theory is to understand the quality of the test instrument by estimating reliability coefficient and standard errors. They further point out that although both theories are considered true score models by some, there are some differences as well.
G Theory and CTT have some commonalities in concepts but use different terms. In G Theory, *true scores*, as we know them in Classical Test Theory, are termed *universe scores* and *facets* are equivalent to CTT *factors* (Shavelson & Webb, 1991).

Various possibilities for facets exist in this theory and include: *items, forms*, and *occasions* (Shavelson & Webb, 1991). Shavelson and Webb (1991) describe the item facet as how well one particular item or question on a test will generalize to all possible items. Certain facets or sources of variation in each study will remain constant, meaning that they affect all persons. Other facets will interact with each person differently. For example, in an observational study, a lenient versus stringent rater remains constant for all persons, but if a rater is more lenient with certain persons than others, this creates a rater-by-person interaction and these two facets are thought to be crossed (Shavelson and Webb, 1991). These researchers consider facets in G Theory as either *random* or *fixed*, and this designation affects the generalizability of the measure.

Studies based in G Theory are considered either Generalizability (G) studies or Decision (D) studies. G studies are intended to fully discover any sources of variation, and D studies are built then to use the information obtained from a G study in order to best design the measure (Shavelson and Webb, 1991). Measurements are typically used either to rate a person’s score in relation to others’ or to establish a level regardless of its relation to that of others. Decisions made based on the standing in relation to others are called *relative decisions* and those based on level of performance regardless of others are called *absolute decisions* (Shavelson and Webb, 1991). The G coefficient reflects how accurate the generalization is from the observed score to the universe score (true score in CTT) or the proportion of variance that is due to universe score variability (Shavelson and Webb, 1991).
According to McWilliams (1994), G Theory goes beyond the traditional, linear conceptualization interrater agreement. He further goes on to explain that G Theory provides the source and magnitude of error variance accounted for by persons and other factors such as measurement occasions, raters, and test forms (McWilliams, 1994). McWilliams also states that typically in CTT two raters watch the same videos (many) and pairs are evaluated. With G Theory, there are more raters and fewer videos. He also indicates that in a G study, an ANOVA is conducted putting the Sums of Squares (SS) into main and interaction effects. Generalizability Theory allows consideration of other factors that may cause variability besides individual difference and measurement error (McWilliams, 1994). Acceptable G Theory results are in the .60-.70 range as compared to traditional reliability results in the .80 and .90s (McWilliams, 1994). G coefficients can be interpreted as an overall indication of reliability in the same manner as reliability coefficients in CTT.

One potential problem of studies using the ADOS is determining whether the reliability of administration is maintained over various testing situations (Gotham, 2007). This problem can be minimized using the G Theory since fewer separate test administrations are required. This theory allows one to rule out poor reliability of test administration as a potential factor for error when looking at interrater reliability. Common terms used in Generalizability Theory (Shavelson & Webb, 1991) can be found in the Appendix.

Summary

Chapter 2 provided a review of literature related to the study. Criteria for diagnosing ASD and instruments used in the process are reviewed. The ADOS is reviewed in detail, including studies examining reliability of the instrument. The theoretical framework of the study, Generalizability Theory, is explained and compared to Classical Test Theory in terms of
the interrater reliability of measures. Common terms used in G Theory are listed. Chapter 3 will provide detailed information about the data and the background of the training from which the data were obtained. Analysis procedures associated with G Theory as used in the study will be further explained.
Chapter 3

Methods

Introduction

Although generalizability approaches do include Analysis of Variance (ANOVA) procedures, Brennan (2003) suggests that thinking of generalizability methods as simply the application of ANOVA procedures to Classical Test Theory is misleading for two reasons. One, G Theory does not utilize hypotheses but focuses on the use and estimation of variance components. Second, this assumption overlooks the unique conceptual framework of G Theory. A G study collects data from a universe of admissible observations (Brennan, 2003). Facets are defined by the researcher. A G study results in a set of random effect variance components (Brennan, 2003). G Theory served as the conceptual framework for this study.

According to Brennan (2003), in a univariate model such as the one in this study, each object of measurement or person will have one universe score. Univariate models utilize random effects models. Variance components are obtained for each facet, interactions of facets, and remaining error. These are estimated by using the expected mean squares in random effects ANOVA (Brenna, 2003). Universe score is denoted as $\sigma^2(\tau)$.

Kim, Park, and Kang (2012) elaborate on the process of using G Theory for interrater reliability. They designate the object of measurement as the person and the rater or items as the facets. They describe the variation due to rater as rater severity and the variation due to the interaction between person and rater as rater inconsistency. Rater severity may differ between raters in that one rater might rate all persons more harshly than another rater. However, there is also an interaction between the rater and each individual person. This is rater inconsistency and means that any one rater might also rate a particular person differently from another person. A method based in Classical Test Theory would not include all of these components.
G Theory (Brennan, 1992) describes the universe of admissible observations as all of the possible $x$ that a researcher might be interested in where $x$ equals a particular observation (e.g., test score). In this study, the facets are raters and items; therefore, we assume that the raters and items used constitute an *admissible condition of measurement* (Brennan, 1992). The universe of admissible observations for this study includes a rater and an item facet. In a G study, it is assumed that the universe is infinite (Shavelson, 1991). For purposes of this study, the universe is an infinite amount of scores on the ADOS-2. In G Theory, it is the researcher who chooses which raters constitute the universe of conditions for the rater facet (Brennan, 1992) and which items constitute the universe of conditions for the item facet. There is no set definition of rater or item that all investigators would accept.

In this study, the researcher chose master’s level students in the field of school psychology who appeared to fit within the expectation of users according to the test creators, who indicate that raters should have some experience with ASD (Lord, 2012). Items in this study are defined as those items on the ADOS-2 measure. Readers are then free to decide if the universe of admissible observations used for this study is relevant to their concerns (Brennan, 1992). In G Theory, it is unnecessary to designate an *admissible universe of persons*. However, it can be deduced that, when considering the rater facet universe, the appropriateness of the persons were considered (Brennan, 1992). In G Theory, *universe* is used instead of *population* for the object of measurement, in this case person or persons. Because ADOS-2 ratings of any person in the universe to any person in the population would be admissible, the design is considered *crossed* (Brennan, 1992).

G Theory would define the universe score in this study as the mean score for every conceivable instance of ADOS administration in the researcher's universe of generalization. The
expected value for any such person or person would be defined as the person’s universe score (Brennan, 1992). The variance of universe scores across all persons in the population is termed *universe score variance*, which is comparable to *true score variance* in Classical test Theory.

Brennan (2001) describes a standardized interrater reliability study as one in which correlation between scores of two raters to student responses to the *same* task as opposed to different tasks in a non-standardized interrater reliability study. A standardized design will be used in this study where raters rate the same tasks.

Brennan (2000) points out that the very nature of a G study is the reason that most G studies consist of single observation per cell. Therefore, he considers the designs used in G studies to be a subset of ANOVA. Brennan further notes this prevalence of single observation per cell design is not due to a failure to collect enough data. Rather, he explains that to replicate within cell observations would simply provide interchangeable findings and would not provide any additional information with regard to a particular facet. In this study, the instrument being studied (ADOS-2) uses multiple raters looking at the same person at the same time. The facets of rater and item are the facets of most concern in this study; therefore, a minimum number of persons is needed to generalize to the universe of admissible raters.

Brennan (2000) also explains that G Theory blurs the line between reliability and validity and can answer questions in both realms and across realms. Gage and Prykanowski (2014) point out that both inter- and intra-observer agreement is necessary to establish validity. One cannot have validity without reliability. They remind readers that reliability does not assume validity but is a precondition of validity. An example they provide is that two observers can agree on the frequency of a behavior they see, but that does not mean that behavior is a true indicator of the construct being investigated. Gage and Prykanowski (2014) further point out that the scores
from measurement instruments are what is being assessed for reliability or validity on the instrument itself. This point reminds us that there are other factors, such as raters, that are independent of the assessment instrument but do impact the reliability of a score. The exception to this concept is that G Theory will not answer questions about construct validity, because it allows the researcher to choose his or her own facets. Therefore, it follows that this study will not answer the question of whether the questions on the ADOS-2 are true indicators of Autism but rather examine the facets of variability in use of the instrument. This study can and will answer the previously proposed research questions regarding overall reliability of the ADOS-2 and proportions of relative variance caused by the designated facets outlined in Chapter 1.

**G Theory Study Design**

In G Theory, when any rater crossed with any item would yield a meaningful measure, the study is considered *crossed*. The design of this study is a two-faceted, crossed design (Brennan, 2003), meaning that all persons were rated by the same raters using the same items. The design would be denoted as \( r \times i \times p \), where \( r \) stands for raters, \( i \) for items, and \( p \) for the person or object of measurement. The \( x \) is read as crossed and indicates that any combination of rater and item would be acceptable to this researcher (Brennan, 1992). Brennan (2001) describes a standardized interrater reliability study as one in which correlation is made between scores of two raters to student responses to the *same* task, as opposed to different tasks in a non-standardized interrater reliability study. According to this description, a standardized design was used in this study where raters rate the same tasks.

In G Theory, it is the investigator who determines which raters and items will be the universe of conditions for the respective facets, and the theory does not assume there is a universally accepted definition of raters or items that would be accepted by all investigators. For
example, the universe of raters consists of master’s level trainees in this study, and this universe might not be of interest to other researchers. Brennan (1992) suggests that it is up to the reader to decide if the universe of admissible observations is relevant to his or her own needs.

Raters and items are sources of unwanted variation on the ADOS 2 assessment. Raters and items are considered to be random as opposed to fixed facets, because they are randomly sampled from an indefinite universe of trainee raters and items that could potentially measure ASD.

**Reliability**

In G Theory, two reliability coefficients can be obtained. Either the Generalizability coefficient or Phi is comparable to a reliability coefficient in Classical Test Theory. The G coefficient is denoted as $E_{p2}$ (Brennan, 2003) and is used when considering data used to make relative decisions. Phi $\Phi$ is the coefficient used when examining absolute decisions. Relative decisions are those that are based on scores representative of a comparison to levels seen in other individuals being given the assessment. A score that reflects a ranking of one child compared to another would be an example of a relative decision in G Theory. Absolute decisions are those that are not dependent on the comparison of performance to the performance of others, such as is the case in this study, because scores are to be given based on the presence of behaviors and not as compared to levels seen in other individuals being given the assessment (Webb & Shavelson, 2005).

Individuals being assessed with the ADOS-2 are not ranked or compared to each other on the levels of behaviors seen. Since the ADOS-2 is used to make absolute decisions, the Phi coefficient was used in this study. In G Theory, the Phi coefficient serves as an indication of how accurate the generalization is from the observed score based on the sample of the behavior to the
universe score, which is what is considered the true score in CTT (Brennan, 2003). This Phi score can also be described as the proportion of variance that is due to universe score variability. Universe score is defined as the expected value of a person’s observed scores over all observations in the universe of generalization (Shavelson and Webb, 1991). Thus, in this study, Phi represents how well the observed score represents the mean score that would be obtained if the person were rated over and over by an infinite number of raters on an infinite number of items.

Sources of Variance

Generalizability Theory allows us to go beyond identifying variability caused by “true” differences in observed test scores and variation due to error that is the combination of both systematic and random sources. It allows us to pull the sources of variation apart from the global error factor that includes omitted variables, interactions between factors, and interactions between factors and the object of interest (Shavelson & Webb, 1991). For this study, it allowed us to further break down sources of error that may cause variance in the observed scores.

Because the items of the ADOS-2 vary in terms of specific behaviors, one can conclude that item variability will be a potential source of error or variability in generalization. Therefore, items were chosen as a facet of the study. ADOS-2 users are interested in the total results of the ADOS-2 scoring rather than specific items scores because this is how diagnostic decisions are made. The universe of items would consist of an infinite amount of items that could possibly measure behaviors associated with ASD. This doesn’t mean that each item has to represent Autism as whole but should mean that each item represents some piece of the disorder that is required for diagnosis. Generalizability Theory seeks to know how accurate the generalization
from the set of items on the ADOS-2 is to all possible items as an indicator of the presence of ASD. The variation in each item will be a source of variability.

Another source of variability is the object of measurement or, in this case, the child’s behavior that is being rated on the video being watched and rated by master’s level students. This reflects variability in the presence of behaviors associated with Autism in each child on the video. Figure 1 demonstrates the sources of variability and their relationships.

Kim, Park, and Kang (2012) elaborate on the process of using G Theory for interrater reliability. They designate the object of measurement as the person and the rater or items as the facets. In this study, the object of measurement was the persons or children on the videos who were being rated. The various sources of error included (a) the persons or children; (b) raters (facet 1) or trainees in the master’s level class; (c) items on the ADOS-2 (facet 2); and (d) the interactions between the person and rater, (e) person and items, (f) rater and items, and (g) a residual consisting of a combination of person, rater, and items coupled with error (which includes unmeasured facets affecting the measurement or random events).

Research Questions

The following eight research questions guided the proposed G study and are being examined for multiple G study configurations.

1. What is the overall reliability of the ADOS-2 when used by trainees with a simple and inexpensive training program?
2. What percentage of variance is caused by raters?
3. What percentage of variance is caused by persons?
4. What percentage of variance is caused by items?
5. What percentage of variance is caused by the interaction of items and raters?
6. What percentage of variance is caused by the interaction of items and persons?

7. What percentage of variance is caused by the interaction of persons and raters?

8. What is the remaining percentage of variance that is caused by an interaction of person, rater, and item plus error?

In addition two questions were posed when considering comparisons across configurations.

9. Do the variance proportions differ across various study configurations?

10. At what number of trainee raters and items is it estimated that acceptable reliability can be achieved?

The research questions as proposed in Chapter 1 are easily answered by the data analysis employed in this study. The variability associated with each facet of the study was calculated using SPSS (IBM, 2013) and methods and matrices provided by Mushquash and O’Connor (2006). Figure 1 represents the different sources of variance that are defined in the current study. This allowed for the calculation of variance components associated with the object of measurement (persons or children seen on videos), raters (graduate students scoring the assessment) as well as items on the ADOS assessment and the interaction of all three. Table 1 provides the variance component estimates for all facets of the study. The magnitude of variability is expressed in terms of variance components according to Shavelson and Webb (1991).

**Data**

ADOS-2 rating scores were obtained from the professor of a master’s level class in a school psychology program at the University of Nevada, Las Vegas (UNLV). Students were enrolled in a spring semester, first-year, 3-credit master’s level class covering assessment topics. Coding scores from the practice sheets completed by the students as part of an ADOS-2 training
assignment were analyzed. The students’ practice codings were based on videos of ADOS-2: Module 3 administration obtained from the Center for Autism Spectrum Disorders at the University of Nevada, Las Vegas. The administration of the ADOS-2 assessments were performed by a qualified, experienced evaluator who completed the ADOS-2 training.

Figure 1 Venn diagram portraying the relationship of the study design

Crossed Design p x r x i

p = object of measurement or child in video
   = items on the ADOS-2
r = raters or master’s level students
e = error
Trainees/Raters

First year students enrolled in a master’s level class at UNLV, also referred to as ADOS-2 trainees, were considered the raters, or facet 1 for purposes of this study, and received an ADOS-2 training in their class. Considering the rater facet universe, the appropriateness of the raters in this study were considered as suggested by Brennan (1992). These raters can represent any random selection of graduate-level trainees, which is considered the universe of admissible raters in G Theory.

An informational questionnaire was collected from the students who were being considered as raters. This questionnaire indicated that raters had prior course experience with individualized assessment measures and general knowledge of ASD. The questionnaire also indicated that student raters had familiarity and experience with typical and atypical development across various ages and disorders. This meets the requirements of the test producers as stated in the ADOS-2 manual (Lord et al., 2012a) and reflects an outcome that would possibly be applicable to beginning practicing school psychologists. Ten of these students provided coding scores after they received a two-session classroom training on the ADOS-2.

Clients

The object of measurement for this study was clients of the UNLV Center for Autism Spectrum Disorders (CASD) whose parents had previously agreed to have their child’s assessment process videotaped and used for educational and research purposes and who were being assessed via the ADOS-2 on the practice videos. The population in G Theory refers to the admissible objects of measurement as designated by the researcher. For this study, any client of appropriate age and language level for the ADOS 2: Module 3 (Lord et al., 2012a) would be admissible as an object of measurement and would be considered within the designated
population. Two client videos were viewed and coded via the ADOS-2 Module 3 protocol. It is these coding data that were obtained from the professor.

**Sampling Plan**

According to Shavelson (1991), in G Theory all facets are assumed to be random when the size of the sample is much smaller than the universe size and the sample is considered to be exchangeable with any other sample of the same size. Another type of sample has the conditions of the facet equal to the number of conditions in the universe of generalization. These are called *fixed samples*. For purposes of this study, the sample of raters will be considered a random sample because it is exchangeable with any other sample of raters of the same size.

**ADOS-2 Training**

Prior to the first class, students were asked to read the pages from the ADOS-2 manual (Lord et al., 2012a, pp. 69-81) relevant to Module 3. In class session 1, raters were trained using a Microsoft PowerPoint® presentation based on the training guidebook (Corsello et al., 2012, pp. 16-22) outlining essential coding topics and the purpose and focus of observation from pages 69-81 (Lord et al., 2012a). Then, the training video (ADOS-2 Training Video, 2012) for Module 3 was shown to the raters as they followed along in the case example pages of the book (Corsello et al., 2012, pp. 69-81). The training video included guiding comments provided by the test administrator during the administration of the ADOS-2. Last, the raters completed a practice rating for the same child (David B.) and exact same ADOS-2 administration that was previously viewed in the training video for Module 3 (but without guiding comments this time) as provided on the practice video (ADOS-2 Practice Video, 2012).

During the second class session dedicated to the ADOS-2 training, students practiced rating ADOS-2 administrations. They watched videos obtained from the CASD of previously
recorded ADOS-2 administrations. All students viewed the same videos. Students took notes in
the note section of the protocol while viewing the administrations. Afterwards, students were
given an opportunity to code the performances seen on the videos according to the criteria
established on the coding forms. Tasks viewed were those outlined in the ADOS-2: Module 3
protocol and included: (a) construction task, (b) telling a story from a book, (c) description of a
picture, (d) conversation and reporting, (e) current work or school, (f) social difficulties and
annoyance, (g) emotions, (h) demonstration task, (i) cartoons, (j) break, (k) daily living, (l)
friends, (m) relationship and marriage, (n) loneliness, (o) plans and hopes and, finally, (p)
creating a story. Raters were not asked to complete the algorithm form. Practice protocols were
collected by the professor for instructional purposes and stored in his office.

Data Analysis

Item codes from the protocols obtained from the course professor were transferred to the
algorithm sheet by the researcher, and the item and total scores were recorded for analysis.

A G study was conducted as well as a D study. The G study was conducted to provide
information about the multiple sources of variance for each configuration, while the D study
estimated the reliability if the number of items or raters were changed. This offers information
that might be useful in designing the best possible application of the ADOS-2 for a particular
purpose.

Statistical analysis procedures resulting in Phi were employed using SPSS (IBM, 2013)
as described in Mushquash and O’Connor (2006). McWilliam and Ware (1994) describe the G
coefficient as the “ratio of the variance for persons to the average observed score variance for
persons and rater” (p. 36). In this G study, the persons or clients were considered the objects of
measurement as termed in G Theory and is identified as \( p \). The raters (master’s students) were
considered facet 1 in this study and are identified as \( r \). Facet 2 was items, identified as \( i \). The term \( x \) means “crossed with” (Brennan, 2003). The design of this study is \( p \times r \times i \) as described by Brennan (2003). In this study, sources of error included the person behaviors, observer (facet 1), items (facet 2), and interactions between the person and rater, person and items, rater and items, and a residual consisting of a combination of person, rater, and items coupled with error (which includes unmeasured facets affecting the measurement or random events). This design is a crossed design rather than a nested design since each person is rated by the same exact raters (Brennan, 2003). The G coefficient denoted as \( E \rho^2 \) and Phi \( \Phi \) (Brennan, 2003) served as an indication of how accurate the generalization was from the observed score based on the sample of the behavior to the universe score. The G coefficient or Phi reflects the portion of variability that is attributable to the construct being measured and no other variables. A D study was then completed with a reliability coefficient of .80 considered acceptable in order to reduce the potential for a Type II error.

**Summary**

This study is a reliability study using ANOVA techniques as understood through the conceptual framework of G Theory. This study made use of data from a master’s-level class in which students rated (coded) videos of ADOS-2 administration given in a university autism center. Descriptive statistics, G coefficients, and variance components were calculated to assess variance sources of the ADOS-2.

Chapter 3 discussed the research methodology including the background justifying the design. Research questions and hypotheses were addressed. Participants, sampling, and data analysis collection were outlined in relation to G Theory. Instrumentation and data collection to
be used in the current study were also explained. The following chapter will present the results for this study.
Chapter 4

Results

Purpose of Study

The primary purpose of this study was to investigate the reliability of the Autism Diagnostic Observation Schedule-Second Edition (ADOS-2). Specifically, this study was intended to determine the interrater reliability of the ADOS-2 with trainees. In addition, this study was intended to determine the proportion of variance attributed to raters and items and the interaction of the two. Because multiple sources of variance or error are suspected when using a behavioral observation rating instrument such as the ADOS-2, Generalizability Theory was chosen as the most appropriate framework for conducting the study. This theory allows for statistical analysis of multiple sources of variance simultaneously.

Data Analysis

The techniques in Generalizability Theory closely resemble ANOVA techniques but focus on variance components rather than F tests, as commonly seen in typical ANOVA analysis. Shavelson and Webb (1991) claim that G Theory does not use F values but rather interprets the magnitudes of estimated variance components. SPSS (IBM, 2013) was used for statistical analysis procedures using expected mean squares in random effects ANOVA (Brennan, 2003) as seen in an example by Mushquash and O’Connor (2006). Resulting data consisted of Phi or the G coefficient used for absolute decisions and variance components for each facet, interactions of facets, and remaining error.

Reliability

The overall reliability score or G coefficient was calculated for each configuration. The G coefficient is denoted as $E\rho^2$ (Brennan, 2003) and will serve as an indication of how accurate the generalization is from the observed score based on the sample of the behavior to the universe.
score. The G coefficient (a) reflects the portion of variability that is attributable to the construct being measured and no other variables, (b) reveals the dependability of the measure under the specified conditions, and (c) reflects the ratio of the universe score variance to the expected observed score variance (Brennan, 2003).

Shavelson and Webb (1991) suggest that researchers set their own standard for G coefficients. Acceptable reliability levels were presented by Bracken in 1987 (as cited by Naglieri and Chambers, 2009) as follows: (a) sub scales .80 or greater and (b) total test scores .90 or greater. Cicchetti (1994) suggested cut-offs for clinical significance of Kappa or intraclass correlations. He considered the reliability coefficient poor if it was below .40, fair if between .40-.59, good if between .60-.74, and excellent if above .75 (Cicchetti, 1994). The cut-off of .80 was used for the G coefficients in this study since no guidelines intended specifically for G coefficients were found. This score falls between the more conservative suggestion of .90 by Bracken (1987) and the less conservative score of .75 by Cicchetti (1994).

Research question 1: What is the overall reliability of the ADOS-2 when used by trainees with a simple and inexpensive training program?

Configuration 1. The first reliability analysis performed was on all 29 items, with scores of 3 left intact but 7s changed to zeros. The absolute G coefficient or Phi for this study was .880, indicating acceptable reliability across raters and items.

Configuration 2. A Phi coefficient was also determined for the same 29 items, but scores of 3 were reduced to a score of 2 and scores of 7 were converted to scores of 0. The absolute G coefficient or Phi for this study was .883, indicating acceptable reliability across raters and items.
Configuration 3. A G coefficient was also calculated on just the 14 items that are currently used to obtain scores for determining level of Autism using the ADOS-2. The Phi coefficient was .829.

**Variance components.**

**Configuration 1: 29 items with scores of 7 converted to 0 but 3s not collapsed.**

Results are provided in Table 1.

- Research question #2. What percentage of variance is caused by raters?
  This is the smallest source of variance ($\sigma^2 = .011$, 1.7 percent of variation).

- Research question #3. What percentage of variance is caused by persons?
  The greatest amount of variation other than error in this study can be explained by the persons or objects of measurement ($\sigma^2 = .113$). This accounts for 17.4 percent of the total variation.

- Research question #4. What percentage of variance is caused by items?
  The variable that contributed the second largest variance was that of the item facet only ($\sigma^2 = .098$; 15 percent of variation). This is a relatively large source of variance in this study.

- Research question #5. What percentage of variance is caused by the interaction of items and raters?
  The variance caused by the interaction of items and raters was .023, which is the second smallest source of variance, reflecting 3.5% of the total variance.

- Research question #6. What percentage of variance is caused by the interaction of items and persons?
  The third largest source of facet variance is the interaction of item and persons ($\sigma^2 = .087$ or 13.3 %). This is a medium-to-large relative proportion of variance in this study.
• Research question #7. What percentage of variance is caused by the interaction of persons and raters?

The variance caused by the interaction of raters and the persons ($\sigma^2 = .042; 6.4\%$) is a medium relative proportion of variance.

• Research question #8. What is the remaining percentage of variance that is caused by an interaction of person, rater, and item plus error?

The 42.6% of total variance that is attributable to measurement error and cannot be attributed to any particular facet of measurement or interaction of facets ($\sigma^2 = .277$) is a large relative variance.

**Configuration 2: 29 items with scores of 7 converted to 0 and 3s collapsed to 2s.**

Results are provided in Table 2.

• Research question #2. What percentage of variance is caused by raters?

Relative variance caused by raters alone in this configuration was the smallest source of variance at .014, or 2.4% of total variance.

• Research question #3. What percentage of variance is caused by persons (persons)?

The second largest source of variance is that caused by the person only facet ($\sigma^2 = .109, 18.4\%$).

• Research question #4. What percentage of variance is caused by items?

Variance caused by items is the third largest source of variance ($\sigma^2 = .90, 15.15\%$).

• Research question #5. What percentage of variance is caused by the interaction of items and raters?

The variance caused by the interaction of items and raters is another small source of variance at .023, or 3.8%.
• Research question #6. What percentage of variance is caused by the interaction of items and persons?

The variance caused by the interaction of items and person in this configuration is relatively large at .90, or 15.1% of total variance.

• Research question #7. What percentage of variance is caused by the interaction of persons and raters?

A medium source of variance ($\sigma^2 = .036, 6.1\%$) in this configuration was that caused by the interaction of person and rater.

• Research question #8. What is the remaining percentage of variance that is caused by an interaction of person, rater, and item plus error?

Error was responsible for the most variance at .240, or 40.2% of total variance.

Configuration 3: 14 items with scores of 7 converted to 0 and 3s collapsed to 2

Results are provided in Table 3.

• Research question #2. What percentage of variance is caused by raters?

Raters was the smallest source of variance at .032, or 5%.

• Research question #3. What percentage of variance is caused by persons?

Persons was the second largest source of variance at .119, or 19%.

• Research question #4. What percentage of variance is caused by items?

Items was responsible for the third largest source of variance at .073, or 11.7%.

• Research question #5. What percentage of variance is caused by the interaction of items and raters?

The interaction of items and raters was responsible for a relatively small amount of variance ($\sigma^2 = .02, 3.4\%$).
• Research question #6. What percentage of variance is caused by the interaction of items and persons?

The interaction of persons and items is a relatively medium source of variance at .059, or 9.5% of variance.

• Research question #7. What percentage of variance is caused by the interaction of persons and raters?

The interaction of persons and raters is a relatively medium source of variance at .046, or 7.4% of variance.

• Research question #8. What is the remaining percentage of variance that is caused by an interaction of person, rater, and item plus error?

Error was responsible for the most variance at .277, or 42.6% of total variance.

Research question 9: Do the variance proportions differ across various study configurations?

The difference in the proportion of variance is small for items, resulting in a difference of less than 1% of variance caused by items when item codings are collapsed. The difference in the variance caused by the interaction of items and persons is slightly larger with 1.8% more variance caused by this interaction. Overall, both changes are small, however.

The person-by-rater interaction and item-by-rater interaction changes are negligible, with less than 1% difference between collapsed and uncollapsed data. Raters alone continue to be responsible for just a small proportion of variance, although slightly more after item codings are collapsed (2.4%) than prior to collapsing data (1.7%). This is an indication of whether some raters code all items higher regardless of the person or if they see “more behavior” associated with ASD than others do (Shavelson and Webb, 1991)
Variance and proportion results can be seen in Table 3. Similar to the previous studies, error was the largest source of variance and persons of children were the second largest source of variance. The variance due to error increased from study 2 with 29 items and collapsed scores to study 3 with only 14 items and collapsed scores by 3.7%. The variance caused by error in study 1 was in between those two amounts at 42.6%. Like the 29-item study after collapsing item scores, the interaction of person and items was the next largest source of variance. The variance caused by the interaction of person and items was 3.4% smaller when only 14 items were analyzed. This is also smaller than the 13.3% variance caused by the person and item interaction on the 29-item study prior to collapsing scores. The item facet was the next largest source of variance on the 14-item study at 7.4%, which is much smaller than the variance caused by items in both the 29-item, collapsed score study and the 29-item, uncollapsed score study. The 7.4% variance caused by items in the third study is less than half of that in the first two studies. The variance caused by the interaction of items and raters was essentially unchanged in all three studies, ranging from 3.4 in study 3 to 3.8 in study 2. The variance caused by raters alone doubled from that in study 2 and almost tripled from that in study 1.

**Research Question 10: At what number of trainee raters and items is it estimated that acceptable reliability can be achieved?**

The G coefficient obtained in this study was reflective of the reliability if 7 raters were used, and this is not always the case when the ADOS-2 is used in common practice. It is questionable that if fewer raters were used the reliability would remain acceptable. This can be determined by looking at the D study results. See Table 4.
Criteria for D study acceptable reliability were set at .80 also. D study results reveal that, with the 14 items recommended for scoring, only 3 raters are needed to approach the acceptable .80 reliability. With just 2 raters and 10 items, .701 is the predicted reliability coefficient.
Chapter 5

Discussion

There have been various studies reporting adequate reliability for the Autism Diagnostic Observation System (Lord et al., 2000; Lord, 2002; Olson et al., 2012; Kim & Lord, 2012; Zander et al., 2015). Most have focused on the prior version of the ADOS as opposed to the current version, the ADOS-2. Techniques used for analysis in previous studies were mainly approached via Classical Test Theory analysis techniques. Raters used in most studies have been experienced and thoroughly trained via the training provided by the test creators. One more recent study (Zander et al., 2015) used Generalizability Theory and attempted to utilize less experienced raters, but there are no studies to date that investigate interrater reliability of trainees who have been given a brief training not provided by the test creators.

Overall, the reliability of the ADOS-2 in the current study was found to be acceptable. This provides evidence that a shorter and less expensive training program might suffice when attempting to train individuals to use this instrument. All three configurations yielded Phi coefficients above .80. There was a slight reduction in the reliability coefficient when the number of items was reduced to 14.

Further investigation into the sources of the variance reveals more specific information regarding the change in overall reliability. When investigating the specific sources of variance, it was discovered that the largest source of variance across all three configurations was that of the object of measurement or person. This indicates that scores differed greatly by person. This could be a reflection of the large variation in how behaviors are manifested in individuals on the Autism Spectrum.
When reducing the number of items to 14, as suggested by the test authors (Lord et al., 2012a) for scoring of the ADOS-2, variance caused by items was reduced from 15% on the first configuration—before scores of 3 were collapsed to 2—to 14% on the second configuration when 3s were collapsed to 2 and further to about 7.4% on the 14-item configuration. This supports the recommendation to both collapse scores of 3 to 2 and to minimize score variation caused by differences in items by throwing out some of the items. Although reducing the items to 14 does decrease the variability caused by items, 7.4% of variance is still a medium source of variation in this study. This variance score is reflective of how well one particular item or question on the ADOS-2 test will generalize to all possible items in the universe of possible items. It reflects differences in how items were scored independent of the person being rated. These results indicate that generalizing from a single item test to the universe of items would not be wise, especially if using all 29 items. We find lower internal consistency on test items when 29 items are used than when 14 items are used, indicating that the chosen 14 items are more representative of a single overall construct reflecting ASD. It is not an indication of whether 14 or 29 items are collectively a good measure of ASD.

When considering the variation in test item scores, it is important to consider multiple factors. One, the general domain or construct of Autism that the item is intended to investigate is important to understand. Another factor to consider is the difficulty of administering each individual task associated with the item. Also, the level of difficulty for the child to perform the task required for the rating of each item is also important. Cognitive, language, and motor levels are some examples of skills that would need to be considered for each task (Gotham, 2007).

Another difference seen when analyzing the proportions of variance in the various configurations was that the 14-item configuration indicates that 11.7% of variance was caused by
the person and item interaction but the 29-item configuration indicated 13.3% of variance caused by this interaction of facets when 3s were not collapsed and even more, 15.1 when scores of 3s were collapsed to 2. This means that the reduction of items to 14 reduced the amount that the rank ordering of persons differed across each item, averaging across raters. However, even with the reduced variance of the 14-item configuration, this is still reflective of a large proportion of variance. This variation is an indication that one child scored much higher than the other child on certain tasks as opposed to scoring higher across all tasks and that this was consistent across raters. This is not surprising given the heterogenetic nature of ASD characteristics present in individuals with ASD and the varying constructs assessed on the ADOS-2.

Next, the medium amount of variance caused by the interaction of raters and the person increased in the 14-item configuration to 9.5% from around 6% in the two 29-item configurations. The medium amount of variance caused by this interaction means that each rater may have somewhat rated a certain child more liberally than another. The relative standing of each child differed somewhat across raters. This is rater inconsistency and means that any one rater might also rate a particular person differently from another person. In other words, the \( r \times p \) variance indicates whether the rank order differs across raters, averaging over items (Kim, Park, and Kang, 2012). If all raters saw one person as higher than the other on all items, this would be a very low source of variance, but here it is a medium source of variance. Because of the low internal consistency of which behavior individual items are measuring, this interaction must be interpreted with caution. It is only natural that some behaviors will be higher in some persons while other items will be lower for that same person. This is reflective of the varying nature of how ASD behaviors present in individuals.
The proportion of variance caused by the interaction of items and raters remained fairly consistent despite changes in the configurations. This means that the raters' average ratings of a person from one item to the next was very consistent.

Although the reliability of the ADOS-2 was found to be acceptable in this study, when investigating the proportion of variance the largest was due to error. This means that the variance attributable to measurement error that cannot be attributed to any particular facet of measurement or interaction of facets is a large variance. This is not desirable since we cannot pull apart this factor to determine any more specific sources of variance. It means that the largest source of variance was caused by raters’ nonsystematic ratings across items and across children. It also indicates that factors other than those that we were able to identify in this study are greatly responsible for the amount of variance that we do see.

Typically, the desired source of the most variance would likely be that caused by the object of measurement or persons, since the assessment is intended to detect differences in these individuals. The highest source of variance in this study was that attributable to error. According to Goodwin and Goodwin (1991), the variance caused by the three-way interaction or error in a study of this type can be interpreted in a few ways. First, some of the error could be caused by the three-way interaction, meaning raters differing in a nonsystematic way across items and persons.

Second, it could be attributed to other factors that were not considered in the study. These possibilities cannot be separated in this type of study. This suggests that a large portion of the variability in item scores is due to other factors besides those that have been systematically accounted for, such as rater, items, or persons. This amount of variance caused by error varied only by 3.7% at the most in all three studies, which is minimal.
Of primary importance to the practical administration of the ADOS-2 and the focus of this study is the variance caused by the raters. Since, in practical administration, often multiple raters are used on any one administration, it is especially vital to be sure that the assessment yields similar results with any given possible rater. The raters used in this study can be generalized to any rater given the training program provided in the graduate class used. The amount of variance caused by raters was relatively small, contributing the least of the facets. Since the rater facet variance was small, one can conclude that the average scores among raters generalizes well to average scores of the entire universe of admissible raters (Shavelson & Webb, 1991). Kim, Park, and Kang (2012) describe the variation due to raters as rater severity and the variation due to the interaction between person and rater as rater inconsistency. Rater severity may differ between raters in that one rater might rate all persons more harshly than another rater. Since the rater variance is small, the stringency of the rater does not appear to have a great impact in this study. This could be an indication that the coding items are objectively defined and do not leave much room for interpretation, which is a desirable characteristic. Raters did not tend to rate certain items more stringently across the children than other items. There was not much variation in the average rating for each item across children. This provides evidence that the training program used for the raters did allow for dependability of the scoring among the raters.

The D study results provided some estimate of reliability, which is consistent with a previous study by Gotham (2008). Gotham’s study (2008) used multi-factor item response analysis to analyze the factor structure within each cell and was used to reorganize items. The reorganization involved reducing items to 14 for scoring purposes, which resulted in a higher reliability. Similarly, the D study in the current study demonstrates that when 14 items are used
the reliability is predicted to be .791. However, the current study provides further information estimating that if 14 items are used, then three raters would be needed to obtain the .791 predicted reliability. With only two raters and the recommended 14 items, reliability is predicted to be .727. With only one rater, it drops to .586.

A consideration in reducing items on the ADOS-2 administration would be deciding whether to reduce the rating items, the tasks presented to the children, or both. One concern with reducing the number of items or tasks is that there are certain rating items that ask the rater to think about the child’s behavior across all tasks (e.g., gestures, eye contact, and affect) (Lord et al., 2012a; p. 11). As such, it must be considered that reducing the tasks gives less opportunity to observe some of these behaviors. Some items are related to one task only (Lord et al., 2012a; p. 11) and eliminating that task might eliminate an opportunity to rate that behavior at all (e.g., joint attention). One possibility is that tasks could be manipulated to include more opportunities to rate more than one behavior within each task. However, depending on the nature of the behavior that the items are seeking to rate, it might be more or less difficult to detect more than one behavior at a time. For example, Initiation of Joint Attention in this assessment is specifically defined as “a 3 point gaze shift between an object and a person with the purpose of sharing interest or pleasure” (Lord et al., 2012; p. 22). This is a very specific criterion that might be difficult for the rater to detect if he or she is looking for multiple behaviors at the same time.

Reducing the number of items will reduce the behaviors being rated, which alone will provide a different result. Also, some items ask the rater to look for just one occurrence of a behavior (e.g., Shared Enjoyment in Interaction) and other items ask raters to rate how much (e.g., several or some) of a behavior (e.g., Amount of Social Overtures/Maintenance of Attention) they observe throughout all tasks presented. This factor may also influence ratings if
certain types of questions are used as opposed to others. Questions that look for some or several occasions of a behavior might be better at detecting behaviors that are more variable within the ASD population, and questions which ask for just one occurrence of a behavior might more easily detect behaviors that are more commonly seen in ASD. This would be important in considering which questions to eliminate or to combine in a shortened version of an ASD assessment.

Decreasing items to decrease the variance caused by this facet by 11.4% doesn’t necessarily move the assigned variance to the desired source, which we call person. In fact, persons as the source of variance only increased by .6% from study 2 to 3. Instead we see as a result of decreasing items that the variance in the combination of facets involving raters (rater only and rater by person) increased by 6.1% variance. The variance caused by facets and interactions involving items decreased by about 11.4%, but all of this lost variance was reassigned mostly to the rater involved facets or interaction of facets. It appears that having more items may have balanced out the variability of scores assigned by raters, allowing for less of a contribution to the total variance.

With this in mind, careful consideration must be taken when deciding if reducing the number of items has improved reliability of the ADOS-2 measure after all. In fact, the resulting Phi coefficients were somewhat consistent in study number 1 (.880) and study 2 (.883) but then the coefficient dropped to .829 in study 3 when we reduced items to 14. The collapsing of scores appears to have made minimal difference in the variances or reliability coefficient, yet we see more significant changes in variance when we reduce items with collapsed scoring to 14.

After reviewing current studies available that addressed reliability of the ADOS-2, none were found that investigated variance proportions of the ADOS instrument. When choosing the
theoretical framework for this study, Generalizability Theory was selected because it allowed for a more thorough investigation of reliability beyond what the studies by Gotham (2007, 2008) were able to provide. This study, using the G Theory analysis techniques, has allowed us to more thoroughly examine the effect of the item reduction that occurred in the newest version of the ADOS (ADOS-2) on other sources of variance. Although the studies by Gotham accurately depict the reduced variability caused by items, they neglect to provide any additional information as to how this change might affect the variance caused by raters and the interaction of facets. The current study is unique in that it allowed for simultaneous study of multiple facets. Accordingly, it has allowed us to closely examine the effects of the change on variance sources.

Brennan (2001) previously indicated that Item Response Theory (IRT) (the method used by Gotham (2007, 2008) was accommodating to single facet designs only but not multi-facet design. He further explained that the emphasis in G Theory is focused on viewing items as a “sampled “condition of measurement. He describes IRT as a scaling model as opposed to G Theory, which is a sampling model more appropriate for looking at absolute measurement decisions. In absolute measure decisions, the consideration of the relationship among the object of measurement is not of interest.

Glas (2012) suggests that when the assessment measure one is analyzing uses rating of behavior but consists of items with selected response formats (e.g., multiple choice), the best analysis might consist of a combination of G Theory and IRT techniques. The results of this study in a sense have allowed us to combine the results from both types of studies, that of Gotham’s study (2007) based on IRT and the current study based in G Theory. The IRT study allowed for the selection of the most reliable items, whereas the current G Theory results would not be able to tell us which specific items were causing the variance. However, the current study
also provided information that the IRT study (Gotham 2007) was not able to provide. For example, it gave us additional information about how the variance caused by other facets is affected when reducing the number of items.

**Limitations**

One limitation of this study is that only two facets were included in the study. Another possible factor might be occasions, even though the ADOS-2 is typically done in only one session. It would be of interest to determine if occasions would account for some level of variance. One can imagine that having observations from more than one occasion on the same person or object of measurement might affect the reliability in some way. Whereas having only one occasion might increase reliability, it might actually decrease validity. One occasion might not be measuring the behavior in question at all but actually another factor, such as the child’s mood.

The small number (2) of videos can also be viewed as a limitation to this study because it might not yield variance components that are stable (Brennan, 2001, p. 58). However, in Generalizability Theory, a large sample is not necessary since the theory postulates that any random sample is representative of the entire population or universe. With this theory in mind, even just one video should be sufficient. Generalizability does not attempt to compare multiple samples but rather uses one sample to represent any random sample. G Theory is intended to be used for a single observation per cell. Brennan (2000) points out that the very nature of a G study is the reason that most G studies consist of single observation per cell. He explains that to replicate within cell observations would simply provide interchangeable findings and would not provide any additional information with regard to a particular facet. In this study, the measure being studied (ADOS-2) uses multiple raters looking at the same person at the same time. The
facets of rater and item are the facets of most concern in this study; therefore, a minimum number of persons is needed to generalize to the universe of admissible raters. The score given by a particular rater using a particular set of items is used as an indication of the average score that would be obtained from many raters using many items.

Another limitation is that the item facet is fixed in the current study, meaning that the items are exhaustive of all possible items in the universe of generalization. According to Brennan (2011), using fixed facets creates a limitation in a study because it lowers error variance and increases coefficients. Brennan (2011) believes that this is done at the expense of narrowing the researcher’s interpretations. Generalizability Theory assumes random sampling, and although Gage and Prykanowski (2014) report that this assumption is often violated when using G Theory, it still must be considered as a limitation since items were all obtained from one assessment and raters were all from one graduate class.

A final limitation is that the study focuses on only one module and results may not be reflective of the interrater reliability of other modules with trainees.

Future Study

Future studies should attempt to divide items by construct or domain and determine if variance changes, rather than considering all items on the assessment.

In consideration of the limitation caused by considering the item facet fixed rather than random, a future study could consider all facets random. If the item facet is considered random, then the universe of generalizability would be all possible items investigating ASD rather than the focus remaining on only the ADOS-2 items as the entire universe of generalizability. The purpose of the current study was to investigate interrater reliability of the ADOS-2, therefore, items were considered fixed, but another possible question to be considered is how well the
particular sample of items on the ADOS-2 would generalize to all possible items assessing traits of ASD. This would be possible if the item facet were considered random.

According to the EDUG Users Guide (Swiss Society for Research in Education, 2010):

You can, in a test situation, treat the facet Questions as fixed. This would mean that you had no intention of generalizing the students’ test performance to a wider question domain. The only questions of interest to you are those in the test itself, corresponding to a particular school assignment, for instance. On the other hand, you might indeed want to generalize to a wider question domain, recognizing that the questions that you happen to have included in your test are merely a sample of many more such questions that you could have used, whether the questions themselves already exist or are yet to be developed (p. 5).

More evidence that both facets in this study could be treated as random is found in an article by Webb, Shavelson, and Haertal (2006). They report that, according to De Finetti, “When the levels of a facet have not been sampled randomly from the universe of admissible observations but the intended universe of generalization is infinitely large, the concept of exchangeability may be invoked to consider the facet as random” (as cited in Webb, Shavelson, and Haertel, 2006, p.24). Webb, et al., (2006) further state that, “Even if conditions of a facet have not been sampled randomly, the facet may be considered to be random if conditions not observed in the G-study are exchangeable with the observed conditions” (p. 24).

In consideration of the amount of variance caused by items, a study examining the variance caused by items when splitting the items into subtests or by construct intended to be measured would be useful.
Summary

Overall, reliability of the ADOS-2 was found to be acceptable among trainees, following a short training. Minus the high proportion of variance caused by error or undifferentiated factors, the proportion of variance is somewhat desirable in that the person contributes the most variance, followed by items that might be unavoidable in the diagnosis of ASD, which includes so many different constructs within one definition. The interactions of facets are responsible for a smaller portion of variance, and raters are responsible for the least variance, indicating well-defined item coding categories.

The current study serves to elaborate upon previous studies that explored ways of increasing reliability of the ADOS-2 by examining not only the overall reliability of the measure but also more closely looking at specific sources of the variance. In particular, this study looks at reliability of trainee ratings on the ADOS-2 and indicates that, given a brief training, ratings are consistent. It suggests that only three raters are needed with the 14 items currently used in scoring on Module 3 in order to achieve reliability close to .80.
References


American Psychiatric Association. (2000). *Diagnostic and statistical manual of mental*


Baron-Cohen, S., Allen, J., & Gillberg, C. (1992). Can Autism be detected at 18 months?: The needle, the haystack and the CHAT. British Journal of Psychiatry,


reliability: Key concepts, approaches, and applications. Research in Social and Administrative Pharmacy, 9, 330–338.


Corporation.

IDEA 2004 Part 300 Subpart A. Retrieved from

http://idea.ed.gov/explore/view/p/%2Croot%2Cregs%2C300%2CA%2C300%252E8%2C
F. R. Volkmar, R. Paul, & S. J. Rogers (Eds.). Handbook of Autism and Pervasive
Developmental Disorders, Assessment, Interventions, and Policy, Volume 2: Assessment,

Parent Rating Scales Preschool Form*. Digital Archive at Georgia State University.


ratings correlate with symptomatology and IQ among individuals with high-functioning
autism spectrum disorders. *Journal of Autism and Developmental Disorders, 40*(4), 416-
423. doi:10.1007/s10803-009-0911-4

Kim, S., & Lord, C. (2012). Combining information from multiple sources for the diagnosis of
Autism spectrum disorders for toddlers and young preschoolers from 12 to 47 months of
7610.2011.02458.x


Reynolds, C. R., & Bigler, E. D. (1994) Test of Memory and Learning (TOMAL). Austin, TX: Pro-Ed..


Developmental Disorders, 10, 91-103.


Appendix

Definition of Terms in Generalizability Theory

Condition The levels of a facet (e.g., Rater 1, Rater 2).

Decision (D) Study Uses information from a G study to design a measurement procedure that minimizes error for a particular purpose.

Facet A characteristic of a measurement procedure--such as task, occasion, observer--that is defined as a potential source of measurement error.

Generalizability (G) Study A study specifically designed to provide estimates of the variability of as many possible facets of measurement as economically and logistically feasible considering the various uses a test might be put to.

Universe of Admissible Observations All possible observations a test user would consider acceptable substitutes for the observation in hand.

Universe of Generalization The conditions of a facet to which a decision maker wants to generalize.

Universe Score (denoted as \( \mu_p \)) The expected value of a person’s observed scores across all observations in the universe of generalization (analogous to a person's "true score" in classical test theory).

Variance Component The variance of an effect in a G study.
Table 1

29 items-Variance components before collapsing item scores

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>MS</th>
<th>Variance</th>
<th>Percentage of Total Variance</th>
<th>Relative Size of Variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>R</td>
<td>6</td>
<td>2.177</td>
<td>.011</td>
<td>1.7%</td>
<td>Small</td>
</tr>
<tr>
<td>i x r</td>
<td>168</td>
<td>.323</td>
<td>.023</td>
<td>3.5%</td>
<td>Small</td>
</tr>
<tr>
<td>p x r</td>
<td>6</td>
<td>.042</td>
<td>.042</td>
<td>6.4%</td>
<td>Medium</td>
</tr>
<tr>
<td>p x i</td>
<td>28</td>
<td>.886</td>
<td>.087</td>
<td>13.3%</td>
<td>Medium to large</td>
</tr>
<tr>
<td>I</td>
<td>28</td>
<td>2.303</td>
<td>.098</td>
<td>15%</td>
<td>Large</td>
</tr>
<tr>
<td>P</td>
<td>1</td>
<td>25.126</td>
<td>.113</td>
<td>17.4%</td>
<td>Large</td>
</tr>
<tr>
<td>p x i x r</td>
<td>168</td>
<td>.277</td>
<td>.277</td>
<td>42.6%</td>
<td>Very large</td>
</tr>
</tbody>
</table>

Note. p = objects of measurement, r = raters, i = items, x = crossed with
### Table 2

29 items-Variance components after collapsing item scores

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>MS</th>
<th>Variance</th>
<th>Percentage of Total Variance</th>
<th>Relative Size of Variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>R</td>
<td>6</td>
<td>2.149</td>
<td>.014</td>
<td>2.4%</td>
<td>Small</td>
</tr>
<tr>
<td>i x r</td>
<td>168</td>
<td>.286</td>
<td>.023</td>
<td>3.8%</td>
<td>Small</td>
</tr>
<tr>
<td>p x r</td>
<td>6</td>
<td>1.290</td>
<td>.036</td>
<td>6.1%</td>
<td>Medium</td>
</tr>
<tr>
<td>p x i</td>
<td>28</td>
<td>.870</td>
<td>.084</td>
<td>15.1%</td>
<td>Large</td>
</tr>
<tr>
<td>I</td>
<td>28</td>
<td>2.085</td>
<td>.90</td>
<td>14%</td>
<td>Large</td>
</tr>
<tr>
<td>P</td>
<td>1</td>
<td>24.140</td>
<td>.109</td>
<td>18.4%</td>
<td>Large</td>
</tr>
<tr>
<td>p x i x r +error</td>
<td>168</td>
<td>.240</td>
<td>.240</td>
<td>40.2%</td>
<td>Very large</td>
</tr>
</tbody>
</table>

Note. p = objects of measurement, r = raters, i = items, x = crossed with
Table 3
14 items—Variance components with collapsed item scores

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>MS</th>
<th>Variance</th>
<th>Percentage of Total Variance</th>
<th>Relative Size of Variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>R</td>
<td>6</td>
<td>2.029</td>
<td>0.032</td>
<td>5.0%</td>
<td>Small</td>
</tr>
<tr>
<td>i x r</td>
<td>78</td>
<td>.316</td>
<td>0.021</td>
<td>3.4%</td>
<td>Small</td>
</tr>
<tr>
<td>p x r</td>
<td>78</td>
<td>1.104</td>
<td>0.046</td>
<td>9.5%</td>
<td>Medium</td>
</tr>
<tr>
<td>p x i</td>
<td>13</td>
<td>.787</td>
<td>0.059</td>
<td>11.7%</td>
<td>Large</td>
</tr>
<tr>
<td>I</td>
<td>13</td>
<td>.046</td>
<td>0.073</td>
<td>7.4%</td>
<td>Medium</td>
</tr>
<tr>
<td>P</td>
<td>1</td>
<td>13.270</td>
<td>.119</td>
<td>19%</td>
<td>Large</td>
</tr>
<tr>
<td>p x i x r + error</td>
<td>78</td>
<td>.274</td>
<td>.274</td>
<td>43.9%</td>
<td>Very large</td>
</tr>
</tbody>
</table>

Note. p = objects of measurement, r = raters, i = items, x = crossed with
### Table 4
Comparison of facet variance across configurations

<table>
<thead>
<tr>
<th></th>
<th>Configuration 1 (29 items)</th>
<th>Configuration 2 (29 items collapsed)</th>
<th>Configuration 3 (14 items)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>R</strong></td>
<td>1.7%</td>
<td>2.4%</td>
<td>5.0%</td>
</tr>
<tr>
<td><strong>i x r</strong></td>
<td>3.5%</td>
<td>3.8%</td>
<td>3.4%</td>
</tr>
<tr>
<td><strong>p x r</strong></td>
<td>6.4%</td>
<td>6.1%</td>
<td>9.5%</td>
</tr>
<tr>
<td><strong>p x i</strong></td>
<td>13.3%</td>
<td>15.1%</td>
<td>11.7%</td>
</tr>
<tr>
<td><strong>I</strong></td>
<td>15%</td>
<td>14%</td>
<td>7.4%</td>
</tr>
<tr>
<td><strong>P</strong></td>
<td>17.4%</td>
<td>18.4%</td>
<td>19%</td>
</tr>
<tr>
<td><strong>p x i x r + error</strong></td>
<td>42.6%</td>
<td>40.2%</td>
<td>43.9%</td>
</tr>
</tbody>
</table>

Note. p = objects of measurement, r = raters, i = items, x = crossed with
Table 5

D-Study G Coefficients

<table>
<thead>
<tr>
<th></th>
<th>0.000</th>
<th>1.000</th>
<th>2.000</th>
<th>3.000</th>
<th>4.000</th>
<th>5.000</th>
<th>6.000</th>
<th>7.000</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.000</td>
<td>.226</td>
<td>.331</td>
<td>.392</td>
<td>.432</td>
<td>.459</td>
<td>.480</td>
<td>.496</td>
<td></td>
</tr>
<tr>
<td>2.000</td>
<td>.338</td>
<td>.469</td>
<td>.538</td>
<td>.581</td>
<td>.610</td>
<td>.632</td>
<td>.648</td>
<td></td>
</tr>
<tr>
<td>3.000</td>
<td>.405</td>
<td>.544</td>
<td>.614</td>
<td>.657</td>
<td>.686</td>
<td>.706</td>
<td>.721</td>
<td></td>
</tr>
<tr>
<td>4.000</td>
<td>.449</td>
<td>.591</td>
<td>.661</td>
<td>.703</td>
<td>.731</td>
<td>.750</td>
<td>.765</td>
<td></td>
</tr>
<tr>
<td>5.000</td>
<td>.480</td>
<td>.624</td>
<td>.693</td>
<td>.734</td>
<td>.760</td>
<td>.779</td>
<td>.793</td>
<td></td>
</tr>
<tr>
<td>6.000</td>
<td>.504</td>
<td>.648</td>
<td>.716</td>
<td>.756</td>
<td>.782</td>
<td>.800</td>
<td>.814</td>
<td></td>
</tr>
<tr>
<td>7.000</td>
<td>.522</td>
<td>.666</td>
<td>.733</td>
<td>.772</td>
<td>.798</td>
<td>.816</td>
<td>.829</td>
<td></td>
</tr>
<tr>
<td>8.000</td>
<td>.537</td>
<td>.680</td>
<td>.747</td>
<td>.785</td>
<td>.810</td>
<td>.828</td>
<td>.841</td>
<td></td>
</tr>
<tr>
<td>9.000</td>
<td>.549</td>
<td>.692</td>
<td>.758</td>
<td>.796</td>
<td>.820</td>
<td>.837</td>
<td>.850</td>
<td></td>
</tr>
<tr>
<td>10.000</td>
<td>.559</td>
<td>.701</td>
<td>.767</td>
<td>.804</td>
<td>.828</td>
<td>.845</td>
<td>.858</td>
<td></td>
</tr>
<tr>
<td>11.000</td>
<td>.567</td>
<td>.709</td>
<td>.774</td>
<td>.811</td>
<td>.835</td>
<td>.852</td>
<td>.864</td>
<td></td>
</tr>
<tr>
<td>12.000</td>
<td>.574</td>
<td>.716</td>
<td>.780</td>
<td>.817</td>
<td>.841</td>
<td>.857</td>
<td>.870</td>
<td></td>
</tr>
<tr>
<td>13.000</td>
<td>.580</td>
<td>.722</td>
<td>.786</td>
<td>.822</td>
<td>.846</td>
<td>.862</td>
<td>.874</td>
<td></td>
</tr>
<tr>
<td>14.000</td>
<td>.586</td>
<td>.727</td>
<td>.791</td>
<td>.827</td>
<td>.850</td>
<td>.866</td>
<td>.878</td>
<td></td>
</tr>
</tbody>
</table>

Note. Columns=number of raters, rows=number of items
Curriculum Vitae

Dorothy R. Parriott
DSchoener1@optonline.net
10640 Snow Lake St
Las Vegas, Nevada 89179
201-572-0918

Education
Anticipated May 2016 Doctor of Philosophy in Educational Psychology, Specialization in School Psychology
University of Nevada, Las Vegas

October 2006 NJ Expedited Certification in Educational Leadership
NJ Principals and Supervisors Association
Foundation for Educational Leadership

January 2001 Master of Arts in Educational Psychology
August 2001 Professional Diploma in School Psychology
Kean University

June 1996 Bachelor of Arts in Psychology
University of Rhode Island

Certifications
November 2008 NV School Psychologist
March 2007 NJ Provisional Principal Certificate
March 2004 NJ Educational Supervisor Certificate
February 2001 NJ School Psychologist Certificate

Teaching Experience
Sept. 2010 - Sept. 2011 Online Instructor/Consultant,
Sierra Nevada College, Las Vegas Nevada
Educational Leadership Program
Duties: Online facilitation of educational leadership courses and course development consultation including the following topics: APA Style Writing, School Finance, System Familiarity, Program Analysis, and Curriculum Design

Sept. 2011 - Dec. 2011 Teaching Practicum
University of Nevada, Las Vegas
Assisting and teaching lessons for an Introductory Graduate level School Psychology course. Topics covered: Role of a School Psychologist, Self-Assessment, Eligibility Criteria, Understanding and Choosing Assessments (reliability and validity, purpose, administration)
Clinical Experience:
August 2013 - present
Clinical Internship
Clark County School District
Supervised work as a school psychologist by a licensed clinical psychologist, Professional development experiences to include clinical and neuropsychological techniques as well as work on an outsource Childfind/Early Childhood team.

Sept. 2012 - Dec 2012
Clinical Practicum
University of Nevada, Las Vegas
Center for Autism Spectrum Disorders
Team evaluation of adult clients for possible Autism diagnosis. Conducted psychosocial and academic evaluations, Clinical style report writing and reporting results at team meeting.

March 2009 - August 2009
Psychosocial Rehabilitative Counseling and Planning,
Group home for adolescents involved in the juvenile justice system

School Setting Experience:
August 2009 to present
School Psychologist,
Clark County School District, NV
Supervisors: Rick Shaw/Keri Altig
Duties: Chair of the Response to Intervention school level team, Analyzing formative and summative assessments, Evaluating reading and math interventions for empirical foundation and validity of delivery, Measuring level and rate of progress to determine eligibility under the Learning Disability category, Multidisciplinary evaluations including cognitive and academic testing, Conducting eligibility meetings, Working with a diverse population, Designing behavioral plans, Conducting ChildFind evaluations of preschool aged children.

Aug. 2006 - Aug.2008
Supervisor of Special Education,
Rockaway Borough Schools, NJ
Supervisor: Emil Suarez
Duties: Budget planning, Presentations, State reporting, Scheduling, Staff evaluations and professional improvement plans, Program planning and evaluation, Interviewing and hiring, Research and legal case analysis, Member of Strategic Planning Action Committee.

Principal Intern,
Byram School District, NJ
Supervisor: Jack Leonard
Duties: Response to Intervention (RTI) Plan Research,
Intervention and Referral Services committee procedures, School evacuation plan, State assessment planning and training, Master scheduling, Curriculum development planning, Student discipline, Design and presentation of paraprofessional training series.

Child Study Team Coordinator,
Byram School District, NJ
Supervisor: Jennifer DeSaye
Duties: Paraprofessional Evaluations, Oversight of IEP Operations, Special services program monitoring, Special Education reporting, Staff interviewing and orientation, Website maintenance, Update of job descriptions.

Sept 2004-Dec. 2004
Supervisor of School Psychologist Counseling Practicum Student,
Byram School District, NJ
Duties: Regularly scheduled supervision meetings with student, Observations of student counseling, guidance on counseling strategies, ethics, and consultation techniques.

School Psychologist,
Byram Township School District, NJ
Supervisor: Jennifer DeSaye
Duties: Assistant Coordinator of Crisis Management Team Steering Committee, Major contributor to Anti Bullying/Anti-Harassment policy, Supervision of Counseling Practicum Student, 504 Coordination and Case Management, Consultation with parents, teachers and administrators, Case management of a variety of classifications and placements (PSD, Autistic, SLD, ABA, self-contained, resource center, inclusion), Psychological evaluations (cognitive, emotional, assessments, ADHD), Individual and group counseling, social skills, Functional Behavior Assessments, In-service presentations, Coordinate “No Name Calling Week,” School Level Planning Team member, Conflict and Resolution Committee member, State Monitoring.

Research/Presentations
Dissertation in process
Using Generalizability Theory to Investigate Sources of Variance of the Autism Diagnostic Observation System-II with Trainees

Poster Presentation,
National Academy of Neuropsychology
Adapting Online Neuropsychological Scales for Online Administration: Study 2
Fall 2012

Master’s Thesis: Teacher Understanding and Opinion of Conjoint Behavioral Consultation
December 2001

110
Presenter,
30 school staff members
Byram School District

"How to Decide if an Intervention is Supported by Research Based Evidence”

Presenter,
150 school district staff
Byram School District

“Asperger Syndrome”

Presenter,
150 school district staff
Byram School District

“Working with Your Classroom Aides for Inclusion”

Presenter,
Rockaway Borough
Board of Education and Community

“Initiative on Autism Grant Application”

Presenter,
Rockaway Borough
Board of Education and Community

“Inclusive Preschool: Planning, Implementation and Evaluation”

Presenter,
Rockaway Borough
Board of Education and Community

“Medicaid Reporting for Related Services Reimbursement”

Accomplishments

2006-2008 Recipient and Coordinator of Initiative on Autism grant for $349,000

2006-2008 Planning, Implementing and Evaluating an Inclusive Preschool Program

2004-2006 President, Sussex-Warren Association of School Psychologists

1996 Recipient, Senior Contributions Award at University of Rhode Island
## References

### Academic

<table>
<thead>
<tr>
<th>Name</th>
<th>Position</th>
<th>Institution</th>
<th>Contact Information</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dr. Tara Raines</td>
<td>Professor, School Psychology</td>
<td>University of Nevada, Las Vegas</td>
<td><a href="mailto:Tara.Raines@du.edu">Tara.Raines@du.edu</a></td>
</tr>
<tr>
<td>Dr. Scott Loe</td>
<td>Professor, School Psychology</td>
<td>University of Nevada, Las Vegas</td>
<td><a href="mailto:Scott.Loe@unlv.edu">Scott.Loe@unlv.edu</a></td>
</tr>
<tr>
<td>Dr. Don Blagg</td>
<td>Doctorate level Intern Supervisor</td>
<td>Clark County School District</td>
<td>702-799-0930</td>
</tr>
</tbody>
</table>

### Professional

<table>
<thead>
<tr>
<th>Name</th>
<th>Position</th>
<th>Institution</th>
<th>Contact Information</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bob Weires</td>
<td>Director of Psychological Services</td>
<td>Clark County School District</td>
<td>702-799-7465</td>
</tr>
<tr>
<td>Melissa Baker</td>
<td>Assistant Principal</td>
<td>Jydstrup Elementary School</td>
<td>702-799-8140</td>
</tr>
<tr>
<td>Kim Sorenson</td>
<td>Assistant Principal</td>
<td>Steele Elementary School</td>
<td>702-799-2201</td>
</tr>
</tbody>
</table>