A PSYCHOMETRIC ANALYSIS OF AN ALTERNATE ASSESSMENT FOR STUDENTS WITH MULTIPLE DISABILITIES

A DISSERTATION

SUBMITTED TO THE FACULTY

OF

THE GRADUATE SCHOOL OF APPLIED AND PROFESSIONAL PSYCHOLOGY

OF

RUTGERS,

THE STATE UNIVERSITY OF NEW JERSEY

BY

SAMANTHA JODIE SCHULMAN

IN PARTIAL FULFILLMENT OF THE

REQUIREMENTS FOR THE DEGREE

OF

DOCTOR OF PSYCHOLOGY

NEW BRUNSWICK, NEW JERSEY OCTOBER 2017

APPROVED: _____
Ryan Kettler, PHD, NCSP

_____
Linda Reddy, PHD

DEAN: _____
Francine Conway, PHD

Abstract

Although students with severe and multiple disabilities have participated in state-level alternate assessments for years, research has shown that their educational achievement has not been measured in a psychometrically sound way (Kettler et al., 2010; Elliott, Compton, & Roach, 2007). Therefore, schools cannot meaningfully evaluate these students' progress. Professionals within PG Chambers, a charter school in New Jersey serving students with severe disabilities, recognized the need for an evaluation tool that could quantify the difference that their school program makes in the lives of their students. To assess both therapeutic and academic progress for students with multiple disabilities, in 2008 a transdisciplinary team created the PG Chambers School Outcomes Measurement Tool (PGS-OMT). By gaining insight into three sets of psychometric properties of the PGS-OMT, this study aimed to determine how meaningfully students at PG Chambers are being evaluated. First, the reliability of the PGS-OMT was evaluated through analyses of internal consistency and cross-informant agreement. Reliability analyses using Cronbach's alpha indicated that scores demonstrated excellent internal consistency, and Pearson correlations indicated that scores demonstrated high cross-informant agreement. Second, the internal structure of the PGS-OMT was assessed through correlations among the subscales, confirmatory factor analysis (CFA), and exploratory factor analysis (EFA). The internal structure analyses were not conclusive. CFA yielded results that were difficult to interpret because of high factor loadings for both the uncorrelated and the correlated three-factor model coupled with poor fit indices. EFA yielded findings supportive of the three-factor model rejected by the CFA. Strengths and weaknesses of the one-, two-, and three-factor models were therefore discussed. Third, change in scores from 2013 to 2014 of the PGS-OMT was assessed using descriptive statistics, paired samples t-tests, and Cohen's d. Change in scores analyses

yielded significant t-tests, indicating that students at PG Chambers are generally improving from year to year on some domains measured by the PGS-OMT.

Dedication

To my mom, Faye Lichbach.

Your everlasting love helped me complete this. Your memory will forever inspire me.

Acknowledgements

Thank you to my father, Mark Lichbach, for granting me endless support throughout this process. I am always learning from you. You are truly a role model for a parent and an academic. Thank you to my sweet husband, Jacob Schulman, for your continuing support and for being proud of me no matter what.

Thank you to my dissertation chair, Ryan Kettler, for giving me the idea to pursue this dissertation and for helping me organize my work. Your feedback has helped me become a better researcher and writer. And thank you to my committee member, Linda Reddy, for your opinions and ideas that enabled me to enhance my work.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter I

Literature Review

States and districts are mandated to hold schools accountable for students meeting basic educational standards (Browder et al., 2003). In 2002, the federal legislation No Child Left Behind (NCLB) required large-scale assessments for the purposes of establishing the goals and having the high standards that can improve student outcomes in education (NCLB, 2002). In 2015, the Every Student Succeeds Act (ESSA) replaced NCLB and had similar goals (ESSA, 2015). To assess their Adequate Yearly Progress (AYP), the Individuals with Disabilities Education Act of 1997 (IDEA) and Individuals with Disabilities Education Improvement Act of 2004 (IDEIA) requires that all students be included in state and district educational assessments [6l2(a)(17)(A)] (Browder et al., 2003, Towles-Reeves, Kleinert, Muhomba, 2009). NCLB, ESSA, and IDEA were enacted to ensure that Americans take responsibility for the quality of education. These legislations encourage educators to "channel more time, resources, and attention to minority students, poor students, and students with special educational needs" by requiring that "their performance [be] made public and schools… held accountable for results" (Jennings & Beveridge, 2009, p. 153).

The reauthorization of the Elementary and Secondary Education Act (ESEA) in 1994 supported state efforts to institute challenging standards, develop aligned assessments, and create accountability procedures for school districts (U.S. Department of Education, 1999). Title I of ESEA mandated improving the academic achievement of underprivileged students, and indicated certain characteristics that state assessment systems must have. Assessments must be aligned with state content and performance standards in at least mathematics and literacy, and must

involve multiple approaches that assess complex thinking skills. In addition, assessments must be administered annually to students in grades 3-8 and once in high school. The ESSA sanctioned more flexible testing procedures in terms of how and when the assessments are administered. All students in the grades being assessed must participate in the assessments, and students who have disabilities and who are Limited English Proficient (LEP) must receive any appropriate accommodations. Assessments must meet nationally recognized professional and technical standards. Finally, the assessments must provide individual student reports that highlight achievement on student performance standards (U.S. Department of Education, 1999).

Although all students are required to participate in these large-scale assessments, some students with disabilities (SWDs) are unable to meaningfully participate in the general assessment system. In the mid 1990's, when these assessments became commonplace, SWDs were therefore excluded from participating in large-scale state and district assessments (Erickson, Thurlow, Thor, Seyfarth, 1996). Guidelines relating to the exclusion of SWDs varied from state to state and the estimated rate of participation of SWDs were often low (Koretz & Barton, 2004). Educators also faced incentives to exclude students who may score poorly from participating in these assessments. The concern was that if students with severe disabilities were excluded from accountability systems they would also be left out of policy decisions.

Defining Alternate Assessment

As a result of IDEA 1997 and IDEIA 2004, students who are unable to participate in the general assessment system, even with appropriate accommodations, still need to be included in the assessment. Alternate assessment is defined as "Data collection procedures used in place of the typical assessment when students cannot take standard forms of assessment" (Ysseldyke &

Olsen, 1997, p. 1). The ESSA requires each states plan to outline how their alternate assessments incorporate a universal design for learning to the maximum extent possible (ESSA, 2015).

There is disagreement in the literature regarding the connection between 'alternate assessment' and 'testing accommodations.' Testing accommodations are "alterations to tests' standard administration procedures that are made to overcome individuals' functional impairments, in order to increase the validity of inferences that can be made from the resulting scores" (Kettler, 2012, p. 53). Some maintain that accommodations refer to "any change or adjustment to standard testing procedures or materials" (Thurlow, Lazarus, Thompson, Morse, 2005, p. 235), and they argue that alternate assessments are a subcategory of testing accommodations. Others maintain that testing accommodations are only utilized if students' instructional goals parallel those of the general curriculum, and therefore students who have instructional goals that differ from those of the general curriculum should participate in alternate assessment (Destefano, Shriner, Lloyd, 2001).

According to the U.S. Department of Education, alternate assessments based on alternate achievement standards (AA-AAS) cover "a narrower range of content (e.g., cover fewer objectives under each content standard) and reflect a different set of expectations in the areas of reading/language arts, mathematics, and science than do regular assessments or alternate assessments based on grade-level achievement standards." (U.S. Department of Education, 2005, p.16). Further, AA-AAS "must establish alternate achievement standards through a documented standards-setting process; the assessments based on alternate achievement standards must yield separate results for reading/language arts, mathematics…and science" (U.S. Department of Education, 2005, p.16).

Even though alternate achievement standards may differ in complexity from grade-level achievement standards, they must still be linked to grade level content. For example, one reading standard is that students demonstrate that they can use reading to learn, communicate, and solve problems independently (i.e., English language arts standard for reading informational text, grades 5-7). SWDs can work toward this standard in a vocational context by having students use photographs to complete the routine of cleaning a hotel room, by having the students rated on each step of the sequence (e.g., knock on door and say "housekeeping," use a key to enter the room, look at the card and complete tasks shown) (Thompson, Quenemoen, Thurlow, Ysseldyke, 2001).

According to the U.S. Department of Education, alternate assessments must assess for at least three levels of fulfillment (e.g., basic, proficient, advanced). In addition, alternate assessments must include descriptions of competencies associated with each level of fulfillment, include cut scores that differentiate among achievement levels, and provide a rationale for how each achievement level is determined. These standards would then be considered during the Department's peer review of each state's assessment system (U.S. Department of Education, 2005a). In addition, alternate assessments must have clear structure, parameters for determining which students participate, defined scoring criteria and procedures, and a report communicating student performance in terms of academic achievement standards (Kettler et al., 2010).

Characteristics of Students who Participate in AA-AAS

One percent of the total student population is qualified to participate in AA-AAS (ESSA, 2015). SWDs taking AA-AAS typically are classified in the following disability categories: autism, cognitive impairment, and multiple-disabilities (U.S. Department of Education, 2003). Not all students classified under these disability categories receive alternate assessments, and this

is not an exhaustive list of classifications. According to IDEA, multiple disabilities refers to "concomitant impairments (such as mental retardation-blindness or mental retardation-orthopedic-impairment), the combination of which causes such severe educational needs that they cannot be accommodated in special education programs solely for one of the impairments. Multiple disabilities does not include deaf-blindness." (U.S. Department of Education, 2006, p. 46756).

The U.S. Department of Education referred to students with severe disabilities as "the small number of students who are (1) within one or more of the existing categories of disability under IDEA (2) whose cognitive impairments may prevent them from attaining grade-level achievement standards, even with the very best instruction" (U.S. Department of Education, 2005, p.23). Students with severe disabilities have significant deficits in intellectual functioning and adaptive behavior (Lowery, Drasgrow, Renzaglia, & Chezan, 2007). Intellectual functioning refers to mental ability, including the ability to problem solve, reason, and comprehend abstract ideas. Adaptive behavior is student functioning in daily life with respect to practical life skills, independence, and coping skills. Some students with severe disabilities also have accompanying medical conditions and therefore require medications and various forms of therapy (Lowery et al., 2007).

Purposes of Alternate Assessment

Browder et al. (2003) conducted a review of the alternate assessment literature and explained that alternate assessments have several benefits. One possibility is that alternate assessment leads to greater consideration of students with severe disabilities in school and state policy decisions. In order to obtain an accurate picture of the educational system all students must be included in the assessment process. Another possibility is that expectations for students

with disabilities will begin to increase. Including these students in large-scale assessments holds the state and district accountable for their education. Alternate assessment also ensures that all students have access to the same curriculum and to be assessed using the same district standards. Some argue that the primary benefit of alternate assessment is improved instruction in the classroom. In short, teachers could use alternate assessment scores to enhance learning. Specifically, if alternate assessment performance is linked to IEP goals then it is feasible to expect students to achieve high levels of proficiency (Browder et al., 2003).

Approaches to AA-AAS

In the early development of alternate assessment, few states actually developed measures that aligned with curricular standards. Some only incorporated functional domains (as opposed to academic domains) in their assessments (Ysseldyke & Olsen, 1999). Brown et al. (1979) introduced the term "functional" to refer to a "new curriculum model that promoted community access by targeting skills needed to function in daily life" (Browder et al., 2004, p. 212). Brown et al. (1979) described four functional domains – community, recreation, domestic, and vocational – that became the new content areas for curriculum and the focus of alternate assessments (Browder et al., 2004). The passage of IDEA triggered a shift in thinking and state education agency personnel realized that alternate assessments needed to incorporate academic domains to focus on students' performance on state standards as a way for students to access the general curriculum (Browder et al., 2005).

There are three typical alternate assessment approaches (Roeber, 2002, Laitusis et al., 2014). The portfolio approach involves a body of student work collected and subsequently evaluated against predetermined scoring criteria. The checklist approach requires teachers to identify whether students are able to perform certain skills. Scoring is based on the number of

skills the students are able to perform proficiently. The performance assessment approach is a direct measure of a skill by teachers observing the students performing assigned tasks. Although most state departments of education chose to use one of these three approaches, some have combined approaches. For example, some states require teachers to complete a checklist and also to directly measure students' skills via a performance assessment (Roeber, 2002).

The mandate to create AA-AAS led states to propose a variety of methods for assessing SWDs (Laitusis et al., 2014). In response, several recommendations about how to implement AA-AAS effectively have been made. For example, it is recommended that states improve the content validity of alternate assessments (Browder et al., 2005). Doing so involves clarifying which standards should be addressed on the alternate assessments and how these standards could be measured. States must define the meaning of reading, math, and science for students who are nonverbal or who use symbolic communication. In addition, it is recommended that AA-AAS assess student acquisition of skills by measuring student learning each year. Alternate assessments should also be linked to students IEP goals (Browder et al., 2005).

Thompson et al. (2001) recommended that states use the same content standards for SWDs, but that they incorporate less rigorous achievement standards that would be more appropriate for these students. The U.S. Department of Education defines content standards as "statements of the knowledge and skills that schools are expected to teach and students are expected to learn" (U.S. Department of Education, 2007, p.13). Content standards must be coherent and encourage teaching advanced skills. Academic achievement standards are "explicit definitions of how students are expected to demonstrate attainment of the knowledge and skills reflected in the content standards" (U.S. Department of Education, 2007, p.13). For example, a standard on communicating ideas through speaking might be redefined as communicating basic

needs using assistive technology while being self-expressive to others. Some SWDs could be responsible for learning the required content standards, yet held to less rigorous achievement standards.

AA-AAS Relation to Common Core State Standards

As previously discussed, an AA-AAS must be aligned with the state's content standards. Although there have been efforts to standardize education since the early 1990s, by the early 2000s each state developed its own learning standards and definition of proficiency for students in grades 3-12. As a result of this inconsistency, remediation rates have been high, especially in mathematics. In response, in 2009 state school chiefs and governors from 48 states, through their membership in the National Governors Association's Center for Best Practices (NGA Center) and the Council of Chief State School Officers (CCSSO), launched the Common Core State Standards (CCSS) (NGA & CCSSO, 2010).

The Common Core set standards for mathematics and English language arts (ELA) that outline skills students should know and be able to perform by the end of each grade. These are areas upon which students build skill sets that are used in other subjects. The CCSS stress that students should learn to read, write, speak, listen, and incorporate language in a variety of content areas because these skills are required for college and career readiness. The ESSA allows states to adopt the CCSS, but does not require it (ESSA, 2015). Forty-two states, the District of Columbia, and the Department of Defense Education Activity (DoDEA) have adopted the CCSS. Creators of the CCSS recognized the value of applicable learning goals and launched this effort to standardize education standards to ensure that all students graduating high school are able to thrive in college and their careers (NGA & CCSSO, 2010).

Despite this drive for standards in education, the CCSS do not provide information on how to apply these standards to SWDs (Browder et al., 2005). Some special education teachers struggle to create access to grade-level academic content. Some teachers question the relevance of grade-level content for students with significant intellectual disabilities (Browder et al., 2007). Not understanding these connections makes it difficult for special education teachers to adapt assessments for these students (Saunders, Bethune, Spooner, Browder, 2013).

One rationale for teaching the CCSS to SWDs is that demands for mathematical competency in today's technological world have increased (e.g., numerically operated machinery). Furthermore, research suggests that SWDs can learn content aligned grade-level standards. Browder et al. (2012) demonstrated that if task analysis, a graphic organizer, and mathematical stories were used, SWDs could learn state standards for mathematics that correspond to their grade level. Task analysis is "a method of breaking down a long, complicated task into its component steps" (Saunders et al., 2013, p. 29). For example, teaching how to measure an object with a ruler can be disaggregated into simple steps that require the student to (1) find the zero on the ruler, (2) find the edge of the object, (3) line the ruler against the object, (4) look at the ruler to find the object's stopping point, and (5) find the number on the ruler that is closest to the object's stopping point. By taking data on each step to determine the areas that are difficult for students, task analysis allows for teaching in a systematic way. The data derived can also be helpful in creating IEPs for students.

Psychometric Properties of Alternate Assessment and Implications

Although the U.S. Department of Education (2005) required that AA-AAS meet standards of technical quality including validity, reliability, accessibility, objectivity, and consistency, the question of standards remains a concern (Elliott, Compton, Roach, 2007,

Laitusis et al., 2014). When AA-AAS was first implemented almost all of the data was anecdotal, for example, teachers describing student performance (Marion, 2006). In addition, different states had variations of portfolio assessments, checklists, and performance assessments. In 2003, over half of the states used portfolios and some states used either one or multiple scoring rubrics. There was also no relation among alternate assessment scores, IEP goals, and post-school outcomes. In addition, some states used mostly functional domains in the assessment, while others were using mostly academic domains (Marion, 2006).

Johnson and Arnold (2004) conducted the first published study that examined the validity of an alternate assessment portfolio. Raters were trained to complete a checklist designed to analyze the Washington (state) Alternate Assessment System (WAAS). Results "indicated serious shortcomings" (Johnson & Arnold, 2004, p. 266). In some of the portfolios sampled, content and structural invalidity were found, including unclear scoring criteria and misalignment to state standards. The connection between the WAAS and the CCSS appeared to be superficial because minimal information was available to teachers that showed how IEP goals were associated with content standards. In addition, teacher subjectivity contributed to student scores, and the basis for scores in general proved to be unclear. Overall, it seemed that the portfolios were more of a reflection of teachers' ability to compose a portfolio than of student progress.

Performance assessment and comprehensive rating scale approaches have demonstrated higher levels of technical soundness because they have more aligned items and because they sample more discrete knowledge and skills (Elliott, Compton, Roach, 2007). Aligned items are linked to state grade level content standards. For example, the history standard to develop historical perspective (i.e., ELA standard for history, grades 9-10) might be redefined as "use a personal calendar" (Browder et al., 2005).

In 2008, 15 states did not meet peer review standards for their AA-AAS (U.S. Department of Education, 2009). This problem partially stems from the nature of alternate assessments being left to interpretation, which led to states using different methods of approaching the assessment process (Browder et al., 2003). Proportionally, very few students participate in alternate assessments, and each student may have a different style of learning. As a result, there are many challenges to documenting the technical quality of AA-AAS. Currently, there is still little research on the psychometric quality of AA-AAS, and nearly all of the existing psychometric research has focused on the impact of testing accommodations for SWDs (Laitusis et al., 2014).

The *Standards for Educational and Psychological Measurement* (2014) (*Standards*) provides criteria for the "development and evaluation of tests and testing practices and to provide guidelines for assessing the validity of interpretations of test scores for the intended test uses" (AERA, APA, NCME, 2014, p. 1). The *Standards* define key two psychometric properties, reliability and validity, which are extremely relevant to constructing AA-AAS.

Reliability. Reliability is defined as "The degree to which test scores for a group of test takers are consistent over repeated applications of a measurement procedure and hence are inferred to be dependable, and repeatable for an individual test taker; the degree to which scores are free of errors of measurement for a given group" (AERA, APA, NCME, 2014, p. 223). In other words, reliability refers to the consistency of scores across replications of a testing procedure. Several measures of reliability can be estimated over replications directly. Although the *Standards* outline many different types of reliability only the types relevant to the current study will be discussed.

One type of reliability is the test-retest reliability coefficient, which is "A reliability coefficient obtained by administering the same test a second time to the same group after a time interval and correlating the two sets of scores" (AERA, APA, NCME, 2014, p. 224). This type of reliability measures the stability of test scores. A test that possesses high test-retest reliability has results about student performance on a domain that can be correctly generalized from one time to another.

Another estimate of reliability is the internal consistency coefficient, which is defined as "An index of the reliability of test scores derived from the statistical interrelationships among item responses or scores on separate parts of a test" (AERA, APA, NCME, 2014, p. 220). Internal consistency measures whether various items that purport to measure the same general construct generate similar scores. Internal consistency is usually measured with Cronbach's alpha, which is a statistic calculated from the correlations among items (AERA, APA, NCME, 2014).

Another type of reliability is inter-rater reliability, which is defined as "The level of consistency in rank ordering of ratings across raters" (AERA, APA, NCME, 2014, p. 220). Possessing adequate inter-rater reliability is dependent on agreement in scoring among comparably trained raters who observe the same phenomena at the same point in time. Inter-rater reliability therefore refers to the degree to which similar types of raters (e.g., two parents) who know the child from the same situational context agree on scores on the same instrument.

It is important to distinguish inter-rater reliability from cross-informant agreement. Cross informant agreement refers to the degree to which different types of raters (e.g., teachers and parents) who are aware of a child's functioning in different situational contexts agree on the

scores on the same instrument (Phye, Saklofske, Andrews, Janzen, 2001; Achenbach & Rescorla, 2007; Sointu, Savolainen, Lappalainen, Epstein, 2012).

Even though psychological assessment has value, achieving inter-rater reliability is an important concern about the technical quality of alternate assessment. Browder et al. (2003) explained that achieving inter-rater reliability in terms of AA-AAS might be especially challenging. Meyer et al. (2001) reviewed more than 125 meta-analyses on test validity and concluded that any single assessment method provides incomplete information for children when various types of knowledgeable informants are compared with each other. For example, teacher ratings have only moderate agreement with clinician ratings ($r = .34$). The authors indicate that any sole assessment method provides an incomplete representation of the constructs it intends to measure and it is difficult to obtain agreed upon information about patients.

Achenbach, McConaughy, and Howell (1987) discuss that low-correlations between different informants have been interpreted as placing doubt on the informants' abilities to evaluate the child. However, this neglects the possibility that different informants contribute different information to their assessments. Cross-informant agreement is therefore not fully intended to assess reliability, which concerns consistency in results, because different informants may contribute different information. The authors reviewed 119 studies on childhood behavioral and emotional problems in a meta-analysis to investigate cross-informant agreement across different types of raters. Mean Pearson correlations were .60 across similar raters (e.g., pairs of parents) and .28 across different types of raters (e.g., parents and teachers). This may indicate that it is difficult to obtain high cross-informant agreement among various types of raters when each rater contributes different information. The low correlations among raters indicate that

childhood behavior and emotional problems cannot be effectively captured by one type of judgment alone.

De Los Reyes et al. (2015) conducted a meta-analysis that reviewed 341 studies that appeared since the year 2000 of cross-informant agreement of children's internalizing and externalizing mental health concerns. The authors discuss how children may display mental health concerns in some situational contexts and not others (e.g., home vs. school). This is important because identifying the specific situational contexts in which children display various behavioral and emotional concerns can help develop treatment planning and enhance treatment efficacy. The multi-informant assessment approach, a critical way to assess context distinctions in mental health, involves taking reports from multiple informants who know the child. Findings indicate low-to-moderate correlations of correspondence (mean internalizing: $r = .25$; mean externalizing: $r = .30$), suggesting that the multi-informant assessment approach may not be a valuable way of judging children's mental health concerns.

Kentucky was the first state to fully include all students in a statewide educational assessment and accountability system. Kleinert, Kearns, and Kennedy (1997) reviewed the development of Kentucky's alternate assessment. Specifically, three years of initial implementation data were analyzed. Reliability was calculated as the measure of agreement between the first rater (teacher) and second round rater (regional level) with the following formula:

$$[\text{number of agreements} \div (\text{number of agreements} + \text{disagreements})] \times 100$$

In the first year, teachers actually scored their students' portfolios lower than the second round raters and only 63% of portfolios were scored the same in both the first and second scoring.

During Years 2 and 3, reliability between the first and second raters actually decreased. In Year 2 and Year 3, only 51% and 52% of the portfolios respectively were scored identically.

Issues of reliability remain critical because alternate assessments continue to remain part of the accountability system (Kleinert, Kearns, and Kennedy, 1997). Teacher judgment became a focus of research because it was theorized that teacher subjectivity could be detrimental in meeting standards of technical quality. In a qualitative study of two school districts, teachers explained that it was difficult to attain adequate inter-rater reliability in terms of assessing student portfolios (Karvonen, Browder, Wakeman, Algozzine, 2006). Some teachers thought that the Alternate Assessment Portfolios (AAP) were too subjective. Teachers indicated that it was difficult to obtain a high student score without also having an attractive portfolio with clean data sheets. In addition, teachers recognized that each student's mastery level depended on the expectations the teachers set. For example, teachers who set higher goals had students who performed poorer. Teachers who set lower and more realistic goals had students who performed higher.

Moreover, teacher judgment can systematically influence the reliability of alternate assessment scores for students at different grade levels (Roach, Elliott, Berndt, 2007; Salvia & Munson, 1986). As students become older the focus becomes more on the subject area rather than child-centered instruction. In addition, teachers' perceptions of inclusive practices are generally less positive as students get older. Furthermore, standards set for early grade levels are better reflected by the associated alternate assessment performance indicators than by the standards and associated alternate assessment performance indicators set for older grades (Browder et al., 2004).

Browder et al. 2005 therefore recommended that AA-AAS assess student acquisition of skills by measuring what students learn each year. *The Vineland Adaptive Behavior Scales*, Second Edition (Vineland-II), is a measure of adaptive behavior that evaluates students' skills needed to take care of oneself in regular daily activities. The Vineland-II includes four domains: Communication, Daily Living Skills, Socialization, and Motor Skills. It also has a Maladaptive Behavior domain that assesses problem behaviors. Items are rated on a 3-point Likert Scale (0 = Never, 1 = Sometimes, 2 = Always).

The Vineland-II manual presents raw and adjusted intraclass correlations between the first and second administration of scores. The interval between occasions ranged from 13 to 34 days. The average test-retest reliability across domains within age groups ranged from .88 to .92, except for ages 14 through 21 in which the average reliability across domains was .76. Internal consistency estimates obtained for this age range were also low. Overall, subdomain retest reliability coefficients were high, with most values exceeding .85. The Vineland-II manual also reports mean scores from the first and second administration. The average difference from the first to second administration was small, which suggests that the second administration was not biased due to familiarity with assessment content.

Validity. Validity is defined as the "Degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests" (AERA, APA, NCME, 2014, p. 11). A test itself cannot be valid, but rather specific interpretations of the test scores can be valid. Validity therefore assesses interpretations of the construct that the test is intended to measure. Although the *Standards* outline many different types of validity only the types relevant to the current study will be discussed.

Content validity is defined as “Evidence based on test content that supports the intended interpretation of test scores for a given purpose” (AERA, APA, NCME, 2014, p. 218). Content validity therefore involves whether the content of the measure representatively samples a domain. A glimpse across the states’ alternate assessment performance indicators raised the issue of content validity (Browder et al., 2005). For example, some states included performance indicators of “cleaning the kitchen” for mathematics and “making eye contact” for language arts. Some of these indicators did not reflect CCSS.

Internal structure validity evidence is defined as “The degree to which the relationships among test items and test components conform to the construct on which the proposed test score interpretations are based” (AERA, APA, NCME, 2014, p. 16). Internal structure validity can be analyzed through factor analysis, which is a data reduction method used to seek underlying variables that are reflected in the observed variables. Factor analytic techniques help demonstrate that the various domains or subscales fit together.

Another way to assess internal structure validity is by analyzing the correlations among the different constructs on the same instrument. This approach helps determine the extent to which different constructs overlap. The Vineland-II manual presents correlations among the different domains. For the 7-11 age range, the Communication domain correlates with the Daily Living Skills domain in the very large range (.81); the Communication domain correlates with the Socialization domain in the large range (.68); the Communication domain correlates in the nearly perfect range with the Adaptive Behavior composite (.92); the Socialization domain correlates in the large range with the Daily Living Skills domain (.69); the Socialization domain correlates with the Adaptive Behavior composite in the very large range (.87); and the Daily

Living Skills domain correlates in the nearly perfect range with the Adaptive composite domain (.92).

Gathering multiple sources of evidence can provide a comprehensive evaluation of construct validity. In the multitrait-multimethod (MTMM) approach researchers use scores from multiple methods that are intended to be indicative of multiple traits as evidence of the validity of a measure. In practice, this has proved useful for evaluating a variety of constructs, such as academic skills and performance, as well as adaptive behavior.

In an important example of this approach, Elliott, Compton, and Roach (2007) provided a MTMM framework. To obtain information about the strength of relationships and the underlying construct(s) being measured, they compared the Idaho Alternate Assessment (IAA) with related standardized measures, including the Academic Competence Evaluation Scales [ACES] and the Vineland Adaptive Behavior Scale [VABS] (Idaho Department of Education, 1999; DiPerna, & Elliott, 2000; Sparrow, Balla, & Cicchetti, 1985). These measures assess multiple traits, including academic skills, academic enablers, and adaptive behavior. Demographic characteristics were also provided. Multiple methods were used to obtain data, including teacher completed rating scales, demographic surveys, and an individualized achievement test for a subset of students. The authors sought to obtain evidence regarding whether the IAA would correlate with related measures and would not correlate with unrelated measures. In general, IAA scales measured skills indicative of the academic content characterized in the state's content standards. Results demonstrated significantly higher correlations between IAA subscale scores in Reading ($r = .59$), Language Arts ($r = .45$), and Mathematics ($r = .75$) with measures of adaptive behavior on the VABS than academic skills. The IAA shares significant variance with measures of adaptive behavior and less variance with a measure of academic skills.

Kettler et al. (2010) further advanced work in this area by examining the relationship between an AA-AAS, a general achievement test, and two norm-referenced teacher rating scales. In the following six states – Arizona, Hawaii, Idaho, Indiana, Mississippi, and Nevada – the authors found a high degree of shared variance between reading and mathematics scores, indicating they could potentially represent the same construct. Analyses also revealed that scores from the states' AA-AAS was strongly related to adaptive behavior as measured by the Vineland Adaptive Behavior Scale (VABS). Over half the correlations were in the very large range or higher. AA-AAS was also found to relate to academic constructs, including readiness and academic skills. Correlations here were within the medium to large ranges.

PG Chambers School Outcomes Measurement Tool (PGS-OMT)

PG Chambers is a charter school in New Jersey that serves students who are classified under the disability category of preschool child with a disability or multiple disability. Professionals within PG Chambers recognized the need for an evaluation tool that could quantify the difference that their school program makes in the lives of their students. In 2008, a transdisciplinary team, including teachers and teaching assistants, an occupational therapist, a physical therapist, a speech language therapist, and a nurse created the PG Chambers School Outcomes Measurement Tool (PGS-OMT) to assess both therapeutic progress and academic progress for students with multiple disabilities. This alternate assessment is used to evaluate students in the following domains: Social, Functional, Physical Navigation, Personal Care, Communication, and Academics.

The PGS-OMT aims to increase the effectiveness of PG Chambers' programs and services, assist the school in fundraising efforts, and communicate the program's value to stakeholders. In order to effectively identify composites that reflect the needs of the student

population at PG Chambers, the PGS-OMT was tested in one official pilot phase in 2009. Three classes, including one preschool, one elementary school, and one middle school, participated. The teams who analyzed the data gave feedback on the tool. The results of the pilot test suggested that the PGS-OMT needed to be modified. Specifically, items identified as ambiguous were reworded, content and skills were operationally defined, additional measures were surveyed and considered.

During 2012, two teachers, one assistant principal, and one case manager at PG Chambers reviewed the PGS-OMT and revised the academic domain to align with key elements from the CCSS for all grade levels. An additional discussion was held with all of PG Chamber's teachers to gain their insight. For example, the seven mathematics items in the PGS-OMT were aligned with the first grade state standards that targeted skills in addition, subtraction, understanding whole number relationships, and basic geometry (NGA & CCSSO, 2010). The items on the PGS-OMT that targeted phonics, reading fluency, and comprehension were aligned with the first grade reading standards for foundational skills.

Zahra (2015) was the first researcher to examine the reliability of scores and the validity of conclusions drawn from the PGS-OMT. Participants in the study included students attending PG Chambers (n = 117) enrolled in prekindergarten through 8th grade. This quantitative study analyzed various psychometric properties including reliability, internal structure validity, relations to other variables, and utility. Coefficient alpha was used to assess whether the PGS-OMT yielded reliable scores that demonstrated adequate internal consistency. Factor analytic techniques were used to assess whether the PGS-OMT demonstrated adequate internal structure validity. In addition, Pearson correlations were calculated to determine whether the PGS-OMT and Vineland Adaptive Behavior Scales, Teacher Report Form-Second Edition (Vineland-II)

converged. Descriptive statistics analyzed demographic characteristics to determine whether there was a significant difference between mean scores of advantaged and disadvantaged ethnic groups for any domain measured by the PGS-OMT. Finally, descriptive statistics were used to analyze teacher impressions of the utility of the PGS-OMT and their impressions of the process of completing the test (Zahra, 2015).

Zahra (2015) found that the PGS-OMT yields scores that have high internal consistency. Coefficient alpha reliabilities were .99 at the total scale level and item-total correlations were high for all six domains. This suggests that the items fit well together. In terms of evidence for internal structure validity, confirmatory factor analysis (CFA) was initially conducted using a six-factor model and demonstrated an overall poor fit. Results of a subsequent exploratory factor analysis (EFA) revealed that a three-factor model accounted for 73% of the variance in the relationships among the items. The first factor was a combination of the Functional and Academic domains (Functional Academics). The second factor was a combination of the Personal Care and Physical Navigation domains (Adaptive Behavior). The third factor was a combination of the Social and Communication domains (Interpersonal). The correlations between factor domains were consistent with the EFA loadings (Zahra, 2015).

The PGS-OMT was also compared to the Vineland-II to demonstrate validity based on relations to other variables (Zahra, 2015). Stronger relationships were observed between the related areas that make up the three-factor model (e.g., Interpersonal factor domain of the PGS-OMT, and Socialization and Communication domains of Vineland-II) and weaker relationships were observed between unrelated domains that did not comprise the three-factor model (e.g., Interpersonal domain of the PGS-OMT and Daily Living Skills domain of Vineland-II). The correlation between the Interpersonal factor domain of the PGS-OMT, and Social and

Communication domains of the Vineland-II were in the very large range. The correlation between the Functional Academics factor domain of the PGS-OMT and Daily Living Skills domain of the Vineland-II was in the large range. The correlation between the Adaptive Behavior factor domain of the PGS-OMT and Motor Skills domain of the Vineland-II was in the very large range.

To assess demographic bias, participants were dichotomized into groups: historically advantaged (i.e., White/Non-Hispanic and Asian) and historically disadvantaged (i.e., Black/Non-Hispanic, Hispanic, and American Indian/Alaskan Native/Pacific Islander), and mean scores across each factor domain were compared (Zahra, 2015). The Adaptive Behavior domain was the only area that indicated significant differences in scores between advantaged and disadvantaged ethnic groups. Specifically, participants in the disadvantaged groups were rated as having stronger gross motor skills (e.g., avoiding obstacles, navigating uneven surfaces) than those in the advantaged group. However, these results need to be interpreted with caution, because the disadvantaged group was ethnically heterogeneous and because the differences may be more characteristic of specific impairments than cultural influences (Zahra, 2015).

In terms of the perceived utility of the PGS-OMT, evaluation results indicated that teachers found the measure to be useful and time efficient for evaluating students with multiple disabilities (Zahra, 2015).

Summary

Alternate assessments are often used to evaluate the academic performance of students with severe and multiple disabilities (Elliott, Compton, & Roach, 2007). Research has shown that even though students with severe and multiple disabilities have participated in state-level alternate assessments for years, their educational achievement has not been measured in a

psychometrically sound way (Kettler et al., 2010; Elliott, Compton, & Roach, 2007). Therefore, schools cannot meaningfully evaluate these students' progress. Since alternate assessments are a key component of each state's assessment system, it is necessary to remedy this deficiency.

The current study is designed to provide insight into the internal consistency, cross-informant agreement, internal structure validity, and change in scores of the PGS-OMT. Findings have the potential to influence the type of assessment administered to the students at PG Chambers. Educators could use findings to better guide decisions concerning goals and objectives for their students.

Research Questions and Predictions

1. *Are scores yielded by the PGS-OMT reliable?*
    a. *Do the PGS-OMT scores from the overall sample demonstrate adequate internal consistency?* Coefficient alpha is predicted to be within the acceptable range (.80 or greater) for all domains, because previous research indicates that the PGS-OMT possesses excellent internal consistency (Zahra, 2015).
    b. *Does the PGS-OMT, as assessed by Team A and Team B, demonstrate adequate cross-informant agreement?* Pearson correlations are predicted to be within the large range, because previous research indicates that the PGS-OMT possesses excellent internal consistency, and because the tool is designed for use by professional educators in various roles (Zahra, 2015).
2. *Does the structure of the PGS-OMT reflect its constructs?* In Zahra's (2015) study, an exploratory factor analysis yielded a three-factor model. The first factor was a combination of the Functional and Academic domains. The second factor was a combination of the Personal Care and Physical Navigation domains. The third factor was a combination of the Social and Communication domains. Correlations among factor domains are predicted to range from the medium to large ranges, because previous research indicates these correlation ranges (Zahra, 2015). Correlations among the original six domains are predicted to range from the medium to very large ranges, because previous research indicates these correlation ranges (Zahra, 2015). Confirmatory factor analysis (CFA) from the overall 2014 sample is predicted to yield an overall good fit for the proposed three-factor model and demonstrate that items appropriately load onto their respective factors.

3. *Did PG Chambers students' performance on the PGS-OMT improve from 2013 to 2014?* Mean and standard deviation indices are expected to be significant indicating that students' performance improved from 2013 to 2014. Paired samples t-tests of 2013 and 2014 mean PGS-OMT scores are predicted to be significant. Cohen's d is predicted to indicate a small effect.

Chapter II

Method

Participants

PG Chambers has Early Childhood, Elementary, and Middle School programs serving students between the ages of three and fourteen. All students at PG Chambers are classified under the disability category of preschool child with a disability or multiple disabilities. PG Chambers educates students under the school districts' permission. Local education authorities placed these students at PG Chambers in accordance with IDEA (2004), because this placement was determined to be the Least Restrictive Environment (LRE) in which the students' educational needs could be appropriately met.

The participants in the 2014 sample include students (n = 121) enrolled in the PG Chambers School, and teachers (n = 14), physical therapists (n = 8), occupational therapists (n = 11), and speech therapists (n = 7) employed at PG Chambers. Passive informed consent forms were distributed to the sending school districts and to students' parents for the purpose of accessing extant data. The students at PG Chambers are the secondary data sources, because the transdisciplinary team identified as the primary data sources assessed them. This study employed four samples including an overall sample, an agreement subsample, a change sample, and a rater sample.

Overall sample. Of the program's total population of 123 students in 2014, the overall sample consisted of 121 students. Parents of two students excluded their children from this study. The student sample was approximately evenly balanced with regard to gender and was predominantly White/Non-Hispanic (70%). Students in the sample were spread evenly across grades. About half of the students had impaired vision (48%) and few had impaired hearing

(12%). Most of the students were non-verbal (63%) and about half of the students were considered to be non-ambulatory (46%). Table 1 depicts the demographic characteristics of the overall sample.

Table 1

Student Demographic Characteristics

	Overall Sample n (%)	Agreement Subsample n (%)	Change Sample n (%)
Gender			
Male	66 (55)	16 (37)	52 (53)
Female	55 (46)	27 (63)	46 (47)
Grade			
Prekindergarten	28 (23)	9 (20)	14 (14)
Kindergarten	9 (7)	7 (16)	9 (9)
1	11 (9)	4 (9)	7 (7)
2	6 (5)	4 (9)	8 (8)
3	24 (20)	11 (26)	22 (22)
4	10 (8)	3 (7)	8 (8)
5	11 (9)	5 (12)	10 (10)
6	10 (8)	0	8 (8)
7	6 (5)	0	6 (6)
8	6 (5)	0	6 (6)
Ethnicity			
White/Non-Hispanic	85 (70)	31 (72)	71 (72)
Black/Non-Hispanic	7 (6)	4 (9)	5 (5)
Hispanic	18 (15)	5 (12)	14 (14)
Native/Pacific Islander	1 (1)		1 (1)
Asian	7 (6)	1 (2)	5 (5)
Other	3 (2)	2 (5)	2 (2)
Vision			
Within Normal Limits	63 (52)	19 (44)	53 (54)
Impaired	58 (48)	24 (56)	45 (46)
Hearing			
Within Normal Limits	107 (88)	38 (88)	85 (87)
Impaired	14 (12)	5 (12)	13 (13)
Communication			
Verbal	45 (37)	17 (40)	34 (35)
Non-Verbal	76 (63)	26 (60)	64 (65)
Mobility			
Ambulatory	65 (54)	23 (53)	52 (53)
Non-Ambulatory	56 (46)	20 (47)	46 (47)
Total Participants	121	43	98

Agreement subsample. Team A and Team B each rated the same 43 students; 78 students were excluded from this agreement subsample due to time constraints. The student agreement subsample was predominantly female (63%) and White/Non-Hispanic (72%). In addition, the sample was approximately evenly balanced with regard to grade level. However, the sample did not include students from the Middle School program. About half the students had impaired vision (56%) and few had impaired hearing (12%). About half the students were considered to be non-ambulatory (47%). Most of the students were non-verbal (60%). Table 1 depicts the demographic characteristics of the agreement subsample.

Change sample. Matching identification numbers (n = 98) appearing in both 2013 and 2014 data sheets were gathered to assess the change in PGS-OMT scores. The change sample was approximately evenly balanced with regard to gender and was predominantly White/Non-Hispanic (72%). In addition, the sample was approximately evenly balanced with regard to grade level. About half the students had impaired vision (46%) and few had impaired hearing (13%). About half the students were considered to be non-ambulatory (47%). Most of the students were non-verbal (65%). Table 1 depicts the demographic characteristics of the change sample.

Rater sample. The rater sample, consisting of teachers (n = 14), speech therapists (n = 7), occupational therapists (n = 11), and physical therapists (n = 8), were female and mainly White/Non-Hispanic (98%). The rater sample was approximately evenly balanced with regard to years working at PG Chambers and how long raters had been certified. The most common highest degree raters earned were bachelor's degrees (40%) or master's degrees (50%).

The teachers in the rater sample were approximately evenly balanced with regard to how long they had been working at PG Chambers and how long they had been certified. The most common highest degree teachers had earned was a bachelor's (71%). Most of the speech

therapists in the rater sample had been working at PG Chambers between 1 and 5 years (57%) and had been certified between 1 and 5 years (57%). The highest degree speech therapists earned was a master's (100%). The occupational therapists in the rater sample were approximately evenly balanced with regard to how long they had been working at PG Chambers. Most occupational therapists had been certified between 11 and 20 years (45%) or between 6 and 10 years (36%). The most common highest degree occupational therapists earned was a master's (64%). Most of the physical therapists in the rater sample had been working at PG Chambers between 1 and 5 years (63%) and had been certified for more than 21 years (63%). The most common highest degree physical therapists earned was a doctorate (50%). Table 2 depicts the demographic characteristics for the rater sample.

Table 2

Rater Demographic Characteristics

	Teachers n (%)	Speech Therapists n (%)	Occupational Therapists n (%)	Physical Therapists n (%)	Combined Rater Sample n (%)
Gender					
Male					
Female	14 (100)	7 (100)	11 (100)	8 (100)	40 (100)
Ethnicity					
White/Non-Hispanic	14 (100)	7 (100)	10 (91)	8 (100)	39 (98)
Black/Non-Hispanic	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
Hispanic	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
Native/Pacific Islander	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
Asian	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
Other	0 (0)	0 (0)	1 (9)	0 (0)	1 (3)
Years Working at PG Chambers					
1-5	4 (29)	4 (57)	4 (36)	5 (63)	17 (43)
6-10	4 (29)	1 (14)	4 (36)		9 (23)
11-20	6 (43)	2 (29)	3 (27)	2 (25)	13 (33)
21+	0 (0)	0 (0)	0 (0)	1 (13)	1 (.03)
Years Certified					
1-5	2 (14)	4 (57)	0 (0)	1 (13)	7 (18)
6-10	5 (36)	1 (14)	4 (36)	1 (13)	11 (28)
11-20	3 (21)	2 (29)	5 (45)	1 (13)	11 (28)
21+	4 (29)	0 (0)	2 (18)	5 (63)	11 (28)
Highest Degree Earned					
High School Diploma	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
Associates	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
Bachelor's Degree	10 (71)		4 (36)	2 (25)	16 (40)
Master's Degree	4 (29)	7 (100)	7 (64)	2 (25)	20 (50)
Doctorate				4 (50)	4 (10)

Measures

PG Chambers School Outcomes Measurement Tool (PGS-OMT). The PGS-OMT is a rating scale professionals use to assess students' progress in the program. The goal of the PGS-OMT is to evaluate whether PG Chambers is helping their students prosper in various areas of development. Students are assessed once a year for their current performance.

The present version of the PGS-OMT assesses performance on items in the following domains: Communication, Social, Personal Care, Physical Navigation, Academics, and Functional. Each item is evaluated on a 7-point Likert scale. For example, an item outlined in the Social domain is acknowledging social greetings. The Likert scales are accompanied by skill level definitions that guide the use of the rating scale. The differences between the anchors that guide the use of this Likert scale will be discussed in the following chapters.

The Social, Communication, Physical Navigation, and Personal Care domains are rated according to frequency and a student's level of support needed for expressing a particular skill. The scores are rated as follows: 0 = Never Expresses the Skill, 1 = Needs Maximum Support, 2 = Needs Moderate Support, 3 = Needs Minimal Support, 4 = The Skill is Emerging (25%), 5 = The Skill is Sometimes Independent (50%), 6 = The Skill is Often Independent (75%), 7 = The Skill is Always Independent (100%).

The Academic and Functional domains are rated differently. The scores are rated as follows: 0 = Never, 1 = Introduced (Participation), 2 = Attempted (Exploring), 3 = Minimal (Infrequent and Inconsistent Performance), 4 = Emerging (Beginning Performance), 5 = Sometimes (Regular Performance/Not Always Correct), 6 = Often (More Often Correct but Not Mastered), 7 = Mastered.

The PGS-OMT has 64 items. There are 10 items in the Social domain, 9 items in the Communication domain, 10 items in the Physical Navigation domain, 10 items in Personal Care domain, 17 items in the Academic domain, and 8 items in the Functional Domain. The items that comprise the Academic domain were compared to the CCSS to determine whether they aligned. Specifically, the principal investigator examined each individual item from the Academic domain to determine whether the content appeared to align with any ELA or mathematic components outlined in the CCSS. Table 3 indicates that each item appropriately aligned with the respective ELA or mathematic components outlined in the CCSS.

Table 3

Academic Domain and CCSS Alignment

Academic Domain Item	ELA	Mathematics
Demonstrates an understanding of concepts in print (letter recognition)	Yes	
Demonstrates word analysis skills (sound blending)	Yes	
Demonstrates phonemic awareness (letter sounds)	Yes	
Demonstrates reading fluency	Yes	
Reads 100 sight word vocabulary	Yes	
Demonstrates reading comprehension skills by answering 3/5 questions correctly	Yes	
Reads 50 functional words	Yes	
Demonstrates listening comprehension skills by answering 3/5 questions correctly	Yes	
Identifies numbers 0-100		Yes
Rote Counts 0-100		Yes
Performs calculations using whole numbers. (Using a number line, manipulatives and/or fingers)		Yes
Solves one-step addition or subtraction word problems		Yes
Performs calculations with fractions. (1/2,1/3,1/4, whole)		Yes
Applies geometric concepts. (Identifies plane and solid figures.)		Yes
Performs measurement concepts. (Selects correct tool to measure: ruler, measuring cup etc.)		Yes
Use number in functional activities. (Phone number, address etc.)		Yes
Uses calculator		Yes

Previous evidence indicates that the PGS-OMT yields scores that have high internal consistency (Zahra, 2015). Coefficient alpha reliabilities were .99 at the total scale level. Item-total correlations were high for all six domains and suggest that the items fit well together. In terms of evidence for internal structure validity, results of confirmatory factor analysis (CFA) demonstrated an overall poor fit. A six-factor model was originally tested, because there are six domains in the PGS-OMT. This six-factor model yielded a poor fit. Results of a subsequent exploratory factor analysis (EFA) revealed that a three-factor model accounted for 73% of the variance. The first factor is a combination of the Functional and Academic domains (Functional Academics). The second factor is a combination of the Personal Care and Physical Navigation domains (Adaptive Behavior). The third factor is a combination of the Social and Communication domains (Interpersonal).

There was variability in the strength of correlations between factor domains (r = .45 to r = .75) (Zahra, 2015). The largest correlations were between the Academic and Functional domains (.91), Social and Communication domains (.87), and Physical Navigation and Personal Care domains (.82), which were consistent with the EFA item loadings. To demonstrate evidence based on relations to other variables, the PGS-OMT was also compared to the Vineland-II (Zahra, 2015). Stronger relationships were observed between related domains, and weaker relationships were observed between unrelated domains.

Student Demographic Questionnaire. Each classroom teacher completed the Student Demographic Questionnaire (Appendix A) for each student they rated. The questionnaire asks for the student's gender, ethnicity, diagnosis, date of enrollment, grade level, number of absences, type of visual impairment, type of hearing impairment, type of communication impairment, and mobility status.

Procedure

Each classroom is comprised of students in a particular grade and of a team of professionals that include one or more speech language therapists, occupational therapists, physical therapists, and teachers. All thirteen classrooms in PG Chambers were involved in the study. The professionals who rated the students worked in the classrooms of the students they rated. However, the raters did not actively participate in this study. They rated the students as part of the usual annual assessment. The principal investigator was given permission to use the de-identified, extant data.

For the agreement subsample, professionals in each classroom were divided into Team A and Team B. Both teams rated the students. Team A was comprised of teachers and occupational therapists and Team B was comprised of speech therapists and physical therapists. PG Chamber's staff claimed that this stratified split among the raters was random by role and the principal investigator did not control the process. While they were stratified by role, raters were assigned to rate the students who they had experience working with.

Stage one in the rating process included Team A and Team B completing the PGS-OMT independently for each student in the agreement subsample they were assigned to rate. The team members collaborated internally to rate each student. Raters were instructed to rate students based on pre-existing impressions from their observations of the students since the previous evaluation was conducted.

Raters were similar because they based their ratings of the students on their performance in the same classroom. However, raters were different because they had different experiences with the students. For example, teachers and occupational therapists had different experiences with the students. Therefore, the analyses that relate to the consistency in ratings across raters

most appropriately fall under the category of cross-informant agreement as opposed to inter-rater reliability.

By the end of stage one, each student in the agreement subsample had two PGS-OMTs completed, one from Team A and one from Team B. The rating scales in stage one took about thirty minutes to complete for each student. In stage two, Team A and Team B joined together to complete a third PGS-OMT for each participant in the agreement subsample. The combined teams also completed one PGS-OMT for all the other participants in the overall sample who were not part of the agreement subsample. Each classroom teacher subsequently completed the Student Demographic Questionnaire for each student who was rated. Each team completed the PGS-OMT in its entirety. No items were skipped. Refer to Figure 1 for a visual description of how the agreement subsample was constructed.

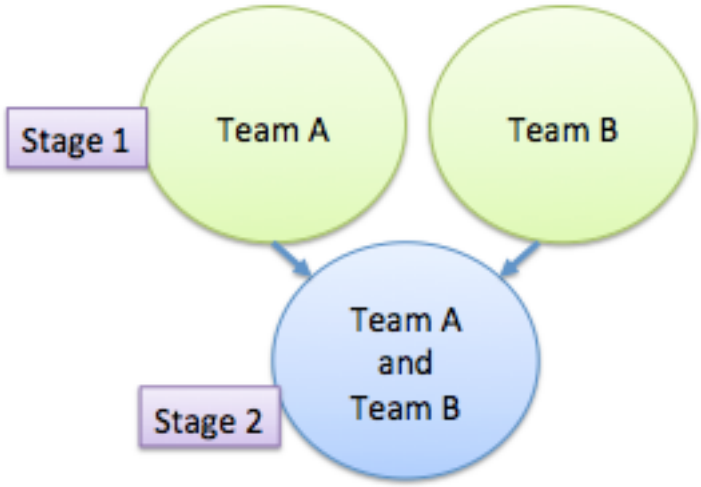

Figure 1. Flow of rating process for students in all classrooms in the agreement subsample. In Stage one, the teams rated the students independently. In Stage two, the teams rated the students together. Team A = Teachers and occupational therapists; Team B = Speech therapists and physical therapists.

All evaluation materials and protocols belong to the PG Chambers School. To analyze the psychometric properties of the PGS- OMT, the transdisciplinary team released the data to the principal investigator. All research information was de-identified and the data received was extant. Transdisciplinary team members gave student participants identification numbers that were used on all evaluation materials. In order to monitor their progress, each participant in the 2013 study was given the same identification number they had in 2014. The principal investigator could not use any research information to identify original participants.

Data Analysis

This study evaluates an alternate assessment for students with multiple disabilities. The reliability of scores from the PGS-OMT, as well as the validity of interpretations drawn from those scores, is assessed. Quantitative techniques are used to analyze the data. Pre-analysis data screening is conducted to assess missing data. Table 4 depicts a summary of the analyses.

Table 4

Data Analytic Plan for Evaluating the Psychometric Properties of the PGS-OMT

	Data Analytic Techniques
Reliability	Coefficient alpha - Overall sample Pearson Correlations - Agreement subsample
Internal Structure	Pearson Correlations - Overall sample Confirmatory Factor Analysis (CFA) - Overall sample, uncorrelated and correlated three-factor models Exploratory Factor Analysis (EFA) - Overall sample, oblique strategy, promax rotation
Change in Scores	Descriptive Statistics - Change sample

Research question 1a. Coefficient alpha is used to determine whether the PGS-OMT yields internally consistent scores. Specifically, coefficient alpha indicates whether the items within their respective domains are adequately correlated. In other words, coefficient alpha indicates the degree of consistency among a set of items that are believed to fit together (Sattler, 2008). According to Murphy & Davidshofer (2005), values .00 to .59 indicate very low reliability, .60 to .69 indicate low reliability, .70 to .79 indicate moderate reliability, .80 to .89 indicate moderately high reliability, and .90 to .99 indicate excellent reliability. To explore items that do not fit as well as other items within the PGS-OMT, item-to-total correlations were calculated at the domain and total scale levels. In other words, the correlations between each item and a scale score that excludes that item were calculated. For example, the correlation was calculated between Social domain item 1 and the total Social domain scale that includes the nine remaining items comprising the Social domain.

Research question 1b. Since students in the agreement subsample were rated three times, Pearson correlations were calculated to determine consistency in ratings. Specifically, Pearson correlations were used to examine cross-informant agreement of the PGS-OMT between the raters from Team A and the raters from Team B, the raters from Team A and the raters from Team A and Team B combined, and the raters from Team B and the raters from Team A and Team B combined. A total of 10 Pearson correlation coefficients, using total domain scores (6), total factor domain scores (3), and the total score (1), were computed for three comparisons – Team A and Team B, Team A and Teams A and B combined, and Team B and Teams A and B combined. Hopkins (2001) gives the following ranges of the magnitudes of correlations: .00-.09 = Trivial, .10-.29 = Small, .30-.49 = Medium, .50-.69 = Large, .70-.89 = Very Large, and .90-1.00 = Almost Perfect.

Research question 2. The internal structure of the PGS-OMT for the overall sample was evaluated via CFA to determine whether the items within the scale appropriately loaded onto the intended constructs (i.e., the three factors). As a preliminary step to CFA, correlation matrices among the PGS-OMT domains and factor domains were produced as shown in Tables 5 and 6. Uncorrelated and correlated three-factor model CFAs were conducted. Absolute fit measures were also calculated, which indicate how well the proposed models fit the data. The four most common absolute fit measures, chi-square, goodness-of-fit index (GFI), the root mean square residual (RMSR), and the root mean square error of approximation (RMSEA), were used (Meyers, Gamst, & Guarino, 2006).

Since the CFA demonstrated an overall poor fit, the next step was to conduct an EFA. Principle axis factoring was first conducted using an unconstrained oblique strategy with a promax rotation, because previous research indicated that all domains comprising the PGS-OMT were correlated (Zahra, 2015). Based on these results, the EFA was conducted a second time by constraining the analyses to three factors, which were the number of factors indicated in the first EFA analysis. The three-factor constrained EFA was similar to the three-factor model rejected by CFA. Since the results of the factor analysis were inconclusive, the EFA was conducted a third time by constraining analyses to two factors. The Kaiser's criterion indicates that only components that have an eigenvalue of 1.0 or more should be extracted. Hair, Black, Babin, and Anderson (2010) indicate that factor loadings .50 or greater are considered significant based on a sample size of 120, power level of 80 percent, and a .05 significance level. See Table 7 for factor analysis predictions.

Table 5
Predicted Correlation Ranges of the PGS-OMT Factor Domains

	Functional Academics	Adaptive Behavior	Inter-personal
Functional Academics	-	-	-
Adaptive Behavior	Medium	-	-
Interpersonal	Large	Large	-

Correlational Ranges
0.10-0.29 = Small 0.30-0.49 = Medium 0.50-69= Large 0.70-0.89 = Very Large
0.90-0.99 = Nearly Perfect

Table 6
Predicted Correlation Ranges of the PGS-OMT Domains

	Social	Communication	Physical Navigation	Personal Care	Academic	Functional
Social	-	-	-	-	-	-
Communication	Very Large	-	-	-	-	-
Physical Navigation	Medium	Medium	-	-	-	-
Personal Care	Large	Very Large	Very Large	-	-	-
Academic	Very Large	Very Large	Medium	Large	-	-
Functional	Very Large	Very Large	Medium	Large	Very Large	-

Correlational Ranges
0.10-0.29 = Small 0.30-0.49 = Medium 0.50-69= Large 0.70-0.89 = Very Large
0.90-0.99 = Nearly Perfect

Table 7

Predicted Factor Loadings for Domain Items for 3F Model

	Factor Loadings		
Domains	Component 1 (Functional Academics)	Component 2 (Adaptive Behavior)	Component 3 (Interpersonal)
Social (1-10)	-	-	High
Communication (1-9)	-	-	High
Physical Navigation (1-9)	-	High	-
Personal Care (1-9)	-	High	-
Academic (1-9)	High	-	-
Functional (1-9)	High	-	-

Note: Factor loadings greater than .50 are considered high

Research question 3. Mean differences and standard deviations among indices were evaluated to determine participants' rate of growth between their 2013 and 2014 PGS-OMT scores. Paired samples t-tests determined whether there were significant differences between mean scores from 2013 and 2014 on each PGS-OMT item. Cohen's d determined the effect size. According to Cohen (1992), an effect size of .20-.49 is a small effect, .50-.79 is a medium effect, and .80 and greater is a large effect.

Chapter III

Results

To address our three research questions about the PGS-OMT, quantitative techniques were used to assess the reliability of the scores, as well as the validity of the interpretations drawn about those scores. Several statistical methods were employed. Pre-analysis data screening was conducted to determine missing data. After assessments of internal consistency and cross-informant agreement were conducted, CFA of two models was used to probe the internal structure of the measures. Finally, the change in scores between 2013 and 2014 was assessed.

Data Cleaning

Frequency tables were examined to determine whether all items were correctly coded in the appropriate 0-7 ranges. Upon examining frequency counts for each PGS-OMT item in the overall 2014 sample, only two data points out of 64 variables for 121 cases in the overall 2014 sample were found to be missing. Data from the 2013 sample was complete.

Based on an examination of the surrounding data points for each participant, the two missing data points were determined to be coding errors. The principal investigator contacted professionals at PG Chambers and was told to change the two missing values to zeros. These data points were therefore subsequently recoded as zeros.

Internal Consistency

To answer research question 1a, data from the overall sample was used to calculate the internal consistency coefficient, Cronbach's alpha, was calculated for each of the six domains, three factor domains, and total scale composite of the PGS-OMT. Refer to Table 8 for reliability coefficients. The PGS-OMT possesses excellent internal consistency at the total scale level (.98)

and Cronbach's alpha values for each domain and factor domain were within the excellent range. As predicted, these scores indicate that items on the scale fit well together. The lower and upper bounds of Cronbach's alpha for each domain, at 95% confidence intervals, were also calculated.

All 64 item-to-total correlations were inspected for all six domains. As Table 8 shows, for each domain, factor domain, as well as the total scale, the lowest item-to-total correlation was well above the commonly noted .30 criterion of acceptability (Field, 2005). Although no items were found to have small item-to-total correlations, seven items at the domain level were found to have large item-to-total correlations of .90 or higher. Five of these item-to-total correlations were in the Physical Navigation domain, one item was in Academic domain, and one item was in the Functional domain. At the factor domain level, four items were found to have large item-to-total correlations of .90 or higher. One of these items was in the Interpersonal factor domain, one was in the Adaptive Behavior factor domain, and two items were in the Functional Academics factor domain.

Table 8
Reliability Coefficients (Cronbach's Alpha) of PGS-OMT Domains, Factor Domains, and Total Scale

	Cronbach's Alpha				
PGS-OMT Factor (# of items)	Estimate	Lower Bound	Upper Bound	Lowest Item-to-Total Correlation	Highest Item-to-Total Correlation
Social (10)	.95	.93	.96	.68	.88
Communication (9)	.95	.93	.96	.72	.86
Physical Navigation (10)	.97	.97	.98	.82	.92
Personal Care (10)	.95	.94	.96	.67	.87
Academic (17)	.97	.97	.98	.69	.92
Functional (8)	.92	.90	.94	.67	.92
Interpersonal (19)	.97	.96	.98	.67	.90
Adaptive Behavior (20)	.98	.97	.98	.69	.90
Functional Academics (25)	.98	.97	.98	.62	.92
Total Scale (64)	.98	.98	.99	.53	.83

Cross-Informant Agreement

To address research question 1b and examine the cross-informant agreement of the PGS-OMT, Pearson correlations were calculated for the agreement subsample data between raters from Team A (Teacher-OT) and the raters from Team B (Speech-PT), between raters from Team A and raters from Team A and Team B combined, and between raters from Team B and raters from Team A and Team B combined. As previously discussed, analyses that relate to the consistency in ratings across different raters are referred to as cross-informant agreement as opposed to inter-rater reliability. Table 9 reports results derived from summing all items in the six domains to create a total domain score for each student for each of the three samples. Pearson correlations were higher than expected and ranged from the large to almost perfect ranges. The

lowest correlation between Team A and Team B was .87 (for the Functional domain), the lowest correlation between Team A and combined raters was .93 (for the Functional domain), and the lowest correlation between Team B and combined raters was .95 (also for the Functional domain).

Table 9
Cross-Informant Agreement

	Pearson Correlation		
Domain (#)	Teacher-OT and Speech-PT	Teacher-OT and Combined	Speech-PT and Combined
Social (10)	.89	.94	.97
Communication (9)	.91	.96	.97
Social + Communication (19)	.92	.96	.97
Physical Navigation (10)	.96	.98	.99
Personal Care (10)	.92	.97	.98
Physical Navigation + Personal Care (20)	.95	.98	.99
Academic (17)	.93	.97	.97
Functional (8)	.87	.93	.95
Academic + Functional (25)	.92	.96	.97
TOTAL	.95	.97	.98

Correlational Ranges
0.10-0.29 = Small 0.30-0.49 = Medium 0.50-69= Large 0.70-0.89 = Very Large
0.90-0.99 = Nearly Perfect
* All correlations are significant at the .05 level (1-tailed test)

Correlations among Domain and Factor Domain Scores within the PGS-OMT

To provide evidence relevant to the second research question, correlations were calculated among all total domain scores based on the combined ratings for the overall sample. Correlations between each domain and the total score were also calculated to determine whether any one domain score is representative of total performance. Correlations among total domain scores are summarized in Table 10. As predicted, correlations were in the medium to nearly perfect range. The correlation between the Academic and Functional domains (r = .88) was in the

large range, the correlation between the Physical Navigation and Personal Care domains (r = .83) was also in the large range, and the correlation between the Social and Communication domains (r = .90) was in the nearly perfect range. All the domains were significantly correlated with each other (.42 or higher). All domains were correlated with the PGS-OMT Total composite score in the very large range (.74 or higher).

Table 10
Correlations of PGS-OMT Domains

	Social	Communication	Physical Navigation	Personal Care	Academic	Functional
Social	-	-	-	-	-	-
Communication	.90*	-	-	-	-	-
Physical Navigation	.48*	.43*	-	-	-	-
Personal Care	.69*	.65*	.83*	-	-	-
Academic	.73*	.79*	.44*	.59*	-	-
Functional	.75*	.79*	.47*	.66*	.88*	-
Total	.87*	.87*	.74*	.87*	.87*	.87*

Correlational Ranges
0.10-0.29 = Small 0.30-0.49 = Medium 0.50-69= Large 0.70-0.89 = Very Large
0.90-0.99 = Nearly Perfect
*Correlation is significant at the 0.05 level (1-tailed)

Summing all items in the same factor domain for each student created a total factor domain score. Correlations were then calculated among all total factor domain scores (Table 11). Correlations between each factor domain and the total score were also calculated to determine whether one factor domain score is representative of total performance. As predicted, correlations were in the large to very large range. The correlations between the three factor

domains and PGS-OMT Total composite score were all in the very large range (r = .84 to r = .89), which suggests that factor domain scores significantly contribute to a student's overall level of performance.

Table 11
Correlations of PGS-OMT Factor Domains

	Functional Academics	Adaptive Behavior	Inter-personal
Functional Academics	-	-	-
Adaptive Behavior	.56*	-	-
Interpersonal	.80*	.59*	-
PGS-OMT Total	.89*	.84*	.89*

Correlational Ranges
0.10-0.29 = Small 0.30-0.49 = Medium 0.50-69= Large 0.70-0.89 = Very Large
0.90-0.99 = Nearly Perfect
*Correlation is significant at the 0.05 level (1-tailed)

Confirmatory Factor Analysis

To further analyze the second research question, data from the overall sample was used to create a three-factor uncorrelated model and a three-factor correlated model via CFA in AMOS. Both correlated and uncorrelated models were divided into the following three latent variables: Functional Academics, Adaptive Behavior, and Interpersonal. The items (observed variables) that comprise the Academic and Functional domains were connected to the Functional Academics latent variable. The items that comprise the Physical Navigation and Personal Care domains were connected to the Adaptive Behavior latent variable. The items that comprise the Social and Communication domains were connected to the Interpersonal latent variable. An error

term was attached to each item. Refer to Figure 2 for the uncorrelated model; the correlated model adds all possible connections between the three latent variables.

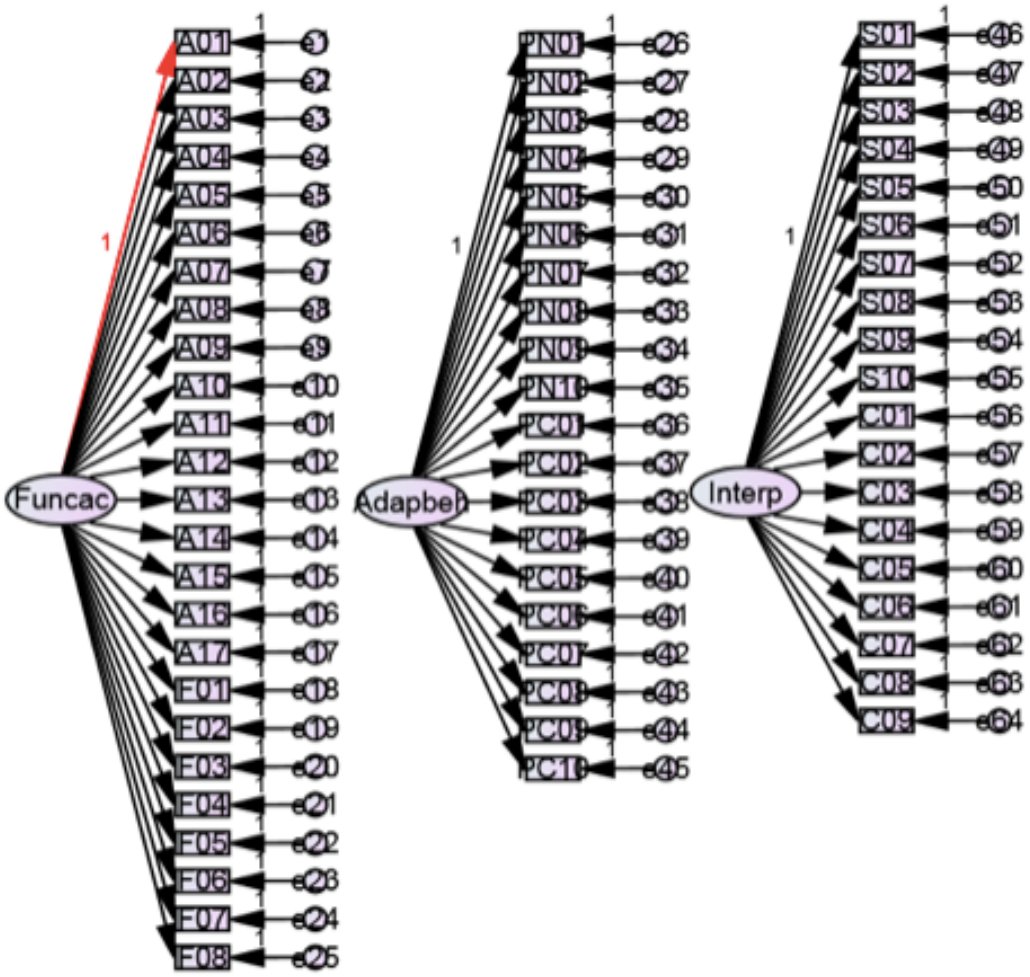

Figure 2. 3F Uncorrelated Model Drafted in AMOS. This figure illustrates the three factor domains, the individual items that comprise each domain, and the corresponding error terms. Funcac = Functional Academics; Adapbeh = Adaptive Behavior; Interp = Interpersonal.

The individual factor loadings were high for both the uncorrelated and correlated three-factor model. Hair, Black, Babin, and Anderson (2010) indicate that factor loadings .50 or greater are considered high based on a sample size of 120, power level of 80 percent, and a .05 significance level.

In the uncorrelated three-factor model, the highest loading in the Functional Academics domain was .93 and the lowest loading was .65. In the Adaptive Behavior domain, the highest loading was .92 and the lowest loading was .68. In the Interpersonal domain, the highest loading was .90 and the lowest loading was .67. Refer to Table 12 for a summary of the uncorrelated factor loadings.

In the correlated three-factor model, the highest loading in the Functional Academics domain was 1.0 and the lowest loading was .24. In the Adaptive Behavior domain, the highest loading was 1.09 and the lowest loading was .52. In the Interpersonal domain, the highest loading was 1.40 and the lowest loading was .85. Refer to Table 13 for a summary of the correlated factor loadings.

Table 12

CFA Loadings for Uncorrelated 3F Model

Domain/ Item	Functional Academics
A01	0.73
A02	0.84
A03	0.76
A04	0.89
A05	0.81
A06	0.71
A07	0.84
A08	0.75
A09	0.93
A10	0.83
A11	0.93
A12	0.89
A13	0.79
A14	0.81
A15	0.90
A16	0.90
A17	0.87
F01	0.90
F02	0.90
F03	0.74
F04	0.75
F05	0.80
F06	0.78
F07	0.77
F08	0.65

Domain/ Item	Adaptive Behavior
	0.82
PN01	
PN02	0.89
PN03	0.92
PN04	0.92
PN05	0.83
PN06	0.89
PN07	0.85
PN08	0.88
PN09	0.90
PN10	0.90
PC01	0.84
PC02	0.83
PC04	0.68
PC05	0.79
PC06	0.74
PC09	0.72
PC10	0.73

Domain/ Item	Interpersonal
S01	0.71
S02	0.86
S03	0.91
S04	0.75
S05	0.88
S06	0.80
S07	0.81
S08	0.80
S09	0.67
S10	0.68
C01	0.87
C02	0.85
C03	0.77
C05	0.85
C07	0.77
C08	0.83
C09	0.83

Table 13

CFA Loadings for Correlated 3F Model

Domain/ Item	Functional Academics	Domain/ Item	Adaptive Behavior	Domain/ Item	Interpersonal
A01	1	PN01	1	S01	1
A02	0.84	PN02	1.04	S02	1.29
A03	1.06	PN03	0.94	S03	1.09
A04	0.77	PN04	0.98	S04	1.06
A05	1.04	PN05	1.05	S05	1.30
A06	0.62	PN06	1.01	S06	1.04
A07	1.12	PN07	0.95	S07	1.17
A08	0.83	PN08	1.09	S08	1.17
A09	1.03	PN09	0.97	S09	0.97
A10	1.02	PN10	1.02	S10	1.02
A11	0.88	PC01	0.78	C01	1.39
A12	0.61	PC02	0.70	C02	1.21
A13	0.35	PC04	0.52	C03	1.14
A14	0.85	PC05	0.94	C05	0.85
A15	0.85	PC06	0.91	C07	1.16
A16	0.79	PC09	0.69	C08	0.90
A17	0.85	PC10	0.85	C09	1.40
F01	0.96				
F02	0.75				
F03	0.58				
F04	0.93				
F05	0.44				
F06	0.55				
F07	0.39				
F08	0.24				

Various fit indices were calculated between the two hypothesized three-factor models and the observed data. Refer to Table 14 for Goodness of Fit Indices for the three-factor uncorrelated and correlated models.

Absolute fit measures. Since a close fit was predicted between the hypothesized three-factor model and observed data, a non-significant chi-square value is expected. Chi-square for the uncorrelated model was found to be significant ($\chi^2 = 5800.98, p < .05$). The Chi-square for the correlated model was also found to be significant ($\chi^2 = 56492.69\ p < .05$). This indicates that neither hypothesized three-factor model is a good fit for the data.

The GFI indicates the proportion of variance in the sample covariance explained by the predicted model. The values range from 0, indicating no fit to 1.0, indicating a perfect fit. To be indicative of an acceptable model, the GFI should be equal to or greater than .90. The GFI for the thee-factor uncorrelated model was .38 and for the correlated model was .37. This again indicates that neither hypothesized three-factor model is a good fit for the data.

The RMSR is a measure of the average size of the residuals between the proposed model and the data. RMSR values that are smaller (< .05) indicate a better fit. The RMSR was 1.78 for the uncorrelated model and 0.52 for the correlated model. This indicates that the correlated model is a better fit to the data.

The RMSEA is the average of the residuals between the observed covariance from the sample and expected covariance estimated for the population. Values less than .08 are considered acceptable and values greater than .10 are considered unacceptable. The RMSEA was 0.13 for the uncorrelated model and 0.13 for the correlated model. This again indicates that neither hypothesized three-factor model is a good fit for the data.

Relative fit measures. Relative fit measures compare the fit between the hypothesized model and a null model that assumes no relationships in the data. Relative fit measures include the comparative fit index (CFI) and the normed fit index (NFI).

The CFI assesses the proposed model by indicating the discrepancy between the hypothesized model and the actual data. Values greater than .90 are considered a good fit, values from .80 to .89 are considered an adequate fit, and values from .60 to .79 are considered a poor fit (Meyers, Gamst, & Guarino, 2006). The CFI value was .66 for the uncorrelated model and .68 for the correlated model. This again indicates that neither hypothesized three-factor model is a good fit for the data.

The NFI reflects the discrepancy between the hypothesized model and the null model. Values greater than .90 are considered acceptable. The NFI value was .57 for the uncorrelated model and .58 for the correlated model. This again indicates that neither hypothesized three-factor model is a good fit for the data.

Table 14

Three-Factor Correlated and Uncorrelated Fit Indices

Fit Indices	3F Model Uncorrelated	3F Model Correlated
Absolute Fit Measures		
Chi Square	5800.98	5642.69
GFI	0.38	0.37
RMSR	1.781	0.52
RMSEA	0.13	0.13
Relative Fit Measures		
CFI	0.66	0.68
NFI	0.57	0.58

Exploratory Factor Analysis. Since both the uncorrelated and correlated three factor models proved to be poor fits to the data, an exploratory factor analysis (EFA) was conducted in SPSS to probe further into the underlying structure of the data. Prior to conducting the EFA, the data was assessed to determine whether it was adequate for factor analysis. The Bartlett's Test of Sphericity was significant (χ^2 [2016] = 10921.04, $p < .05$), indicating that factor analysis was suitable for this data.

Since factor correlations were in the high range, principal axis factoring was first conducted using an unconstrained oblique strategy with a promax rotation (See Table 15). Coefficients less than .30 were suppressed.

Table 15

Unconstrained EFA

Factor Model	Eigenvalue	High Loadings	Low Loadings	Total Variance Explained
First Factor	33.99	1.00	-.00	53%
Second Factor	7.93	.96	-.00	65%
Third Factor	3.72	.94	-.00	71%
Fourth Factor	2.18	.89	-.00	75%
Fifth Factor	1.57	.92	-.00	77%
Sixth Factor	1.35	.79	.00	79%
Seventh Factor	1.00	.64	.01	81%

In the unconstrained EFA model, seven factors had an eigenvalue greater than 1.0, and the seven factors accounted for 81% of the variance (See Table 15). The first six factors had an eigenvalue greater than 1.35 and accounted for 79% of the variance. However, the sixth factor had minimal item loadings and when the loadings were above .70 the Communication and Academic domains were divided. The fifth factor had an eigenvalue of 1.57 and accounted for 77% of the variance, although all loadings above .50 were in the Personal Care domain. The fourth factor had an eigenvalue of 2.18 and accounted for 75% of the variance, although it generally had low loadings and all factors above .50 were in the Academic domain. The third factor had an eigenvalue of 3.72 and accounted for 71% of the variance, and divided the Social and Communication domains. The second factor had an eigenvalue of 7.93 and accounted for 65% of the variance, and divided the Physical Navigation domain and two items of the Personal

Care domain. The first factor had an eigenvalue of 33.99 and accounted for 53% of the variance, and divided the Academic and Functional domains.

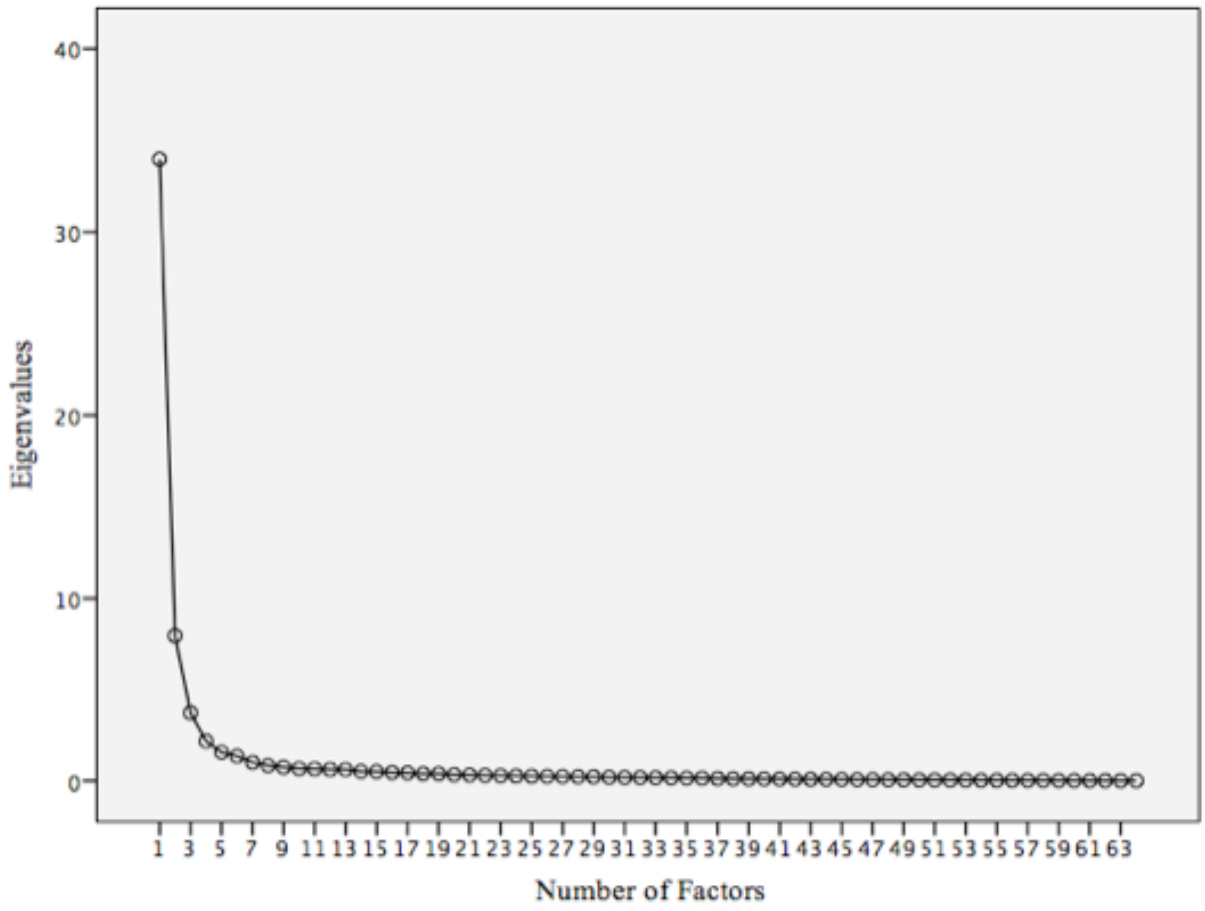

Figure 3. EFA Scree Plot. This figure illustrates the number of extracted factors and corresponding eigenvalues from the overall 2014 data sample.

A scree plot of the eigenvalues (Figure 3) demonstrated that the loadings were best explained by three factors with eigenvalues ranging from 33.99 to 3.72. The EFA was therefore conducted a second time by constraining the analyses to three-factors and suppressing small

coefficients with an absolute value less than .40. This model had an eigenvalue of 3.72 and accounted for 71% of the variance (See Table 16 for factor loadings).

On the first factor, 15 out of the 17 items of the Academic domain loaded and 7 out of the 8 items of the Functional domain loaded. The highest factor loading was 1.03 and the lowest loading was 0.48. Academic domain item 8 loaded more on the third factor and item 1 loaded solely on the third factor. Functional domain item 4 loaded on the third factor. This factor is Functional Academics.

On the second factor, 9 out of the 10 items of the Personal Care domain loaded and all 10 items of the Physical Navigation domain loaded. The highest factor loading was 1.01 and the lowest loading was 0.43. Personal Care domain item 6 loaded more on the third factor. This factor is Adaptive Behavior.

On the third factor, 7 out of the 9 items of the Communication domain loaded and all 10 items of the Social domain loaded. The highest factor loading was 0.98 and the lowest loading was 0.56. Communication domain items 6 and 7 loaded more on this factor and slightly less on the first factor. This factor is Interpersonal.

Table 16

EFA Item Loadings for Constrained 3F Model

Item	Factor 1
A02	.90
A03	.48
A04	.95
A05	.71
A06	.65
A07	.76
A09	.88
A10	.57
A11	.93
A12	1.03
A13	1.02
A14	.61
A15	.92
A16	.97
A17	.93
F01	.84
F02	.91
F03	.74
F05	.91
F06	.53
F07	.90
F08	.74

Item	Factor 2
PC01	.77
PC02	.76
PC03	.43
PC04	.71
PC05	.71
PC07	.47
PC08	.64
PC09	.80
PC10	.80
PN01	.87
PN02	.96
PN03	.94
PN04	.94
PN05	1.01
PN06	.98
PN07	.92
PN08	.97
PN09	.92
PN10	.88

Item	Factor 3
C01	.86
C02	.87
C03	.90
C04	.63
C05	.91
C08	.62
C09	.81
S01	.97
S02	.68
S03	.84
S04	.73
S05	.86
S06	.65
S07	.98
S08	.89
S09	.56
S10	.77

The EFA was conducted a third time by constraining the analyses to two factors and suppressing small coefficients with an absolute value less than .40. This model had an eigenvalue of 7.92 and accounted for 65% of the variance (See Table 17 for factor loadings).

On the first factor, all Social, Communication, Academic, and Functional items loaded. The highest factor loading was .998 and the lowest loading was 0.455. Personal Care item 3 loaded more on the first factor.

On the second factor, all the Physical Navigation and Personal Care items loaded. The highest factor loading was 1.00 and the lowest loading was .442. Personal Care item 7 loaded more on the second factor and slightly less on the first factor.

Table 17

EFA Item Loadings for Constrained 2F Model

Item	Factor 1
A16	.998
A12	.951
A09	.936
C06	.933
A11	.911
A04	.896
F01	.894
F02	.878
A15	.870
A17	.862
A13	.860
A02	.858
A08	.839
F05	.837
C08	.820
C01	.814
A10	.813
C07	.809
F07	.809
A07	.797
A14	.786
C09	.774

Item	Factor 1
F06	.757
A06	.752
A03	.750
S03	.749
A05	.746
F08	.726
C04	.719
F03	.711
S06	.711
A01	.691
S02	.671
C02	.664
S04	.655
F04	.653
S05	.633
S09	.581
S08	.563
C03	.547
S10	.54
C05	.538
S07	.533
S01	.455
PC03	.455

Item	Factor 2
PC03	.442
PN05	1.004
PN06	.986
PN08	.984
PN09	.945
PN02	.928
PN04	.909
PN03	.901
PN10	.881
PC09	.876
PN07	.865
PN01	.842
PC10	.804
PC01	.790
PC02	.785
PC04	.742
PC05	.731

Change in Scores

To address the third research question regarding change in student scores between 2013 and 2014, data from the change sample was used to calculate descriptive statistics (e.g., ranges, means, and standard deviations) on the domain and factor domain levels (See Table 17). Differences between each pair of 2013 and 2014 domain, factor domain, and total scores were calculated. For example, the total score for an individual on the Social domain in 2013 was subtracted from the total score for that individual on the Social domain in 2014. Descriptive statistics were calculated for the differences for each domain, factor domain, and total score. Paired-sample t-tests of the differences of means were also calculated.

The Social domain demonstrated no change in scores between 2013 and 2014. The mean score for 2013 was 29.28 and the mean score for 2014 was 29.19. The t-test showed no significant difference and the Cohen's d indicated no effect. For domain scores, the Personal Care and Academic domains demonstrated significant improvement from 2013 to 2014. For the factor domain scores, the Adaptive Behavior and Functional Academics factor domains demonstrated significant improvement from 2013 to 2014. The total scale scores also demonstrated significant improvement from 2013 to 2014. All these t-tests were significant and all Cohen's d tests indicated small effects.

The remaining domain scores, including the Communication, Physical Navigation, and Functional domains, yielded a significant effect that is very small. The Interpersonal factor domain also yielded a significant effect that is very small. While the t-test of differences were significant, Cohen's d showed no effects.

Table 18

Change in Scores 2013-2014

	Change (2014 - 2013 Scores)					t-test		Effect Size
Domain (#)	Range	Min	Max	Mean	SD	Mean 2013	Mean 2014	Cohen's d
Social (10)	59.00	-29.00	30.00	-.08	12.75	29.28	29.19	-0.01
Communication (9)	62.00	-30.00	32.00	.97	9.13	22.32	23.29	0.11*
Social + Communication (19)	118.00	-59.00	59.00	.89	20.18	51.59	52.48	0.04*
Physical Navigation (10)	51.00	-27.00	24.00	.44	8.56	28.79	29.22	0.05*
Personal Care (10)	38.00	-12.00	26.00	2.49	6.81	18.87	21.36	0.37*
Physical Navigation + Personal Care (20)	71.00	-34.00	37.00	2.93	12.30	47.65	50.58	0.24*
Academic (17)	71.00	-29.00	42.00	5.11	14.28	22.78	27.89	0.36*
Functional (8)	46.00	-22.00	24.00	1.24	6.55	6.89	8.12	0.19*
Academic + Functional (25)	101.00	-51.00	50.00	6.35	18.94	29.66	36.01	0.34*
Total (64)	219.00	-108.00	111.00	10.16	37.94	128.91	139.07	0.27*

* t-tests of differences of means (paired samples) were significant at the 0.05 level (1-tailed)

Chapter IV

Discussion

Although alternate assessments are mandated for students with severe disabilities, previous research has shown that these measures generally do not represent students' academic skills in a psychometrically sound way (Browder et al., 2005; Elliott, Compton, Roach, 2007; Laitusis et al., 2014). The study's aim was therefore to determine how meaningfully students at PG Chambers are being evaluated by gaining insight into the psychometric properties of the PGS-OMT.

The reliability of this tool was assessed by analyses of its internal consistency and cross-informant agreement. Reliability analyses using Cronbach's alpha indicated that scores demonstrated excellent internal consistency, and Pearson correlations indicated that scores demonstrated high cross-informant agreement.

The internal structure of the PGS-OMT was assessed through correlations among the subscales, confirmatory factor analysis (CFA), and exploratory factor analysis (EFA). Internal structure analyses were not conclusive. CFA yielded results that were difficult to interpret because of high factor loadings for both the uncorrelated and the correlated three-factor model coupled with otherwise poor fit indices. EFA yielded findings that were similar to the three-factor model rejected by the CFA.

Change in scores from 2013 to 2014 of the PGS-OMT was assessed through descriptive statistics, paired samples t-tests, and Cohen's d. Change in scores analyses yielded significant t-tests indicating that students at PG Chambers are generally improving in particular areas from year to year.

Reliability

The reliability of the PGS-OMT was evaluated through analyses of internal consistency and cross-informant agreement. The internal consistency of the PGS-OMT was evaluated for the overall sample using Cronbach's alpha and item-to-total correlations. The goal was to determine whether items within each domain, as well as all the items comprising the complete scale, fit well together. Results indicate that the PGS-OMT possesses excellent internal consistency at the domain and total scale levels, which suggests that the items within each domain, as well as all the items comprising the complete scale, fit well together. No items had small item-to-total correlations. Seven items had large item-to-total correlations at the domain level and four items had large item-to-total correlations at the factor domain level.

The cross-informant agreement of the PGS-OMT was evaluated for the agreement subsample using Pearson correlations. The goal was to determine whether professionals at PG Chambers are in agreement about students' mastery levels of all areas measured by the PGS-OMT. Pearson correlations were higher than expected, and indicate that the PGS-OMT yields scores that demonstrate adequate cross-informant agreement.

Internal Consistency. Research question 1a addressed the internal consistency of the PGS-OMT for the overall sample (See Table 8). Cronbach's alpha was calculated for the scales for each domain and for the total scale. As predicted, the PGS-OMT possesses excellent internal consistency at the domain (.92 and higher) and total scale levels (.98). This indicates that the items within each domain, as well as all the items comprising the complete scale, fit well together.

Overall, and taking account of the potential modifications suggested above, these results indicate that the PGS-OMT yields scores that demonstrate adequate internal consistency. The findings were similar to Zahra's (2015) results.

Cross-informant agreement. Research question 1b addressed the cross-informant agreement of the PGS-OMT as assessed by the agreement subsample data (Table 9). As previously discussed, although raters are similar because they based their ratings of the students on student performance in the same classroom, the raters are different because they had different experiences with the students due to the different roles they played in the classroom. Therefore, these analyses most appropriately, but as seen shortly, not completely, fall under the category of cross-informant agreement, which refers to the degree to which different types of raters (e.g., teachers and therapists) who are aware of a child's functioning in different environments agree on the scores on the same instrument (Phye, Saklofske, Andrews, Janzen, 2001; Achenbach & Rescorla, 2007; Sointu, Savolainen, Lappalainen, Epstein, 2012).

Pearson correlations were calculated for each of the three combinations of raters – Team A (Teacher-OT), Team B (Speech-PT), and Team A and Team B combined. Pearson correlations were higher than expected and ranged from the large to the almost perfect range. These results indicate that the PGS-OMT yields scores that demonstrate adequate cross-informant agreement.

This finding should be understood in the context of Meyer et al.'s (2001) review of more than 125 meta-analyses on test validity. Specifically, the authors indicate that individual assessment methods provide unique information for children when various types of knowledgeable informants are compared with each other. For example, teacher ratings have only moderate agreement with clinician ratings (r = .34). The authors conclude that any single assessment method provides an incomplete representation of the constructs it intends to measure

and it is therefore difficult to obtain consistent information about patients from different types of raters. The results of the current study are more optimistic and suggest that teams of different types of raters at PG Chambers can provide a complete representation of the constructs that PGS-OMT intends to measure and that consensual agreement about student performance can indeed be obtained between teams of different types of raters.

In a meta-analysis, Achenbach, McConaughy, and Howell (1987) reviewed 119 studies on childhood behavioral and emotional problems to investigate cross-informant agreement across different types of raters. The authors indicate that cross-informant agreement is generally low (r = .28) across different types of raters (e.g., parents and teachers), and that these judgments are not usually meaningful. In spite of their conclusions, the results of our study are again more optimistic. Even though the raters had different experiences with students in the same classroom, their evaluations converged to a great degree. Our findings therefore indicate that different types of professionals at PG Chambers are in agreement about students' mastery levels of all areas measured by the PGS-OMT.

One reason for the discrepancy between our results and the previous research might be that the PGS-OMT raters observed students in the same classroom context. De Los Reyes et al. (2015) discuss how children may display mental health concerns in some contexts and not others. Their meta-analysis reviewed cross-informant agreement of children's internalizing and externalizing mental health concerns and found low-to-moderate correlations of correspondence (mean internalizing: r = .25; mean externalizing: r = .30). It is important to recognize that the PGS-OMT raters know the students they rated from the same classroom in PG Chambers. If raters had tried to rate students across a variety of different contexts (e.g., school and home), results may have been different.

In sum, although the raters had different experiences with the students they knew them from the same classroom context. Moreover, there were teams of raters rather than individual raters. Therefore these analyses do not fall completely under the category of cross-informant agreement, nor do they fall completely under the category of inter-rater reliability. This overlap helps explain why our results may differ from the previously published research of cross-informant agreement.

Internal Structure Validity

The internal structure of the PGS-OMT for the overall sample was evaluated using correlations among domains and factor domains (Tables 10 and 11), confirmatory factor analysis (Tables 12-14), and exploratory factor analysis (Tables 15-16). The goal was to determine whether items in the scales appropriately loaded on to the following three factors: Functional Academics (combination of the Functional and Academic domains), Adaptive Behavior (combination of the Personal Care and Physical Navigation domains), and Interpersonal (combination of the Social and Communication domains).

Correlations among all six domains were at least in the medium range. This demonstrates that all the domains comprising the PGS-OMT are inter-related. Correlations between the two domains that were hypothesized to comprise each of the three factor domains ranged from the very large to the almost perfect ranges. While this correlational finding supports the three-factor model, results from the CFAs of the three-factor model were inconclusive, as indicated by cross-factor loadings and poor fit indices (i.e., chi square, GFI, RMSR, RMSEA, CFI, and NFI). Nevertheless, based on the constrained and unconstrained EFAs, the three-factor model may indeed be one of a couple of defensible models.

Correlations. To provide evidence on the second research question about the internal structure validity of the PGS-OMT, correlations were calculated among domain and factor domain scores using the combined ratings for the overall sample. Correlations among the domains were consistent with the three-factor model. The correlation between the Social and the Communication domains was .90, between the Physical Navigation and Personal Care domains .83, and between the Academic and Functional domains .88. Domains from the six-factor model that were unrelated to the three-factor model correlated in the medium to very large ranges. For example, the correlation between the Communication and Physical Navigation domains was .43 and the correlation between the Academic and Communication domains was .79. All domains correlated with the total scale in the very large range. The Physical Navigation domain had the lowest correlation with the total scale (.74).

Correlations among the factor domains ranged from the large to the very large ranges. The largest correlation was between the Functional Academics and Interpersonal factor domains (.80). This is a very high correlation between two domains that are supposedly measuring different constructs. The smallest correlation was between the Adaptive Behavior and Functional Academics factor domains (.56). All factor domains correlated with the total scale in the very large range. The Adaptive Behavior domain had the lowest correlation with the total scale (.84).

Some of the high correlations between domains and factor domains may indicate that these domains and factor domains are measuring the same construct. For example, the correlation between the Social and Communication domains is .90; the 81% shared variance between the scores may indicate that these domains are measuring a similar construct. The correlation between the Functional Academics and Interpersonal factor domains is .80; the 64%

shared variance between the scores may indicate that these factor domains are measuring a similar construct.

Alternatively, the high correlations between domains may suggest that while the domains are highly inter-related they still measure different constructs because one construct is a prerequisite of another construct. For example, students who perform low on the Social domain may be more likely to perform low on the Communication domain. Students who perform low on the Functional domain may be more likely to perform low on the Academic domain. Our findings are consistent with Zahra's (2015) study, which provides more evidence about these possibilities.

The strong correlation between the Academic and Functional domains is of particular importance. In the early development of alternate assessment, most states incorporated functional domains in their assessments. IDEA eventually encouraged states to incorporate more academic domains in alternate assessment and AA-AAS was then mandated to be aligned with the CCSS (U.S. Department of Education, 2005). In Elliott, Compton, and Roach's (2007) multitrait-multimethod (MTMM) validation study of the Idaho Alternate Assessment (IAA), the IAA reading, language arts, and mathematic scales all shared more variance with adaptive behavior than with measures of academic skills. Kettler's et al. (2010) MTMM study also revealed that scores from states' AA-AAS reflect a construct that is strongly related to adaptive behavior.

The current study also implies that the PG Chambers measure does not distinguish academic performance from functional skills. A possible reason in our data is that these two domains share the same anchors that guide the use of the 7-point Likert scale, and that these anchors are different than the anchors applied to the other four domains. Suggestions for applying the same skill definitions to all domains will be discussed shortly.

Confirmatory Factor Analysis. To provide further evidence on the second research question, confirmatory factor analysis (CFA) in AMOS was conducted. The choice to use CFA to evaluate the PGS-OMT was based on both the measure's theoretical framework and Zahra's (2015) findings. Specifically, the PGS-OMT is divided into six domains and 64 items that professionals at PG Chambers thought would be representative of students' mastery levels. Zahra (2015) tested the original 64 items using an EFA and found support for three-factor model, which was the basis for the theoretical three-factor model used in this study.

Using data from the overall sample, a three factor uncorrelated model and a three factor correlated model were therefore tested. Unfortunately, results were difficult to interpret due to the poor fit indices for both models. Based on Meyers, Gamst, and Guarino's (2006) guidelines for appropriate values of fit indices for factor models, both the uncorrelated and correlated CFAs were therefore rejected.

Exploratory Factor Analysis. As previously stated, the CFA was used to test the three-factor model that had been found in the previous study by Zahra (2015). Since EFAs are conducted based on empirical observation rather than theoretical deduction, the EFA statistical method was chosen as the logical next step to explore the second research question about the internal structure validity of the PGS-OMT.

An EFA unconstrained oblique strategy with a promax rotation analysis revealed that the loadings could possibly be explained by three factors. Several pieces of evidence led to this possibility. First, the scree plot showed a decline in variance explained after the third factor was extracted. Second, the fourth, fifth, and sixth factors either had minimal item loadings or divided a domain whose items were better explained by another factor, leaving many items poorly unexplained. For example, while the fourth factor accounted for 75% of the variance, it generally

had low loadings, and all loadings above .50 were in the Academic domain. Third, when the EFA was conducted a second time by constraining the analyses to three factors, the model accounted for 71% of the variance. Finally, our findings were very similar to the three-factor model found in Zahra's (2015) work.

However, the constrained EFA also revealed that the three-factor model consisted of the two domains, comprising each factor domain loading together, that the CFA rejected. The second factor accounted for 65% of the variance and divided the Physical Navigation domain and two items of the Personal Care domain. The first factor accounted for 53% of the variance, and divided the Academic and Functional domains.

Given these inconclusive results, and given that all domains are inter-related, one might argue that the PGS-OMT measures one domain instead of three. After reviewing all the findings from the correlations, CFA, and EFA, the one-factor conclusion seems unwarranted for three reasons. First, if there were only one underlying factor then all the correlations would be in the very large range. However, the data show that some domains only have medium (e.g., Physical Navigation and Communication) or large (e.g., Academic and Personal Care) correlations. Second, one factor explained only 53% of the variance. Third, since the three-factor uncorrelated and correlated models were rejected, results from the CFA might have indicated that there was only one underlying factor. However, the subsequent EFA revealed that there are three underlying factors.

Although a one-factor model is inconsistent with our results, another possibility is a two-factor model, which combines the Functional Academics and Interpersonal factor domains and separates the Adaptive Behavior factor domain. This two-factor structure may not have been apparent in the three-factor constrained EFA because the Functional Academics and

Interpersonal factor domains may only be weakly related under certain circumstances. For example, students with social and communication skills may be better able to learn functional and academic skills. Therefore, high performance on the Social and Communication domains may be a prerequisite for high performance on the Academic and Functional domains. The two-factor constrained EFA indeed demonstrated this split. However, the two-factor model accounted for 65% of the variance, while the third-factor model raised the variance explained to 71%.

Overall, the results of the factor analyses are therefore inconclusive. The CFA rejected the three-factor model, the subsequent EFA revealed a three-factor model, and a more sophisticated two-factor model might possibly explain the different findings. Additional factor analyses with larger sample sizes may be warranted. However, a larger sample size may only make this misfit more apparent, because the cross loadings of the items will make it difficult to separate these constructs in the population. It is most likely that the PGS-OMT would have to be redesigned to revise items with high cross loadings. Refer to Appendix B for specific items that should be revised.

Change in Scores

The third research question addressed change of student scores on the PGS-OMT between 2013 and 2014 (Table 16). Descriptive statistics including paired samples t-tests and Cohen's d were calculated for the participants in the change sample at the domain, factor domain, and total scale level.

The Social domain was the only score that demonstrated no significant change between 2013 and 2014 scores. Four scores, representing the Communication domain, Interpersonal factor domain, Physical Navigation domain, and Functional domain, had relationships indicated by significant t-tests, yet were substantively negligible because Cohen's d was in the no effect

range (d = .04-.19). As shown by significant paired samples t-tests and effect sizes in the small range (d = .24-.37), the other five scores, representing the Personal Care domain, Adaptive Behavior factor domain, the Academic domain, the Functional Academics factor domain, and the total, demonstrated some improvement from 2013 to 2014.

Overall, results indicate that students' PGS-OMT scores increased in the small or negligible range from year to year. The exception is the Social domain, which showed no significant change.

Practical Implications

The results of the current study have practical implications for reformatting the PGS-OMT and for enhancing education at PG Chambers.

Reformatting the PGS-OMT. Guidelines for reformatting the PGS-OMT differ depending on how transdisciplinary team members want to revise the tool. If transdisciplinary team members want to keep the six constructs measured in each domain separate and are willing to revise items, revisions to the PGS-OMT should ensure that the content measured by one domain does not require skills from another domain. Correlation results indicate that the targeted constructs across domains and factor domains were unclear and require refinement. The domain and factor domain correlations ranged from .43 to .90, with 65% of the correlations higher than .70. Most of the correlations account for almost 50% of the variance explained, which indicates that most of the domains and factor domains are highly inter-related.

Cross loadings from the unconstrained EFA reveal that content measured by one domain may require the skills from another domain. For example, Functional domain item 4 (i.e., demonstrates memory of routines/structures from day to day) loaded on the Functional Academics factor (.36) and the Interpersonal factor (.38). This item incorporates both functional

and communication skills, which may explain this cross loading. Social domain item 6 (i.e., copes with disappointment) loaded on the Functional Academics factor (.40) and the Interpersonal factor (.65). This item incorporates both functional and social skills, which may explain this cross loading. Some items may be measuring additional constructs as seen by items loading on the fourth, fifth, and sixth factors. For example, Personal Care domain item 9 (i.e. completes steps for hand washing) loaded on the Adaptive Behavior factor and the fifth factor (.66), which may indicate that this item measures personal care skills and unrelated skills such as following directions. In sum, if transdisciplinary team members want to keep the six constructs measured in each domain separate, they need to revise items to ensure that each construct is assessed independently.

On the other hand, if transdisciplinary team members are open to combining domains, perhaps because they are not willing to revise items, specific recommendations are made regarding which domains to combine. The largest correlation is between the Social and the Communication domains (r = .90). With 81% of the variance explained, this result indicates that these two domains are inter-related and some individual items comprising the Social domain may be redundant with some individual items comprising the Communication domain. One suggestion is therefore that items from both the Social and Communication domains be incorporated into a single larger domain, possibly titled as the Interpersonal domain. Items from the Functional and Academic domains should also be incorporated into another single larger domain, possibly titled as the Functional Academics domain. The correlation between these domains explains more than 77% of the variance, and transdisciplinary team members should also consider combining these domains. Items from the Physical Navigation and Personal Care domains should also be incorporated into another single larger domain, possibly titled as the

Adaptive Behavior domain. The correlation between these domains explains almost 70% of the variance, and transdisciplinary team members should also consider combining these domains.

There is another possible way to combine domains. The rejection of the three-factor model by the CFA indicates that three factors are too many rather than too few and a two-factor model, which combines the Interpersonal and Functional Academics factor domains and separates the Adaptive Behavior factor domain, may more appropriate. The correlation between the Interpersonal and Functional Academics factor domains is .80, which explains 64% of the variance. Furthermore, the two-factor constrained EFA demonstrated that the Interpersonal and Functional Academics factor domains loaded together on one factor and the Adaptive Behavior domain loaded on a second factor. A fourth suggestion is therefore that transdisciplinary team members think about how to combine items from the Social, Communication, Functional, and Academic domains into one large domain, and about how to combine the Physical Navigation and Personal Care domains into a second large domain.

Further evaluations could be conducted with the changes transdiscipinary team members want to make. The purpose would be to determine whether a positive change results in the empirical evaluation of the tool.

Two final suggestions relate to the wording and structure of the PGS-OMT measurement instrument. Transdisciplinary team members should consider applying the same anchors that guide the use of the 7-point Likert scale to all domains. Currently, the Academic and Functional domains have different anchors than the other four domains. Since the Academic and Functional domains are highly correlated, the different anchors may impact results. Therefore, the same anchors should be applied to all domains. Our last suggestion is to redesign the order of the domains listed on the PGS-OMT. Currently, raters complete the PGS-OMT in the following

order: Social, Communication, Physical Navigation, Personal Care, Academic, and Functional. The domains that are highly correlated are adjacent on the actual tool. To rule out that the ordering of the domains influences the results, it is recommended that transdisciplinary team members reorder the domains on the tool randomly, possibly having three or four different domain combinations.

Enhancing Education at PG Chambers. The results regarding change in scores indicate that in some areas PG Chambers students are improving year to year. As indicated by significant t-tests and small effect sizes, students improved the most on the Personal Care and Academic domains from 2013 to 2014. Instruction and IEP goals should be tailored to reflect that students are generally improving in these areas. Students demonstrated no change on the Social domain between 2013 and 2014. Perhaps lessons that teach students the social skills evaluated in the Social domain should become more of a focus at PG Chambers. Browder et al. (2005) recommended that alternate assessments be aligned with IEP goals. Professionals at PG Chambers should therefore tailor IEP goals to be more reflective of students' changing performance on the PGS-OMT.

Limitations

The generalizability of this study's results is limited due to a variety of factors. All participants attend PG Chambers, and this convenience sample limits the applicability of this study's results beyond PG Chambers to the general population of students who participate in AA-AAS. In addition, all of our student samples were comprised mostly (at least 70%) of White/Non-Hispanic students, which also limit the generalizability of our results. Moreover, the rater sample was homogeneous in terms of race (98% were White/Non-Hispanic) and gender (all female), which may limit the generalizability of the conclusions drawn from the cross-informant

agreement analyses. Additionally, the rater sample was divided into Team A and Team B based on profession (i.e., all teachers and occupational therapists were in one group and all speech therapists and physical therapists were in another group). If a random assignment of raters was employed, results from the cross-informant agreement analyses may be more valuable because they would ensure that any differences between and within the groups are not systematic from the onset of the study. Raters between teams may have also collaborated regarding PGS-OMT ratings. The principal investigator was not able to monitor this potential clinical collaboration, which may limit the internal validity of this study.

The number of participants in this study is another important limitation. Comrey and Lee (1992) provide a general evaluation of the adequacy of various sample sizes for factor analysis and explain that sample sizes within the 100-200 range are considered poor. Meyers, Gamst, & Guarino (2006) suggest that analyses with greater than 10 variables require no less than 200 subjects to ensure stable parameter estimates. The PGS-OMT has 64 variables, which suggests that the sample size of 121 is not a substantial base to draw adequate conclusions.

Future Research

More research regarding the psychometric properties of alternate assessments should be conducted. Specifically, additional research is needed to analyze the reliability and validity of the PGS-OMT.

Several suggestions for future research involve improving the external validity of the evidence regarding the PGS-OMT. One recommendation is to administer the PGS-OMT to other students outside of PG Chambers. Other charter schools, private schools, or public schools that have a population of students who participate in AA-AAS could qualify to participate in this future study. A comparison of the samples could facilitate generalizability of the PGS-OMT

results to another context. Another suggestion is to divide the raters randomly into groups, which will make results more generalizable across groups. Raters could also be divided into two groups based on years of professional experience. Groups could rate students and results could be compared to determine whether more experienced raters rate students differently than raters with less experience.

To gather additional information about the cross-informant agreement of the PGS-OMT, it is also recommended that future studies include parents/guardians in the evaluation process. A parent version of the PGS-OMT should be created and parents could complete a measure for their children. Correlations between parent ratings and teacher ratings can then be compared and performance will be evaluated in two contexts (i.e., school and home). Adding parental input could not only create a more thorough evaluation procedure; it can also lead to greater understanding of student performance on the part of teachers and parents.

It is also recommended that future studies analyzing the PGS-OMT employ a MTMM approach. Using established measures of adaptive behavior and academic achievement could help further explore whether these traits align for the student population at PG Chambers and may help determine whether the two-factor theory is more appropriate. Zahra (2015) compared the PGS-OMT to the Vineland-II and it would be helpful to include other established measures of adaptive behavior (e.g., Adaptive Behavior Assessment System (ABAS) and academic achievement (e.g. New Jersey Alternate Proficiency Assessment (APA), Academic Competence Evaluation Scales (ACES)) in future studies.

Since the Vineland-II is an established measure that is similar to the PGS-OMT, it is also recommended that a future study compare these measures and be responsive to Zahra's (2015) work in this area. Nearly all the domains on the PGS-OMT seem to align with the domains on

the Vineland-II. The Communication domain of the PGS-OMT and the Communication domain of the Vineland-II align. The Functional and Personal Care domains align with the Daily Living Skills domain of the Vineland-II. The Social domain of the PGS-OMT aligns with the Socialization domain of the Vineland-II. The Physical Navigation domain of the PGS-OMT aligns with the Motor Skills domain of The Vineland-II. The Vineland-II also has subdomains, which could help categorize the items within domains on the PGS-OMT and enhance the content covered on the tool. For example, the Communication domain of the Vineland-II has the following subdomains: Receptive, Expressive, and Written Language. Comparing results of this similar established measure to the PGS-OMT could help create more informed decisions regarding revision of the PGS-OMT.

Conclusions

To better inform educationally related decisions made at the school, district, and state levels, federal regulations require the use of psychometrically sound assessments that reflect the required content areas (NCLB, 2001; IDEA, 2004; ESSA, 2015). Despite this mandate alternate assessments have not met this standard. As a result, important educationally related decisions for students with severe disabilities cannot be meaningfully made. To their credit, staff members at PG Chambers recognized this deficit and created an alternate assessment to better evaluate their students' progress and hopefully show that their program is successful in helping students achieve academic and functional milestones.

Evidence from the current study indicates that the PGS-OMT possesses adequate reliability. Unfortunately, the evidence regarding the validity of the inferences drawn from the scores is inconclusive. Nevertheless, the PGS-OMT is new and the revision process is continuing. The current study provides encouragement that the PGS-OMT can eventually be

used as an alternate assessment of academic achievement and functional skills for the students enrolled at PG Chambers.

References

Achenbach, T. M., McConaughy, S. H., & Howell, C. T. (1987). Child/adolescent behavioral and emotional problems: implications of cross-informant correlations for situational specificity. *Psychological bulletin*, *101*(2), 213.

Achenbach, T.M., & Rescorla, L.A. (2007). Multicultural understanding of child and adolescent psychopathology. New York: Guilford Press.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (2014). Standards for educational and psychological testing. Washington, DC: Author.

Barrett, P. (2001). Assessing the reliability of rating data. Retrieved from http://www.pbarrett.net/presentations/rater.pdf

Browder, D., Ahlgrim-Delzell, L., Flowers, C., Karvonen, M., Spooner, F., & Algozzine, R. (2005). How states implement alternate assessments for students with disabilities: recommendations for national policy. *Journal Of Disability Policy Studies*, *15*(4), 209 220 12p.

Browder, D. M., Fallin, K., Davis, S., & Karvonen, M. (2003). Consideration of what may influence student outcomes on alternate assessment. *Education and Training in Developmental Disabilities*, 255-270.

Browder, D., Flowers, C., Ahlgrim-Delzell, L., Karvonen, M., Spooner, F., & Algozzine, R. (2004). The alignment of alternate assessment content with academic and functional curricula. *The Journal of Special Education*, *37*(4), 211-223.

Browder, D. M., Spooner, F., Algozzine, R., Ahlgrim-Delzell, L., Flowers, C., & Karvonen, M.

(2003). What we know and need to know about alternate assessment. *Exceptional Children*, *70*(1), 45-61.

Browder, D. M., Trela, K., Courtade, G. R., Jimenez, B. A., Knight, V., & Flowers, C. (2012). Teaching mathematics and science standards to students with moderate and severe developmental disabilities. The Journal of Special Education, 46, 26-35.

Browder, D. M., Wakeman, S. Y., Flowers, C., Rickelman, R. J., Pugalee, D., & Karvonen, M. (2007). Creating access to the general curriculum with links to grade-level content for students with significant cognitive disabilities an explication of the concept. *The Journal of Special Education*, *41*(1), 2-16.

Cohen, J. (1992). Quantitative methods in psychology. *Psychological Bulletin*, *112*(1,155-159).

Comrey, A. L., & Lee, H. B. (1992). *A first course in factor analysis* (2nd ed.). Hillsdale, NJ: Erlbaum.

De Los Reyes, A., Augenstein, T. M., Wang, M., Thomas, S. A., Drabick, D. A., Burgers, D. E., & Rabinowitz, J. (2015). The validity of the multi-informant approach to assessing child and adolescent mental health.

Destefano, L., Shriner, J. G., & Lloyd, C. A. (2001). Teacher decision making in participation of students with disabilities in large-scale assessment. *Exceptional Children*, *68*(1), 7-22.

DiPerna, J. C., & Elliott, S. N. (2000). *Academic competence evaluation scales.* San Antonio, TX: Psychological Corporation.

Elliott, S. N., Compton, E., & Roach, A. T. (2007). Building Validity Evidence for Scores on a State-Wide Alternate Assessment: A Contrasting Groups, Multimethod Approach. *Educational Measurement: Issues And Practice*, *26*(2), 30-43.

Erickson, R. N., Thurlow, M. L., Thor, K., & Seyfarth, A. (1996). *State special education outcomes, 1995*. Min- neapolis: University of Minnesota, National Center on Educational Outcomes. (ERIC Document Reproduction Service No. ED 385 061)

Every Student Succeeds Act of 2015, PL 114-195, Stat. 1177, 20 U.S.C. §§6301 *et seq.*

Field, A., (2005). *Discovering Statistics Using SPSS*. 2nd ed. London: Sage.

Goldstein, J., & Behuniak, P. (2011). Assumptions in alternate assessment: An argument- Based approach to validation. *Assessment For Effective Intervention*, *36*(3), 179-191.

Hair, J., Black., W., Babin, B., Anderson, R. (2010). *Multivariate Data Analysis.* Upper Saddle River, NJ: Prentice Hall.

Hanzlicek, V. A. (2006) The Kansas Alternate Assessment and instructional planning for special education teachers: A case study of implications for students with severe disabilities (Doctoral dissertation), AAT 3223355, DAI-A 67/06, Dec 2006. Retrieved from ProQuest Dissertations and Theses database.

Hopkins, W.G. (2001). A scale of magnitudes for effect statistics. A New View of Statistics. Retrieved from http://sportsci.org/resource/stats/effectmag.html

Idaho Department of Education (1999). *Idaho Alternate Assessment*. Boise, identification: Author.

Individuals with Disabilities Education Act, 20 U.S.C. §1400 (1997)

Individuals with Disabilities Education Improvement Act, 20 U.S.C. § 1400 (2004)

Jennings, J. L., & Beveridge, A. A. (2009). How Does Test Exemption Affect Schools' and Students' Academic Performance? *Educational Evaluation and Policy Analysis*, (2). 153.

Johnson, E., & Arnold, N. (2004). Validating an alternate assessment. Remedial and Special Education, 25 (5), 266– 275.

Karvonen, M., Flowers, C., Browder, D. M., Wakeman, S. Y., & Algozzine, B. (2006). Case study of the influences on alternate assessment outcomes for students with disabilities. *Education and Training in Developmental Disabilities*, 95-110.

Kearns, J., Towles-Reeves, E., Kleinert, H., Kleinert, J., & Thomas, M. (2011). Characteristics of and implications for students participating in alternate assessments based on alternate academic achievement standards. *Journal of Special Education*, *45*(1), 3-14.

Kettler, R. J. (2012). Testing accommodations: Theory and research to inform practice. *International Journal of Disability, Development and Education*, *59*(1), 53-66.

Kettler, R. J., Dickenson, T. S., Bennett, H. L., Morgan, G. B., Gilmore, J. A., Beddow, P. A., ... & Palmer, P. W. (2012). Enhancing the accessibility of high school science tests: A multistate experiment. *Exceptional Children*, *79*(1), 91-106.

Kleinert, H. L., Kearns, J. F., & Kennedy, S. (1997). Accountability for all students: Kentucky's alternate portfolio assessment for students with moderate and severe cognitive disabilities. *The Journal of The Association for Persons with Severe Handicaps*, 22, 88-101.

Kleinert, H. L., Kennedy, S., & Kearns, J. F. (1999). The Impact of Alternate Assessments: A Statewide Teacher Survey. *Journal Of Special Education*, *33*(2), 93-102.

Koretz, D., & Barton, K. (2004). Assessing students with disabilities: Issues and evidence. *Educational Assessment*, *9*(1-2), 29-60.

Laitusis, C. C., Maneckshana, B., Monfils, L., & Ahlgrim-Delzell, L. (2014). Differential item functioning comparisons on a performance-based alternate assessment for students with severe cognitive impairments, autism and orthopedic impairments. *Association of Test Publishers*, *10*(2), 1-33.

Marion, S. F., & Pellegrino, J. W. (2006). A validity framework for evaluating the technical quality of alternate assessments. *Educational Measurement: Issues and Practice*, *25*(4), 47-57.

Meyer, G. J., Finn, S. E., Eyde, L. D., Kay, G. G., Moreland, K. L., Dies, R. R., ... & Reed, G. M. (2001). Psychological testing and psychological assessment: A review of evidence and issues. *American psychologist*, *56*(2), 128.

Murphy, K. R., & Davidshofer, C. O. (2005). *Psychological testing: Principles and applications* (6th ed.). Upper Saddle River, NJ, Pearson Education.

National Governors Association Center for Best Practices (NGA Center), & Council of Chief State School Officers (CCSSO). (2010). *Common Core State Standards*. Retrieved from: www.corestandards.org

No Child Left Behind Act of 2001, PL 107-110, 115 Stat. 1425, 20 U.S.C. §§6301 *et seq.*

Phye, G. D., Saklofske, D. H., Andrews, J. J., & Janzen, H. L. (2001).*Handbook of Psychoeducational Assessment: A Practical Handbook A Volume in the EDUCATIONAL PSYCHOLOGY Series*. Gulf Professional Publishing.

Roach, A., Elliott, S., & Berndt, S. (2007). Teacher perceptions and the consequential validity of an alternate assessment for students with significant disabilities, *18,* 168-175.

Roeber, E. (2002). *Setting standards on alternate assessments* (Synthesis Report 42). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes. Retrieved [today's date], from the World Wide Web: http://education.umn.edu/NCEO/OnlinePubs/Synthesis42.html

Salvia, J., & Munson, S. (1986). Attitudes of regular education teachers toward mainstreaming mildly handicapped students. *Mainstreaming handicapped children: outcomes, controversies and new directions*, 111-128.

Saunders, A. F., Bethune, K. S., Spooner, F., & Browder, D. (2013). Solving the Common Core Equation: Teaching Mathematics CCSS to Students with Moderate and Severe Disabilities. *TEACHING Exceptional Children*, *45*(3), 24-33.

Sointu, E. T., Savolainen, H., Lappalainen, K., & Epstein, M. H. (2012). Cross informant agreement of behavioral and emotional strengths between Finnish students and teachers. *Scandinavian Journal of Educational Research*, 56(6), 625-636.

Sparrow, S. S., Balla, D. A., & Cicchetti, D. V. (2006). *Vineland adaptive behavior scales: Teacher rating form manual, second edition*. Minneapolis, MN: Pearson.

Thompson, S. J., Quenemoen, R. F., Thurlow, M. L., & Ysseldyke, J. E. (2001). *Alternate assessments for students with disabilities.* Thousand Oaks, CA: Corwin Press.

Thurlow, M. L., Lazarus, S. S., Thompson, S. J., & Morse, A. B. (2005). State policies on assessment participation and accommodations for students with disabilities. *The Journal of Special Education*, *38*(4), 232-240.

Towles-Reeves, E., Kleinert, H., & Muhomba, M. (2009). Alternate assessment: Have we learned anything new?. *Exceptional Children*, *75*(2), 233-252.

U.S. Department of Education. (1999) *For Evaluating Evidence of Final Assessments Under Title I of the Elementary and Secondary Education Act*. Washington, DC: Author.

U.S. Department of Education. (2003). *Education week* analysis of data from the Office of Special Education Programs, Data Analysis System. Washington, DC: Author.

U.S. Department of Education. (2005). *Alternate achievement standards for students with the most significant cognitive disabilities: Non regulatory guidance.* Washington, DG: Office of Elementary and Secondary Education.

U.S. Department of Education. (2005a). *Alternate achievement standards for students with the most significant cognitive disabilities: Non regulatory guidance.* Washington, DG: Office of Elementary and Secondary Education.

U.S. Department of Education (2006). *Assistance to States for the Education of Children With Disabilities and Preschool Grants for Children With Disabilities; Final Rule*. Federal Register, Volume 71, Number 156.

U.S. Department of Education (2007). *Modified Academic Achievement Standards. Non Regulatory Guidance.* Washington, DG: Office of Elementary and Secondary Education.

Ysseldyke, J. E., & Olsen, K. (1997). Putting alternate assessments into practice: What to measure and possible sources of data. (Synthesis Report 28). Minneapolis: University of Minnesota, National Center on Educational Outcomes.

Zahra, L. (2015). *An Alternate Assessment for Students with Multiple Disabilities: A Psychometric Evaluation and Measurement Validity Study* (Unpublished doctoral dissertation). Rutgers, the State University of New Jersey.

Appendix A
Student Demographic Questionnaire

Student Demographics for Outcomes Study

Thank you for your participation in the Outcomes Measurement Study. Please complete the following survey for each student to the best of your ability based on your experience with the student.

1. **D.O.B.**: _____ **Age**: _____ years _____ months
2. **Diagnosis**: _____
3. **Date of enrollment**: _____
4. **Gender**: Male ___ Female ___
5. **Grade**: _____

   Grade Level: Early Childhood ____ Elementary _____ Middle School _____
6. **Ethnicity**:

   White/Non-Hispanic ___ Black/Non-Hispanic ___ Hispanic ___

   American Indian/Alaskan Native ___ Asian ___ Pacific Islander ___
7. **Has this student had extended absences? Length?** _____

   _____
8. **Vision** (check one):

   Within Normal Limits ___ Wears glasses ____

   Impaired-Unaided ___ Cortical Visual Impairment ___

   Eligible for services through Commission for the Blind ___
9. **Hearing** (check one):

   Within Normal Limits_____ Impaired-aided _____ Impaired-Unaided_____
10. **Communication**:

    Verbal ____ Non-verbal _____ Augmentative Device _____

    Student's Primary Mode of Communication _____
11. **Mobility**: Ambulatory _____ Non-Ambulatory _____

    Primary Mode of Mobility _____

Appendix B
PGS-OMT Items to Revise

Domain/Factor Domain	Cross Loadings*
Social	item 6, 9
Communication	Item 6
Physical Navigation	
Personal Care	item 6, 9
Functional	item 4
Academic	

* For the unconstrained exploratory factor analysis, all cross loadings greater than 0.3 on the first three factors. These items should be considered for revision.