2018

Quality Indicators for Classrooms Serving Students with Autism Spectrum Disorder: A Construct Validation Study

Rebecca Hopkins
University of Central Florida

QUALITY INDICATORS FOR CLASSROOMS SERVING STUDENTS WITH AUTISM SPECTRUM DISORDER: A CONSTRUCT VALIDATION STUDY

by

REBECCA HOPKINS
BA Rutgers University, 1993
MA Fairfield University, 2006

A dissertation submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy
in the Department of Child, Family, and Community Sciences
in the College of Education and Human Performance
at the University of Central Florida
Orlando, Florida

Summer Term
2018

Major Professor: Eleazar Vasquez

ABSTRACT

A rise in the prevalence of students with ASD points to the need for more qualified and effective teachers to meet the needs of this population. Existing research delineates evidence-based practices and teaching standards positively improve educational outcomes for students with ASD. Teacher evaluation systems have the potential to highlight strengths and areas for improvement in special education teaching practices. Research on observation instruments to evaluate the unique skills and knowledge of special education teachers of students with ASD is limited. A need exists for high quality observation instruments to measure teacher performance in special education classrooms serving students with ASD. The purpose of this study was to examine the internal consistency reliability and the construct validity of the Quality Indicators for Classrooms Serving Students with Autism Spectrum Disorder (QIASD) scores. The researcher used a confirmatory factor analysis framework to determine if the QIASD quality indicators load onto the seven factors as hypothesized in the measurement model. The researcher found promising results but was not able to identify an acceptable model with this sample.

My dissertation is dedicated to my loving husband, Scott, who believed in me from the beginning and has encouraged me with unconditional love, and to my brilliant sons, Thomas and Daniel, who lighted my way with their joy and smiles. Thank you for all the support, I love you.

ACKNOWLEDGMENTS

To my dissertation chair, Dr. Trey Vasquez. You have been a constant source of calm direction throughout my doctoral journey. Your ability to encourage and challenge me has helped me to grow as a scholar and to become a better researcher. Thank you for everything.

To my committee, Dr. Dieker, Dr. Marino, and Dr. Rosenburg. Thank you for your guidance and mentorship throughout this dissertation process. You have provided incredible support, honest feedback, and constructive questioning to challenge me to do my best. I am fortunate to have such a compassionate, knowledgeable, and insightful team serving on my committee.

To Dr. Cynthia Pearl, thank you for being an amazing mentor. You have been an incredible source of intellectual and emotional support throughout the past three years. Your passion and dedication for what you do will always be an inspiration to me.

To the UCF faculty and colleagues I have had the pleasure of learning from and working with during my time here. Thank you for sharing your knowledge and compassion, and for welcoming me to participate in so many wonderful research and service opportunities. I have learned so much from you all. Your voices and hearts shape my future role as a faculty member.

To my fellow doctoral students, especially my cohort, Celeste, Angelica, and Faith. I am fortunate to have taken this doctoral journey with such an amazing group of scholars. You have inspired me with your passionate dedication for what you believe in and the work that you do. Thank you for being there through the tough times and for sharing the fun times too.

To the students, educators, and families I have been fortunate to know and work with. Your strength, passion, and dedication inspire me every day. Thank you.

To my family and friends near and far, I acknowledge your endless kindness, support, and understanding. Your caring words and faithful encouragement gave me the strength and motivation to stay on my path. To my parents, thank you for your unwavering love and for always believing in me. I could never have done this without you. To my darling boys, Thomas and Daniel, who are a constant source of joy and inspiration in my life.

And finally, to my forever husband, Scott, thank you for taking this journey with me. You encouraged me to take that first step on this doctoral path and have stayed next to me the whole way. You always knew when I needed a hug, some words of wisdom, or a piece of chocolate. Your strength and encouragement have lead me to where I am today. Thank you for loving me and believing in me when I was sure I could not do it, and then I did. You are my other half, and I could not have done this without you. I love you with all my heart.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER ONE: INTRODUCTION

Background and Need for Study

Over the past two decades, federal legislation in the United States has increased focus on school accountability and teaching effectiveness to improve quality and equitable learning opportunities for all students. The No Child Left Behind Act (NCLB, 2002) mandated teachers be highly qualified and held schools accountable for the academic achievement of all students. Under NCLB, minority students, including students with disabilities, were required to receive equal access to experienced, highly qualified teachers (Goe, Bell, & Little, 2008). The reauthorization of the Individuals with Disabilities Education Act (IDEA, 2004) supported the NCLB highly-qualified teacher requirements by ensuring special education teachers are fully certified and competent in the subject areas they teach. The U.S. Department of Education (2009) defined an effective teacher as a "teacher whose students achieve acceptable rates (e.g., at least one grade level in an academic year) of student growth" (p. 12).

In response, states devised evaluation systems to measure teaching effectiveness with the required emphasis on student growth, and many adopted a value-added model based on students' standardized test scores of academic achievement (Economic Policy Institute, 2010; Johnson and Semmelroth, 2014). Through these evaluation methods, the positive relationship between teacher effectiveness and student achievement was demonstrated (Aaronson, Barrow, & Sander, 2007; Cantrell, 2013; Goldhaber, 2010; Kane & Staiger, 2012; Leigh, 2010). Problems have since arisen with using value-added and standardized test scores for evaluating special education teacher effectiveness (Buzick & Jones, 2015; Kearns, Kleinert, Thurlow, Gong, & Quenemoen,

2015; Johnson & Semmelroth, 2014; Jones & Brownell, 2014; McCaffrey & Buzick, 2014; Woolf, 2015). The complex responsiblities, specialized pedagogy, and diverse needs of students with disabilities create challenges for appropriately evaluating special educators.

With the recent passage of the Every Student Succeeds Act (ESSA, 2015), the focus has shifted from federally mandated teacher qualifications to state definitions of quality and effective teachers. While the emphasis on accountability and student achievement remains, states have increased flexibility to develop unique teacher evaluation systems (ESSA, 2015). The challenge now is for states to determine accurate ways of identifying and measuring ineffective and effective teaching in both general and special education settings (ESSA, 2015). Researchers agree teacher effectiveness is influenced by multiple factors, including content knowledge, pedagogical skills, student characteristics, family support, school climate, and classroom learning environments (Cantrell, 2013; Connor, 2013; Little et al., 2009; Marshall, Smart, & Alston, 2016). Yet, special education teachers differ from general education teachers in many ways, such as in specialized knowledge and skills, diversity of roles, student populations, and demands on their time (Boe, Cook, & Sunderland, 2008; Semmelroth & Johnson, 2014; Stempien & Loeb, 2002). The distinctive position of special education teachers should be considered in the development of teacher effectiveness evaluations aimed at improving outcomes for students with disabilities.

One group of special educators with such specialized roles are teachers of students with autism spectrum disorder (ASD). The most recent Diagnostic and Statistical Manual of Mental Disorders (DSM-5; American Psychiatric Association, 2013) defined autism spectrum disorder in terms of both (a) persistent deficits in social communication and social interactions, and, (b)

patterns of restricted or repetitive behavior, interests, or activities. Specialized instructional strategies are necessary for students with ASD to learn individualized skills and to meaningfully access educational curriculum (Koegel, Koegel, Ashbaugh, & Bradshaw, 2014; Spencer et al., 2014; Spooner et al., 2012). The United States Department of Education (2016) reported the percentage of students with ASD receiving special education services increased from 0.3% in 2005 to 0.8% in 2014. Given the significant increase in the prevalence of students with ASD, states are in need of more special education teachers with the knowledge and skills to effectively teach this population (U. S. Department of Education, 2016).

Evidence-based practices, quality indicators, and effective teaching standards are abundant in the literature on educating students with ASD (Cook et al., 2014; Council for Exceptional Children, 2015; National Research Council, 2001; Wong et al., 2015). Teacher use of evidence-based practices can positively affect student outcomes (Brownell et al., 2009; Mesibov & Shea, 2011). But, teachers often lack the knowledge, training, and skills to implement evidence-based practices for students with ASD (Brock, Huber, Carter, Juarez, & Warren, 2014; Garland, Vince Garland, & Vasquez, 2013; Loiacono & Allen, 2008). The goal of teaching evaluation systems is to improve teacher knowledge and skills to positively impact student outcomes (Darling-Hammond, 2014; Woolf, 2015). Widely used teacher evaluation instruments are based on general education teaching and have not been validated to rate the specific quality and effectiveness of special education teachers (Brownell & Jones, 2015; Crowe, Rivers, & Bertilli, 2017; Johnson & Semmelroth, 2014; Jones, Buzick, & Turkan, 2013). A need exists for high quality observation instruments validated to measure the quality of special education teachers of students with ASD (Pearl et al., 2017).

Teacher Evaluation

Recent educational policy changes and a continued emphasis on standards-based accountability led to teacher evaluation reforms across the United States. In response, researchers and practitioners developed numerous instruments for measuring effective teaching and quality classroom practices (Goe & Croft, 2009; Marshall, Smart, & Alston, 2016; Semmelroth & Johnson, 2014). The U.S. Department of Education, Office of Innovation and Improvement (DOE, OII, 2017) initiated the Teacher and School Leader Incentive Program, which "promotes comprehensive evaluation and support systems for all educators" and provides funding for "performance-based compensation" (para. 2). The ultimate goal of this funding is to improve outcomes for all students, including those with special needs. Many states are re-purposing current evaluation systems or redesigning new methods to effectively encompass the diversity of schools, teachers, and students in the current educational system (DOE, OII, 2017; ESSA, 2015). Most teacher performance measures are developed for general education teachers and academic content areas (Casabianca et al., 2013; Ho & Kane, 2013; Kane & Staiger, 2012). Research on appropriate evaluation methods for special education teachers and classrooms is in the early stages of development (Johnson & Semmelroth, 2014; Kraft & Gilmour, 2017; Woolf, 2015).

Special Education Teacher Evaluation

Special educators must be highly skilled and provide individualized instruction to meet the needs of students with exceptionalities (Johnson & Semmelroth, 2014; Yell, Drasgow, & Lowrey, 2005). Teaching context, classroom composition, and student characteristics influence

the quality of teaching instruction (Cohen & Goldhaber, 2016; Jones & Brownell, 2014; Steinberg & Garrett, 2016). The challenges associated with evaluating special education teachers include varied instructional responsibilities, heterogeneous student populations, specialized knowledge, and a range of teaching conditions and environments (Buzick & Jones, 2015; Goe, Bell, & Little, 2008; Jones & Brownell, 2014; Woolf, 2015). For example, special education teachers of students with significant disabilities may provide focused, individualized instruction on multiple subjects in resource rooms or self-contained classrooms (Johnson & Semmelroth, 2014; Jones & Brownell, 2014). Using one universal quality measure for special education teachers does not account for the specialized teacher skill sets and the classroom differences to meet the heterogeneous needs of students with disabilities (Economic Policy Institute, 2010; Crowe, Rivers, & Bertoli, 2017; Johnson & Semmelroth, 2014).

Effective instructional methods to support the learning strengths and needs of students with ASD are crucial for successful education and life outcomes (Koegel, Koegel, Ashbaugh, & Bradshaw, 2014). Special education teaching necessitates specialized and individualized instructional strategies for students with ASD to access the educational curriculum (Spencer, Evmenova, Boon, & Hayes-Harris, 2014; Spooner, Knight, Browder, & Smith, 2012). Teachers have a wide range of quality indicators and evidence-based practices to choose from as a means to develop effective teaching environments for students with ASD (NCR, 2001; Wong et al., 2015). The problem is most teaching evaluation tools do not take into account these specialized teacher roles and environmental contexts of special education classrooms serving students with ASD (Crowe et al., 2017; Goe, Bell, & Little, 2008; Pearl et al., 2017).

Valid and reliable measures of evidence-based teaching for students with ASD are necessary to inform and improve educational practices (Brock et al., 2014; Coggshall, Bivona, & Reschly, 2012; Iovannone et al., 2003; Johnson & Semmelroth, 2014; Jones & Brownell, 2014; Woolf, 2015). Pearl et al. (2017) developed the Quality Indicators for Classrooms Serving Students with Autism Spectrum Disorders (QIASD) instrument for measuring the presence of quality teaching indicators and evidence-based practices in special education classrooms serving students with autism spectrum disorder. The QIASD has been validated for content (Pearl et al., 2017) in alignment with CEC educator preparation standards (CEC, 2015) and was designed specifically to provide discrete and actionable feedback to special education teachers working in classrooms serving students with ASD (Pearl et al., 2017). This dissertation explores the internal structure validity of the QIASD scores as a measure of quality teaching practices in special education classrooms serving students with ASD.

Statement of the Problem

A rise in the prevalence of students with ASD points to the need for more qualified and effective teachers to meet the needs of this population. Research exists to delineate evidence-based practices and teaching standards to positively improve educational outcomes for students with ASD. There also is research on developing, validating, and implementing measures of effective teaching to evaluate and improve educational practices. The problem is a gap in the research on psychometrically sound observation instruments to measure teaching effectiveness in special education classrooms serving students with ASD.

Rationale

Special education teachers of students with ASD face the challenge of implementing specialized knowledge and skills necessary to improve learning outcomes for this population. Teacher evaluation systems should provide meaningful feedback to special educators based on their unique roles and the population of students they serve. The interpretation of teaching evaluation scores should be validated for the context in which it will be used. Research on teaching evaluation instruments designed for special education teachers serving students with ASD is limited. In this study, the researcher addresses this critical area of need by exploring the construct validity of the Quality Indicators for Classrooms Serving Students with Autism Spectrum Disorder (QIASD) measure.

Overview of Methodology

The researcher used a confirmatory factor analysis (CFA) framework to investigate the internal consistency reliability and the construct validity of the QIASD scores with a sample of K-12 special education classrooms observations serving students with ASD.

Research Questions

RQ 1a: To what extent does the QIASD produce reliable scores as measured by internal consistency reliability?

RQ 1b: To what extent is construct validity of the QIASD achieved as measured by a confirmatory factor analysis?

Organization of the Dissertation

This dissertation is divided into five chapters. Chapter One introduces the background and need for the study including some foundational information on the construct of effective teaching, educating students with autism spectrum disorder, special education teacher evaluation, and the importance of validity assessment. Chapter Two provides a systematic literature review of the existing empirical research on the development and psychometric validation of observation instruments designed for measuring teaching effectiveness in special education classrooms serving students with ASD. Chapter Three provides a detailed description of the methodology for this study including the research questions, research design, sample description, study procedures, and data analysis techniques. Chapter Four presents the results of the data analysis as well as methodological adjustments evolving from an iterative data analysis process. Chapter Five offers a discussion of the findings and methods including implications of the analysis, limitations, and future research recommendations.

CHAPTER TWO: LITERATURE REVIEW

Chapter Two displays the results of a systematic literature review on the intersection of K-12 special education teaching evaluation instruments, autism spectrum disorder, and evidence of psychometric properties. This chapter includes an overview of students with ASD and the importance of quality measures of effective teaching for this population. The researcher provides a detailed summary of the empirical literature on teaching evaluation instruments for special education classrooms serving students with ASD and the evidence of psychometric qualities of those instruments.

Introduction

Autism Prevalence

Approximately one in 59 children in the United States has ASD (Baio et al., 2018). This category was added as a separate special education disability category in the 1990 amendments to the Individuals with Disabilities Education Act (IDEA; 1990). According to the Office of Special Education and Rehabilitative Services (OSERS) 38th report to Congress, over 450,000 students in the U.S. ages 6 to 21 have autism and receive special education services (DOE, OSERS, OSEP, 2016). The IDEIA (2004) mandated students with disabilities, including ASD, be educated within the least restrictive environment (LRE) and to the same high academic standards as their typically developing peers. However, students with more severe symptoms of ASD are often educated solely by special education teachers within self-contained classrooms or

resource room settings (Hart & Whalon, 2011; White, Keonig, & Scahill, 2007). Over 32%, or approximately 140,000 students with ASD, are educated in public school settings outside of the general education classroom for 60% or more of the school day (DOE, OSERS, OSEP, 2016). The rise of students with ASD over the past two decades has created a growing demand for special education teachers to serve this population.

Special Education and Autism

Researchers responded to the need for special educators with expertise in teaching students with ASD by identifying quality indicators and standards of effective teaching. Iovanonne, Dunlap, Huber, and Kincaid (2003) offered the following core components of effective teaching of students with ASD:

- individualized supports and services for students and families;
- systematic instruction;
- comprehensible and/or structured learning environments;
- specialized curriculum content;
- a functional approach to problem behavior; and,
- family involvement (p. 153).

The National Research Council (2001) proposed 12 quality indicators for educational programs serving students with autism, three of which relate directly to educators: "(a) highly trained staff, (b) comprehensive professional resources, and (c) staff supervision and program evaluation mechanisms" (Morrier, Hess, and Heflin, 2011, p. 2).

The Council for Exceptional Children (CEC; 2015) identified initial, advanced, and specialty standards delineating the critical knowledge and skills required for effective special education teaching. The CEC Practice and Preparation Standards for special educators are approved by the Council for the Accreditation of Educator Preparation (CAEP) to support quality teacher preparation programs (CEC, 2012). One CEC Specialty Set focuses on the specialized knowledge and skills of special education teachers of students with developmental disabilities and autism spectrum disorder (CEC, 2015). These specialty standards highlight quality indicators for teaching students with ASD under seven main categories: (a) learner development and individual learner differences, (b) learning environments, (c) curricular content knowledge, (d) assessment, (e) instructional planning and strategies, (f) professional learning and ethical practice, and (g) collaboration (CEC, 2015). These best practice standards provide a sound basis of professional guidelines for developing and evaluating effective special educators to supporting students with ASD (CEC, 2012).

Evidence-Based Practices

The Individuals with Disabilities Education Improvement Act (IDEIA, 2004) authorized grant funding to states and institutions of higher education to support research and initiatives focused on improving outcomes for students with disabilities (U.S. Department of Education, Institute of Education Sciences, National Center for Special Education Research, 2004). Over the past two decades, researchers have identified a plethora of evidence-based practices designed for students with disabilities (Cook et al., 2014; Rogers & Vismara, 2008; Spooner, McKissick & Knight, 2017; Walker & Gresham, 2013; Wong et al., 2015). A substantial increase in the

number of students identified with ASD (Centers for Disease Control, 2014) led to a particular emphasis on identifying evidence-based interventions specific to students with ASD (National Autism Center, 2009; Rogers & Vismara, 2008; Smith & Iadarola, 2015; Wong et al., 2015).

The U.S. Department of Education, Institute of Education Sciences formed the What Works Clearinghouse (2002) to identify evidenced-based practices for students with disabilities, including ASD, based on high-quality empirical research. Researchers at the National Autism Center (NAC; 2009) and the National Development Center on Autism Spectrum Disorders (NPDC; Odom et al., 2010; Wong et al., 2015) reviewed hundreds of studies and identified 27 evidence-based interventions for students with ASD. Students with ASD are a heterogeneous population, and they have unique abilities and challenges requiring educators to use specialized knowledge and skills for effective instruction (APA, 2013; Begoli, DeFalco, & Ogle, 2016; Iovannone et al., 2003; Johnson & Semmelroth, 2014; Sansosti & Sansosti, 2013).

Despite the wealth of evidence-based interventions and best teaching standards, teachers are inadequately prepared and lack sufficient knowledge and skills to effectively meet the educational needs of students with ASD (Brock, Huber, Carter, Juarez, & Warren, 2014; Iovannone et al., 2003; Jennett, Harris, & Mesobov, 2003; Johnson & Semmelroth, 2014; Shyman, 2012; Westling, 2010). In a survey of 234 teachers, less than 5% of teachers reported implementing best practices for students with ASD (Morrier, Hess, & Heflin, 2011). The lack of skillful teachers for students with ASD is a concern as differences in teacher effectiveness levels lead to differences in student learning and achievement outcomes (Leigh, 2010; Reddy, Fabiano, Dudek, & Hsu, 2013; Rockoff, 2004). Evaluation methods are needed to inform continuous

improvement and professional development initiatives for special education teachers serving students with ASD (Marshall, Stuart, & Alston, 2016; Reddy, Fabiano, Dudek, & Hsu, 2013).

Educational and life outcomes of individuals with ASD are inconsistent (Anderson, Liang, & Lord, 2014). Some individuals with ASD have gained greater independence and positive improvements in social, language, and cognitive skills as they progressed through school years into adolescence and adulthood (Anderson et al., 2014; Billstedt, Gilberg, & Gillberg, 2005; Eaves & Ho, 2008; Farley et al., 2009; Farley & McMahon, 2014; Fein et al., 2013; Sutera et al., 2007). Studies of longitudinal data (Fein et al., 2013; Pellicano, 2012) showed individuals with ASD who had "extremely positive outcomes had milder symptoms and received more treatment as young children" (Anderson et al., 2014, p. 485). Greater positive outcomes are predicted for students with ASD with normal IQ's compared to those with co-occurring intellectual disabilities (Anderson et al., 2014; Howlin, Goode, Hutton, & Rutter, 2004). Quality of life outcomes, such as independence, employment, and socialization for individuals with ASD are relatively poor (Bishop-Fitzpatrick et al., 2016). Individuals with more severe ASD and co-occurring intellectual disabilities face even greater challenges to obtain successful life outcomes (Anderson, Liang, & Lord, 2014). Effective special education teachers are necessary to provide a quality education for students with ASD that will positively impact their educational achievement and quality of life.

Challenges of Evaluating Special Education Teachers

Researchers agree current approaches to teacher evaluation designed for general education teachers may not be successful in identifying quality and effectiveness of special

education teachers (Brownell & Jones, 2015; Johnson, 2015; Johnson & Semmelroth, 2014; Jones, Buzick, & Turkan, 2013). One challenge is defining an effective special education teacher. Johnson and Semmelroth (2013) defined an effective special education teacher as being "able to identify a student's needs, implement evidence-based instructional practices and interventions, and demonstrate student growth" (Johnson, 2015, p. 83). Another challenge is fair and accurate measurement of the diverse roles of special educators who require specialized knowledge and skills to serve a variety of students with disabilities in a range of learning contexts (Council for Exceptional Children, 2012; Johnson & Semmelroth, 2013). These elements highlight the need for evaluation systems to account for the unique and complex roles of special educators that result in actionable feedback for positive professional development (Johnson & Semmelroth, 2014; Reddy, Fabiano, Dudek, & Hsu, 2013).

Teacher evaluation systems currently used within special education often do not accurately identify variations in teacher performance, effectiveness, or ability to improve student achievement (Cohen & Goldhaber, 2016; Kraft & Gilmour, 2016; Rockoff, 2004). Researchers indicate value-added scores are not appropriate for evaluating special education teachers due to small class sizes, context variability, and the uncertainty of standardized test scores to reflect the true abilities of students with disabilities (Buzick & Jones, 2015; CEC, 2012; Gansle et al., 2015; Goldhaber, 2015; Jones & Brownell, 2014; Steinbrecher, Selig, Cosbey, & Thorstensen, 2014). Special education teachers focus on multiple important outcomes for students with disabilities including academic, social, communication, behavioral, and adaptive goals (Jones and Brownell, 2014). Classroom observations can provide evidence of teaching practices and student learning across a variety of contexts (Crowe, Rivers, & Bertoli, 2017; Kane, McCaffrey, Miller, &

Staiger, 2013). Classroom observation tools are an important alternative to value-added methods for measuring quality teaching in special education environments (Goe et al., 2008; Jones & Brownell, 2014; Pianta & Hamre, 2009).

Effective Teacher Evaluation

A universal quality measure for special education teachers does not account for the specialized teacher skill sets and the classroom differences to meet the heterogeneous needs of students with disabilities (Economic Policy Institute, 2010; Crowe et al., 2017; Johnson & Semmelroth, 2014). The Council for Exceptional Children (CEC, 2012) released a position paper identifying five main components of effective evaluation of special education teachers. CEC recommended evaluation systems incorporate: (a) fundamental system wide components (i.e., research-based standards, fidelity of implementation, and continuous improvement); (b) the complex role of special educators (i.e., based on specific responsibilities, have clear performance standards, and take into account the population of students); (c) teacher use of evidence-based practices (i.e., to include multiple indicators of effectiveness and not based only on student growth); (d) recognize professionalism (i.e., teacher involvement in the evaluation process and provision of constructive and actionable feedback); and, (e) continually incorporate research findings (i.e., collaboration between evaluation leaders, use of valid and reliable measures, and continued research to link evaluations to improvement) (CEC, 2012). A need exists to create more consistent measures of teaching effectiveness sensitive to the distinctive expertise,

environment, and responsibilities of special educators (Darling-Hammond, 2015; Goe et al., 2008; Jones & Brownell, 2014; Woolf, 2015).

Special Education Classrooms

The Individuals with Disabilities Act (2004) provides a definition of special education as "specially designed instruction … to meet the unique needs of a child with a disability" (IDEA, 2004, §1401(29)). Students with significant disabilities, including many with autism spectrum disorder, often receive specialized instruction within separate, self-contained classrooms (DOE, OSERS, OSEP, 2016). Self-contained classrooms typically have a smaller student to teacher ratio, ranging from 6:1 to 12:1, and a minimum of one paraprofessional to assist with classroom management and instruction (Crowe et al., 2017). Self-contained special education classrooms are characterized by specialized supports and intensive, systematic instruction based on individual student needs (Jones & Brownell, 2014). Students with ASD who exhibit severe social communication impairments and behavioral challenges are more likely to be educated within self-contained classrooms (Lyons et al., 2011; White et al., 2007).

Students with ASD with less access to inclusive educational environments are still held to the same challenging academic standards as their general education peers (ESSA, 2015; Hart & Whalon, 2011; Westling, 2010). Special education classrooms serving students with ASD need skillful teachers to ensure access to core curriculum and individualized instruction. High quality education for students with ASD includes a classroom environment with specialized teaching and focused evidence-based instruction and supports (CEC, 2015; Odom & Wong, 2015). Valid

and reliable measures of effective teaching practices for classrooms serving students with ASD are important to inform professional development and continuous quality improvement efforts.

Current Evaluation Trends

Legislative efforts to hold schools and teachers accountable for student achievement fuel current educational reforms focused on standards for quality teaching, teacher preparation programs, and professional development to improve teaching (ESSA, 2015; Holdheide, 2015). States and school districts are tasked with developing valid and reliable evaluation systems to measure the quality of education for all students (USDOE, Office of Planning, Evaluation, and Policy Development [OPEPD], 2016). In recent years, teacher effectiveness has largely been measured through a combination of administrative classroom observations, self-assessment, and value-added models (Benedict, Thomas, Kimerling, & Leko, 2013; Fall, 2010; Kersting, Chen, & Stigler, 2013; Pearl et al., 2017; Tandy, Whitford, & Hirth, 2016). States have based comprehensive teacher evaluation systems on instruments designed for teachers of academic subjects and test scores of students in general education settings (Holdeheide, 2015; Danielson, 2007; Goe et al., 2008; Semmelroth & Johnson, 2014). Widely used evaluation systems do not take into account the unique skills of special education teachers and the individualized learning characteristics of students with disabilities (Danielson, 2013; Hamre, Pianta, Mashburn, & Downer, 2007; Kane & Steiger, 2012). For example, the Framework for Teaching (Danielson, 2007) model used in over 30 states was not intended for special education teachers and classrooms (Crowe et al., 2017).

In the recent Study of Emerging Teacher Evaluation Systems (USDOE, OPEPD, 2016), eight district evaluation systems were examined. All eight systems contained a classroom observation and a student growth component. Four districts in the study incorporated Danielson's Framework for Teaching, while the other four districts created evaluation systems based on multiple existing frameworks. For example, the District of Columbia consulted 20 other frameworks to develop IMPACT, which delineates nine domains of high quality teaching practices (USDOE, OPEPD, 2016). None of the districts in this review customized classroom observation tools to reflect different grade-levels or subjects. The District of Columbia used standards based on applied behavior analysis methodology for evaluating special education teachers of students with autism (DCPS, 2012). Two districts (District of Columbia and Hillsborough County) sought outside validation of their teacher evaluation systems to examine correlations between observation measures of teacher practice and student growth measures (USDOE, OPEPD, 2016). All districts in this report provided some form of informational training packet on their evaluation system and five districts developed online trainings for observers. Only two of the eight districts gathered interrater reliability data on their observation measures (USDOE, OPEPD, 2016).

The National Council on Teacher Quality (Doherty & Jacobs, 2015; NCTQ) published a report on how states and districts are evaluating teacher effectiveness in K-12 classrooms. According to the NCTQ, 43 states required measures of student growth as part of their teacher evaluation systems, and 28 states determined ineffective teachers may be eligible for dismissal. Differentiating between effective and ineffective teachers is an ongoing challenge in teacher observation systems. Most teachers across the nation are being rated by administrators as

effective or highly effective (Doherty & Jacobs, 2015), which leaves little room for teacher improvement efforts or professional development. This trend of overly positive teacher evaluations contradicts policy supporting the need for accountability to improve teacher effectiveness and equality in student outcomes (ESSA, 2015). Observation instruments with valid and reliable ratings are needed to effectively distinguish levels of quality teaching occurring in special education classrooms.

Importance of Validity

A need exists for high quality tools to measure the performance of special education teachers of students with ASD. These tools would provide valuable insights into the teaching strengths and areas of need of students with ASD in special education classrooms. A high-quality teaching evaluation instrument should meet standards to validly interpret scores as intended. Validity is defined in the Standards for Educational and Psychological Testing (AERA, APA, & NCME, 1999) as "the degree to which accumulated evidence and theory support specific interpretations of test scores entailed by proposed uses of a test" (p. 184). Evidence of validity is considered the most crucial element in developing and evaluating the quality of an educational assessment (Camara, 2003; Millett, Stickler, Payne, & Dwyer, 2007).

Many states and districts have not evaluated the quality of their teacher evaluation systems, particularly in the area of special education (Semmelroth & Johnson, 2014). Very little empirical research has been published on the validation of classroom observation measures intended for use within special education classrooms (Crowe et al., 2017). Determining the validity and reliability of classroom observation ratings to measure the critical indicators of

quality special education teaching is necessary to provide relevant feedback and design professional development geared to improving teaching (Camara, 2003; Conner, 2013; Grossman & McDonald, 2008; Holdeheide, 2015).

Classroom Observations

Classroom observations are considered a crucial component to any teaching evaluation system (Benedict et al., 2013; Brownell & Jones, 2015; Holdeheide, 2015). Teachers report feedback from classroom observations helps improve their teaching (USDOE, OPEPD, 2016). Classroom observations can be used to measure a variety of teacher practices and environmental contexts that may impact quality teaching. Effective observation systems to improve teaching practices should focus on specific subject or content areas, such as special education, and stimulate productive, actionable feedback (Johnson, Crawford, Moylan, & Ford, 2016). While considerable research has been conducted on classroom observation tools, the majority focus on general education teachers, students in general education classrooms, and academic subject areas (Kane & Staiger, 2012; Semmelroth & Johnson, 2014). The U. S. Department of Education Office of Planning, Evaluation and Policy Development (2016) reported teachers and administrators across eight states had district-wide classroom observation rubrics that were inappropriate for special education teachers. Quality teaching in special education classrooms must incorporate the unique learning characteristics of students with disabilities, variations in instructional content and delivery, paraprofessional involvement, and individualized services and instruction based on evidence-based practices (Crowe et al., 2017; Odom, et al., 2005; Johnson et

al., 2016). These specialized components must be considered in classroom observation systems if they are to yield valuable feedback for improving special education teacher practices.

The Measures of Effective Teaching (MET) project consisted of a large-scale investigation on classroom observations to identify valid and reliable methods for informing teacher feedback and professional development (Kane & Staiger, 2012). The MET researchers evaluated five classroom observation instruments: Framework for Teaching (Danielson, 2012; FFT), Classroom Assessment Scoring System (CLASS; Pianta, La Paro, & Hamre, 2003), Protocol for Language Arts Teaching Observations (PLATO; Grossman, 2009), Mathematical Quality of Instruction (MQI; Hill et al., 2008), and UTeach Teacher Observation Protocol (UTOP; Marder & Walkington, 2012). Data were collected by 90 trained raters who scored 7,491 videos of math and English language arts lessons from 1,333 teachers across six states. The results indicated reliable scores based on multiple raters and observations, and all five instruments positively related to student achievement gains. All of the observations in this study focused on academic subjects within general education classrooms. The validity of the observation instrument ratings may not generalize to other subject areas, such as special education, where the "model of effective instruction could be very different" (Kane & Staiger, 2012, p. 58).

Two recent studies examined the extent of special education teacher evaluation measures currently available. Holdheide (2015) reviewed five emerging practices being developed by states and districts to evaluate special education teaching. For example, the Massachusetts Department of Elementary and Secondary Education (2015) designed a teacher observation rubric focused on inclusive educational practices for use with general and special education

teachers. Colorado created the Practical Ideas for Evaluating Special Educators (Colorado Department of Education, 2015), a set of flexible examples to guide districts in adapting the general observation rubric to address the unique roles of special educators. The Pennsylvania Teacher Effectiveness Evaluation System, based on the Framework for Teaching (Danielson, 2013), included supplemental guidelines to support the general observation rubric within the context of a specific teacher's role (Pennsylvania Department of Education, 2014). Holdheide described an example of how follow-up questions and examples may be used to supplement the evaluation of teachers serving students with autism. Holdheide's review highlighted how many states are tacking on supplements and guidelines to existing teacher evaluation frameworks, resulting in observation rubrics without empirical evidence to support their effectiveness (Holdeheide, 2015).

Crowe, Rivers, and Bertoli (2017) reviewed literature and federal websites for existing classroom observation protocols in published research or being used in school districts. Crowe et al. (2017) identified 104 observation tools used to evaluate general or special education environments. Thirteen tools were designed for special education classrooms. Of those, only two were found in published research: the Classroom Climate Scale (McIntosh, Vaughn, Schumm, Haager, & Lee, 1994) and the Classroom Observation Scale (Stanovish & Jordan, 1998).

Two main concerns evident from these studies (Crowe et al., 2017; Holdheide, 2015) are (a) the continued lack of teaching observation tools designed for special education classrooms, and (b) the dearth of research on the psychometric properties of special education observation instruments. States and districts have a limited range of high quality, specialized teacher evaluation tools to choose from in the literature. A need exists for researchers to develop and

validate measures of special education teaching in the unique contexts of special education classrooms serving students with disabilities.

Since the passage of ESSA (2015), states and districts are revamping or redesigning evaluation systems with the goal of improving teaching practices and student achievement. There is minimal research in the area of special education teacher evaluation to guide state policy and district practices (Brownell & Jones, 2015; Tandy et al., 2016). Observation measures designed for special education classrooms should capture the unique context, student characteristics, and specialized teacher skills that represent quality teaching and influence student growth (Semmelroth & Johnson, 2014; Tandy et al., 2016). Standards for quality special educator practice (CEC, 2012) and evidence-based practices for students with ASD (Cook & Odom, 2013; Wong et al., 2015) provide a solid foundation for developing valid and reliable special education classroom observation tools for teachers of students with ASD.

Purpose

The purpose of this systematic literature review is twofold: (a) to identify existing empirical literature on observation instruments for measuring teaching effectiveness in special education classrooms serving students with autism spectrum disorder, and (b) to determine the extent of psychometric evidence provided for these instruments.

Research Questions

1. To what extent does a systematic review of empirical literature reveal observation instruments for measuring effective teaching practices in K-12 special education classrooms serving students with autism spectrum disorder?
2. To what extent is psychometric evidence reported for the observation instruments identified in this systematic literature review?

Methods

Criteria

A systematic review of the literature was conducted to identify existing psychometrically validated observation tools used to measure the quality of special education classrooms serving students with autism in grades K-12. The researcher searched the following electronic databases: ERIC, PsychINFO, Education Source, and Google Scholar. The search criteria included empirical articles published between 2006 to 2017 in scholarly, peer reviewed journals and in English language. Three separate electronic searches were conducted in attempt to find the most relevant articles.

1. Initial search terms included special education, evaluation or classroom observation, and tool, instrument, scale, measure, rating, or indicators. This search produced 1,024 results. The search was further refined to include the terms psychometrics or validity or reliability, resulting in 175 empirical studies.

2. A second search was conducted using similar but slightly more focused search terms in attempt to find studies related specifically to assessing teaching of students with autism spectrum disorder. The second search terms were autism, teacher or classroom, effective or quality, and tool or instrument or scale or measure or rating or indicators. This search identified 276 empirical studies. The search was narrowed to 211 results by excluding the term preschool.
3. A third search was conducted in Google Scholar using the advanced search option with the terms: special education, teacher evaluation, classroom observation, quality, effective, validity AND autism AND tool, instrument, scale, measure, rating indicator. The search was narrowed to exclude the terms medical, nursing, and preschool. This search produced 40 results.

Data Extraction

Instrument and measurement characteristics were extracted from the selected articles to include: author, instrument name, setting being observed, purpose of the instrument, number of teachers/classrooms observed, number of observations, number of raters, type of rater (administrator, faculty, peer teacher), response process (i.e., observation type, number of indicators, scoring scale), psychometric evidence (validity or reliability).

Search Reliability

A second doctoral student in the field of education provided a reliability check on October 12, 2017 by independently performing the same three searches using the specified search terms and criteria. The researcher provided a written description of the three search processes, including (a) accessing the data bases, (b) inputting the search terms, (c) selecting other inclusion criteria such as the date range and empirical research, and (d) inputting exclusion terms. The researcher then provided an excel spreadsheet with a list of the articles identified for inclusion in this review. The second doctoral student's searches resulted in (a) 171 articles, (b) 211 articles, and (c) 40 results. Although there was a small discrepancy with the number of results in the first search (171 compared to 175), it did not impact the check for articles. Once the researcher completed the list of articles to be included in the review, the second doctoral student reviewed the search results to corroborate the identify of these final articles within the given search parameters.

Results

The researcher examined the abstracts for the final 426 results to identify potentially relevant articles for this review. Articles with the following criteria were included:

- empirical articles published in scholarly, peer reviewed journals between 2006-2017;
- articles reported on the development, validation, or use of an observation instrument; focused on teachers and students in special education settings; and,
- articles written in English and published in the United States.

Twenty-nine articles were selected for review. The author excluded 23 articles with instruments developed or validated only for use in general education or preschool settings. Six articles met the search criteria and were retained for this review.

Instrument Characteristics

The researcher discovered five special education observation instruments in published empirical works between 2006-2017 (see Table 1). The results support previous research findings of a lack of valid and reliable observation measures for special education classrooms (Crowe et al., 2017; Semmelroth & Johnson, 2014), including those serving students with ASD (Pearl et al., 2017). All five measures focused on special education classrooms were rooted in the use of evidence-based practices as important components of teaching quality. The APERS (Odom et al., 2013) and the QIASD (Pearl et al., 2017) were developed specifically to measure effective and evidence-based teaching practices in classrooms serving students with ASD.

Table 1

Instrument Characteristics

Observation Tool	Author	Article Name	Setting Observed	Purpose Tool	Items	Scoring	Data type
Recognizing Effective Special Education Teachers (RESET)	Johnson & Semmelroth, 2012	Examining interrater agreement analyses of a pilot special education observation tool.	Special Education - Elementary and Secondary	Evaluate teacher use of EBPs aligned with content, nature of disability, and grade level of students	28-67 items based on EBPs being observed	Four-point Likert scale	Observation video coded
	Semmelroth & Johnson, 2014	Measuring rater reliability on a special education observation tool.	Special Education Elementary and Secondary	Evaluate special education teacher effectiveness through use of EBPs	Three Subscales	Four-point Likert scale	Observation video coded
Participatory Evaluation and Expert Review for Classrooms Serving Students with EBD (PEER-EBD)	Tsai, Cheney, & Walker, 2013	Preliminary psychometrics of the participatory evaluation and expert review for classrooms serving students with emotional/behavioral disabilities (PEER-EBD)	Special Education	Evaluate programs serving students with EBD	Four Domains and 19 practices each with 3-9 indicators	Five-point Likert scale	Individual self-review, peer review, and expert consultant classroom observation

Observation Tool	Author	Article Name	Setting Observed	Purpose Tool	Items	Scoring	Data type
Behavioral Classroom Needs Assessment	Leaf, Leaf, McCray, Lamkins, Taubman, McEachin, & Cihon, 2016	A preliminary analysis of a behavioral classrooms needs assessment.	Special education classrooms, autism classrooms, resource rooms PreK – HS	Measure quality implementation of ABA principles, teaching, and classroom	Nine domains, 40 questions	Five-point Likert scale	Observation
Autism Program Environment Rating Scale (APERS)	Odom, Cox, Brock, & National Professional Development Center on ASD, 2013	Implementation science, professional development, and autism spectrum disorders.	Special education classrooms and general education inclusive classrooms PreK-HS	Assess quality in classrooms using a specific program serving students with ASD	PreK/Elem 11 domains, 64 indicators Middle/High 12 domains, 66 indicators	Five-point behaviorally anchored Likert scale	Observation, interviews, record review
Quality Indicators for Classrooms Serving Students with Autism Spectrum Disorder (QIASD)	Pearl, Vasquez, Marino, Wienke, Donehower, Gourwitz, . . . Duerr, 2017	Establishing content validity of the Quality Indicators for Classrooms Serving Students with Autism Spectrum Disorders instrument.	Special education classrooms	Evaluate presence of quality indicators in classrooms serving students with ASD	Seven Standards, 52 indicators	Rating scale 0-4 and N/A	Observation, interview, artifact review

Measurement Characteristics

Of the five instruments reviewed (*N*= 5), four researchers reported some form of psychometric evidence of validity or reliability of scores (*N*=4) (see Table 2). In two studies (Pearl et al., 2017; Tsai, Cheney, & Walker, 2013) researchers collected feedback from national experts in the field to provide evidence of content validity. Only Tsai, Cheney, and Walker (2013) assessed construct validity by analyzing the proposed factor structure of the PEER-EBD. Semmelroth and Johnson (2014) collected interrater reliability on teachers coding videos of classroom instruction with the RESET tool, but no further validity evidence was reported.

Table 2

Measurement Characteristics

Tool	Number of Teachers/ Classrooms Observed	Number of Obsvs	# of Raters	Type of Rater	Psychometric Evidence	
					Validity	Reliability
Recognizing Effective Special Education Teachers (RESET)	12 special education teachers across 3 districts over one school year	Two coding sessions of observing videos of classroom instruction	6	Special education teachers	None reported	Interrater reliability
	9 special education teachers across 5 districts over two school years	Two coding sessions of observing videos of classroom instruction	5	Special Education Teachers		Interrater reliability
Participatory Evaluation and Expert Review for Classrooms Serving Students with EBD (PEER-EBD)	23 center-based classrooms serving students with EBD	145 self-evaluations	145	Administrators, special education teachers, school psychologists, counselors, social workers, paraprofessionals, and related service providers	Content, Construct Validity	Internal Consistency
Behavioral Classroom Needs Assessment	68 teachers	128 observations	18	Consultant and first author	None reported	Interrater reliability

Tool	Number of Teachers/ Classrooms Observed	Number of Obsvs	# of Raters	Type of Rater	Psychometric Evidence	
					Validity	Reliability
Autism program environment rating scale.	58 school programs across nine states	2 Pretest – posttest observations	Not reported	Authors	None reported	None reported
Quality Indicators for Classrooms Serving Students with Autism Spectrum Disorder (QIASD)	51 quality indicators reviewed	N/A	103 national, subject experts	59 teachers, 5 school administrators, and 39 university faculty	Content Validity – Lawshe's methodology	None reported

Results of Individual Studies

The Recognizing Effective Special Education Teachers (RESET) was developed and piloted in Idaho to measure teacher use of evidence-based practices as a means to evaluate the effectiveness of special education teachers (Semmelroth & Johnson, 2014). The RESET tool has been repeatedly studied by the developers to gather evidence of interrater reliability and validity (Johnson & Semmelroth, 2012; Semmelroth & Johnson, 2014). The more recent study produced acceptable levels (.65 or above) of score reliability for Subscales 1 and 3, but not for Subscale 2 (Semmelroth & Johnson, 2014). Semmelroth and Johnson emphasized the need for multiple observations conducted by multiple raters in order to achieve higher levels of interrater reliability of measurement scores.

The Behavioral Classroom Needs Assessment (Leaf et al., 2016) was designed to measure the quality of special education instruction based on applied behavior analysis (ABA) methodology. The instrument consisted of nine domains with a total of 40 questions scored using a five-point Likert scale. The nine domains were: age appropriateness, curriculum, reinforcement, behavior plans, teaching strategies, Discrete Trial Teaching, shadow support, data, and classroom environment (Leaf et al., 2016). Observations for interrater reliability were conducted by 17 school-based consultants and the first author. The observers completed 128 classroom observations within 69 special education, autism, and resource classrooms, each lasting about 20 minutes. The results indicated overall high interrater reliability with Intraclass Correlation Coefficient ranges from 0.528 to 0.845 and Cronbach's Alpha from 0.691 to 0.916.

Odom, Cox, Brock, and the National Professional Development Center on ASD (2013) conducted a study on the implementation of the Evidence-Based Individualized Program for Students with Autism (EBIPSA), a specific model for classrooms serving students with autism spectrum disorder. The model was designed to guide states in developing quality programs for students with ASD and in the provision of professional development to improve teacher use of evidence-based practices. A component of this model was the Autism Program Environment Rating Scale (APERS). There are two versions of the APERS: one for preschool/elementary grades with 11 domains and 64 quality indicators; and one for middle/high school grades with 12 domains and 66 indicators (National Professional Development Center on ASD, 2012). Odom et al. (2013) used the APERS as a pre-post measure of the quality of programs serving students with ASD before and after implementing the EBIPSA model.

The Participatory Evaluation and Expert Review for Classrooms Serving Students with EBD (PEER-EBD; Tsai, Cheney, & Walker, 2013) was developed to evaluate the quality of school-based programs serving students with EBP. The PEER-EBD was designed as a participatory and collaborative feedback tool to gain multiple perspectives from a variety of educators using a teacher self-evaluation, peer evaluation, and expert consultant observation. Tsai et al. (2013) validated the content of the PEER-EBD through a national panel of experts in the field of EBD. The authors reported evidence of internal consistency and construct validity with a sample of 145 staff members over 23 special education K-12 classrooms specifically developed to serve students with EBD. Cronbach's alpha values for the four domains were .876, .943, .917, and .900, respectively and .965 for the overall measure, supporting good internal consistency reliability (Tsai et al., 2013). To examine construct validity, a confirmatory factor

analysis of the proposed four-factor model was revised and resulted in mediocre fit to the data (CFI = .943, SRMR = .065, RMSEA = .083) with adequate loadings of items to factors (> .30).

Limitations and Implications

The main limitation of this systematic review is the likelihood of failing to identify some instruments. The range of teacher evaluation and observation measures in the field of education is extensive. In addition, searching for specific measures of teaching quality and effectiveness was challenging due to the assorted terminology, wide variety of measurement tools, and large amount of research on interventions that is intertwined with relevant publications targeted by the researcher. With states and districts updating or redesigning evaluation systems and researchers devising innovative ways to measure effective teaching, it is also possible new observation measures were developed and studied after this review was conducted. Another limitation is the search was limited to articles in English language and published in the United States, which may have introduced publication bias and led to the exclusion of some potentially useful observation tools developed in other countries.

The researcher specified parameters for this systematic review in attempt to make the process more manageable and to focus in on the special education instruments most relevant to the proposed study. Although this systematic review was not exhaustive of all the instruments being used for evaluating special educators, the results support previous findings (Crowe et al., 2017) by highlighting the dearth of published research available on special education classroom observation instruments. While it was promising to see some special education classroom observation measures being developed and used within school district teacher evaluation systems

(Jacob et al., 2016; Semmelroth & Johnson, 2014), a clear gap exists in the research on developing high quality observation measures for special education classrooms serving students with ASD and on investigating the psychometric validity of those measures.

QIASD Development

This proposed validity research on the QIASD instrument is guided by the Standards for Educational and Psychological Testing (AERA, APA, & NCME, 1999), a widely acknowledged reference for developing and evaluating educational assessments (Camara, 2003). Gathering psychometric information on teaching effectiveness measures is a crucial step towards making appropriate inferences about test scores (AERA, APA, & NCME, 1999; Gall, Gall, & Borg, 2007). The process of validating an instrument involves determining the reliability and validity of scores. Reliability was defined in the MET study as "the proportion of the variance in instrument scores reflecting consistent differences in practice between individual teachers" (Kane & Staiger, 2012, p. 4). Test reliability indicates the "consistency, stability, and precision of test scores" (Gall, Gall, & Borg, 2007, p. 151) as influenced by the level of measurement error. According to Classical Test Theory, all observed scores are comprised of a "true" score and some sources of random error (Gall et al., 2007; Kline, 2010; Wu, Tam, & Jen, 2016). One way to estimate the reliability of test scores is by examining the internal consistency of the correlations between individual indicators on the assessment (Cook, Thomas, & Beckman, 2006; Gall et al., 2007). If observed variables are positively and highly intercorrelated under the corresponding latent variables, the instrument is considered to have internal consistency (Kline,

2010, 2014). In the proposed study, the researcher aims to establish internal consistency reliability of the QIASD scores as a necessary step towards validating the instrument's quality.

Validity is defined in the Standards for Educational and Psychological Testing (AERA, APA, & NCME, 1999) as "the degree to which accumulated evidence and theory support specific interpretations of test scores entailed by proposed uses of a test" (p. 184). Evidence of validity is considered the most crucial element in developing and evaluating the quality of an educational assessment (Camara, 2003; Millett, Stickler, Payne, & Dwyer, 2007). Two important components for determining validity are evidence of appropriate content informed by experts in the field and examining whether the internal structure supports the constructs being measured (AERA, APA, & NCME, 1999; Kline, 2010, 2014). Recently, Pearl et al. (2017) completed a content validity study of the QIASD standards and quality indicators with a national sample of experts in the field. This researcher aims to extend this line of research to provide evidence of construct validity of the QIASD measure to produce valid interpretations of the scores.

Purpose of the QIASD

Evidence for validity must take into account the purpose of the measure and the inferences that will be made from an instrument's scores (Goe, Bell, &Little, 2008; Kane, 2006; Messick, 1989). Classroom observation data are important for gauging the quality of instruction and identifying professional development needs (Kane & Staiger, 2012). The QIASD was developed as a measure of the level of quality indicators present in classrooms serving students with ASD. The QIASD is meant to be implemented by administrators, teachers, or higher education faculty to provide actionable feedback to guide teacher improvement and determine

professional development needs. The QIASD was originally designed for use with special education teachers in self-contained special education classrooms serving students with ASD (Pearl et al., 2017). Ultimately, the purpose of the QIASD as a classroom observation measure is to improve teaching by assessing the specific qualities of effective instructional environments for students with ASD. The current research study on the internal consistency reliability and construct validity of the QIASD is a logical next step in gathering preliminary evidence supporting the psychometric properties of this instrument for use in classrooms serving students with ASD.

CHAPTER THREE: METHODOLOGY

Introduction/Statement of Problem

Over the past five years, value-added models for teacher evaluation, including those for special educators, have increasingly come into question (American Statistical Association, 2014; Harris & Herrington, 2015; Gansle et al., 2015). The recently reauthorized Every Student Succeeds Act (ESSA, 2015) maintains a focus on accountability systems and teacher effectiveness, but provides increased flexibility, placing the responsibility for developing and implementing teacher evaluation systems in the hands of states and local education agencies (LEA's). The ESSA provides federal funds via the Teacher and School Leader Incentive Program. The purposes are to support state and district innovation to "develop, implement, improve, or expand comprehensive performance-based compensation systems or human capital management systems for teachers, principals, or other school leaders…who raise student achievement" (section 2211 (a)(1)); and "To evaluate the effectiveness, fairness, quality, consistency, and reliability of the systems" (section 2211, (a)(2)).

Despite the strong evidence-base for a number of practices for teaching students with ASD, researchers have shown teachers lack preparation and support for the implementation of those EBPs (Belfiore, Fritts, & Herman, 2008; Brock, Huber, Carter, Juarez, & Warren, 2014; Browder & Cooper-Duffy, 2003; National Research Council [NRC], 2001). Project ASD's Quality Indicators for Classrooms Serving Students with ASD (QIASD) is an observational tool specifically designed to support special education teachers serving students with ASD with what Johnson and Semmelroth (2014) refer to as "detailed, actionable feedback to improve their

practices" (p. 68).

The results of the systematic literature review support the need for further research on developing psychometrically sound teaching evaluation instruments designed for special education classrooms serving students with ASD (Johnson & Semmelroth, 2012; Leaf et al., 2016; Odom et al., 2013; Pearl et al., 2017; Tsai et al., 2013). The problem is existing research in this domain is limited. Only two empirical studies were found at the intersection of teaching evaluation instruments designed for K-12 special education classrooms serving students with ASD and evidence of psychometric properties (validity and reliability). The purpose of this work is to establish the internal consistency reliability and the construct validity of the QIASD measure using a confirmatory factor analysis (CFA) framework.

Research Questions

RQ 1a: To what extent does the QIASD produce reliable scores as measured by internal consistency reliability?

RQ 1b: To what extent is construct validity of the QIASD achieved as measured by a confirmatory factor analysis?

Table 3

Research Questions Matrix

Research question	Data Type	Instrument	Sample	Data analysis
RQ 1a: To what extent does the QIASD produce reliable scores as measured by internal consistency reliability?	Item, subgroup, and total scores; means and standard deviations ($N > 100$)	QIASD instrument	Trained Graduate students in the Special Education Master's Program/ Observations in classrooms serving students with ASD	Descriptive statistics and Cronbach's alpha
RQ 1b: To what extent is construct validity of the QIASD achieved as measured by a confirmatory factor analysis?	Individual item scores, factor scores ($N > 100$)	QIASD Instrument	Trained Graduate students in the Special Education Master's Program / Observations in classrooms serving students with ASD	Confirmatory Factor Analysis (CFA)

Method

Sample

A purposeful sample of data was selected from special education graduate student observations of classrooms serving students with ASD. Permission to conduct the study was acquired through the institutional review board (IRB). This study was exempt from human research by the IRB because direct contact with observers was made and no participant-identifying information was accessed.

The sample in this study is composed of the QIASD classroom observations completed by graduate student observers. The observers were all enrolled in graduate level special

education courses at a central Florida university. A purposefully selected sample from special education graduate student observers was based on the following: (a) they were enrolled in coursework relevant to teaching students with ASD, (b) they had experience in classrooms as teachers, administrators, or other educational positions, and (c) they suited a main purpose of the QIASD to assess the gap between teacher skills and classroom application. The observers were from a range of counties and school districts across central Florida. Non-identifying demographic information was collected on the observers and the classroom observation settings (see Chapter Four, Table 6).

Sample Size

The target sample size for this study was 100-150 special education classroom observations utilizing the QIASD. The observations were conducted by special education graduate students across two semesters in order to obtain a large enough sample size for a confirmatory factor analysis (N > 100). Sample size requirements were based on a power analysis, literature, and available resources for practically acquiring the sample (Kline, 2015). Multiple rules of thumb are proposed to identify minimum sample size needed for confirmatory factor analysis (Field, 2013; Harrington, 2009; Hoyle, 2000; Kline, 2015). Kline (2015) recognized the number of model parameters, the estimation method, the normality of the data distribution, and the number of indicators per factor all impact sample size requirements. Jackson (2003) suggested the ratio of the number of participants (*N*) to the number of model parameters (indicators) should be at least 10:1 when using maximum likelihood estimation. Tabachnik and Fidell (2012) proposed models with "strong expected parameter estimates and reliable variables

may require fewer participants" (p. 688). Further, Gignac (2006) viewed sample size in factor analysis as a suggestion that should be tested and that a sample of 100 may be appropriate (Harrington, 2009).

Previous validation studies of teaching practice measures were reviewed to assist with identifying an acceptable sample size. Marshall, Stuart, and Alston (2016) reported on the development and validation of the Teacher Intentionality of Practice Scale (TIPS) as a measure of teaching practice and growth over time. The TIPS was designed with seven core indicators of teacher instructional practice and 14 sub-indicators. The authors used a sample of 76 observations of 37 teachers conducted by three observers and based the adequacy of their sample size on a 3:1 ratio of observations to variables for a factor analysis (Osborne & Costello, 2005). Tsai, Cheney, and Walker (2013) created the Participatory Evaluation and Expert Review for Classrooms Serving Students with EBD (PEER-EBD) comprised of four main constructs and 19 best practices with three to nine indicators. In this study of content validity, internal consistency, and test of model fit with confirmatory factor analysis, the authors used a sample of 145 staff raters across 23 K-12 classrooms. Reddy, Fabiano, Dudek, and Hsu (2013) used a sample of 317 general education teachers with 67 observers in their development and construct validity study of the Classroom Strategies Scale-Observer Form.

Power Analysis

MacCallum, Browne, and Sugawara (1996) explained a method to identify the minimum sample size needed in factor analysis to attain a desired level of power. Statistical power refers to

the probability of rejecting a null hypothesis when it is false (Gall et al., 2007; Kline, 2015). MacCallum et al. (1996) determined power estimates for various samples sizes and degrees of freedom that indicated "the probability of rejecting the hypothesis of close fit under these conditions" (p. 141). The estimated sample size necessary for tests of close model fit with 100 degrees of freedom and a power of .08 was 132. In this study, the hypothesized QIASD model had 1224 degrees of freedom, suggesting a sample even smaller than 132 may be acceptable. Further, MacCallum and colleagues (1996) provided an estimate for power of about 0.650, based on an alpha of .05, for a sample size of 100, and 0.955 for a sample size of 200.

The proposed sample of N = 100-150 for this study is comparable with the literature and is within range of the power analysis based on Cohen's (1988) recommendation of acceptable power at .80.

Setting

Observations occurred in K-12 special education classrooms serving a minimum of two students with autism spectrum disorder. The graduate students conducted observations within special education classrooms other than their own to improve the validity of the data. Demographic information on the observation settings is summarized in Table 7 (Chapter Four).

Instrument

The Quality Indicators for Classrooms Serving Students with Autism Spectrum Disorders (QIASD) is a content validated instrument designed to evaluate the presence of specific

educational quality indicators in classrooms serving students with autism spectrum disorder (Pearl et al., 2017). The QIASD observation tool is intended to provide actionable feedback to special educators serving students with ASD (Pearl et al., 2017). The QIASD was developed as a product of Project ASD, a teacher preparation program at the University of Central Florida, funded through the Office of Special Education Programs (Pearl et al., 2017). A copy of the QIASD instrument may be viewed in Appendix B.

The QIASD includes indicators from the Observation Assessment for Teachers Providing Services to Students with Autism Spectrum Disorders (OAASD), the product of a PEPSA (Partnership for Effective Programs for Students with Autism) grant project and Florida Centers for Autism and Related Disabilities (CARD). The QIASD incorporates revisions and additions to the quality indicators based on field testing of the OAASD, review of the literature, and alignment with the Council for Exceptional Children's Initial Special Educator Standards Specialty Set: Developmental Disabilities and Autism Spectrum Disorder (CEC, 2015; Pearl et al., 2017). The CEC Initial Preparation Standards delineate the pedagogical knowledge and skills teacher candidates must master to effectively teach in a classroom (CEC, 2015).

The CEC standards have been iteratively developed since the 1980's with input from stakeholders to encompass teaching principles deemed important to professionals in the field (CEC, 2015). The CEC Specialty Sets were built to meet the specialized needs of teachers focused on different disability areas. The CEC standards include precisely described knowledge and skills that are continuously validated for content by a team at CEC with input from external professionals and experts in the related specialty set fields (CEC, 2017). The Council for Exceptional Children partners with the Council for the Accreditation of Educator Preparation

(CAEP) to inform teacher preparation programs and recognize those aligned with the CEC preparation standards (cec.sped.org).

The Initial Specialty Set for students with developmental disabilities and ASD describes essential knowledge and skills required for special education teachers to serve this population of students. The set includes seven standards: (1) learner development and individual learning differences, (2) learning environments, (3) curricular content knowledge, (4) assessment, (5) instructional planning and strategies, (6) professional learning and ethical practice, and (7) collaboration. (CEC, 2015). These seven standards describe critical teacher knowledge and skills based on the learning characteristics of students with developmental disabilities and ASD.

The QIASD contains 51 quality indicators aligned with the seven CEC Specialty Set standards for students with ASD (CEC, 2015; Pearl et al., 2017). Indicators are scored on a rating scale of N/A to 4, representing the degree each indicator is present in the classroom during a one-hour observation session. Quality indicator ratings may be derived from three different data collection methods – direct observation, interview, or artifacts – to ensure all items have the opportunity to be scored across observations (see Table 4).

Table 4

Quality Classroom Indicator (QI) Scoring and Data Collection

Quality Indicator Rating	Data Collection Methods
4: Highly Effective (Very Much Present) 3: Effective (Present) 2: Needs Improvement (Somewhat Present) 1: Developing (Not Present) 0: Unsatisfactory (Not Present) N/A: Unrated (No opportunity to observe)	Direct Observation Teacher Interview Artifact Review

Observers have the option for interviewing the teacher to gather information on indicators that may not be directly observed in the classroom (i.e., family involvement in the IEP or students referred for a functional behavior assessment). Observers also may examine artifacts as evidence of an indicator (i.e., lesson plans, behavior intervention plans, or data on IEP goals). A *Comments* section is provided in the QIASD instrument for each of the seven standards to allow observers to record specific examples supporting their ratings.

For the purpose of this study, the QIASD was configured into an on-line format using the university's Qualtrics platform. Qualtrics (www.qualtrics.com) is a web-based survey software allowing the researcher to have a consistent data collection method and to easily track participant responses. The QIASD instrument was accessible to participants in this study through a provided URL link. A sample of the QIASD Instrument in Qualtrics is presented in Figure 1.

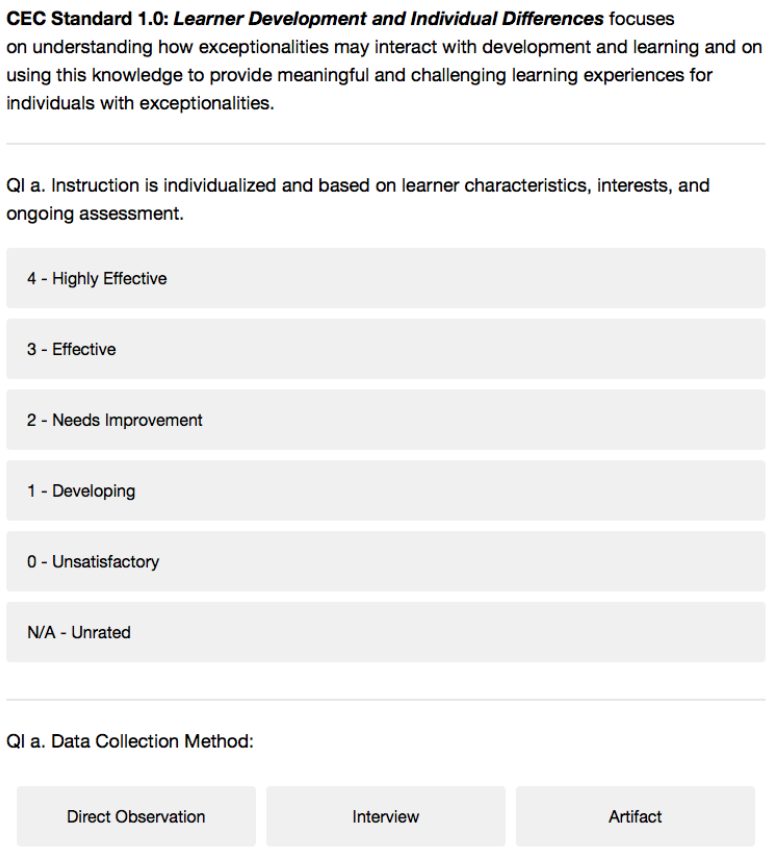


Figure 1: Sample of QIASD Instrument in Qualtrics

Procedures

Observer Training and Interrater Reliability

A QIASD training protocol was created to enhance reliable observations and minimize rater bias (Little et al., 2009; Wu, Tam, & Jen, 2016). The training protocol consisted of three main components: (1) Adobe tutorial, (2) tutorial quiz, and (3) scoring reliability with a practice video. The training protocol was incorporated into the assignment modules within relevant graduate courses. Prior to conducting on-site classroom observations, observers

completed all three components of the training protocol and met criterion for rater reliability with practice video scoring.

Observers watched a 40-minute Adobe Connect tutorial created by this researcher and accessed by observers through a link in the course modules. The QIASD tutorial introduced the elements of the QIASD instrument and provided a detailed description of each quality indicator. Specific verbal and visual examples were included to demonstrate what evidence of the quality indicators might look like during the classroom observation.

The tutorial explained the rating scale for scoring the quality indicators and the data collection methods. Observers were trained to recognize the indicators as present in a classroom at different levels of effectiveness (i.e., 4 = highly effective, 1 = unsatisfactory). Observers also learned to identify when there was no opportunity to observe an indicator during the one-hour session and to follow-up with teacher interview and/or artifact review. A hard copy of the training protocol was provided in the course module for reference. A sample of the Adobe tutorial training is provided in Appendix 10.

The Adobe tutorial was followed by a quiz designed to assess observer ability to identify correct and incorrect quality indicator ratings based on 10 sample classroom scenarios. Observers were required to meet 90% criterion on the post-tutorial quiz in order to receive a link to the on-line QIASD. Specific feedback was provided for correct and incorrect responses to maximize learning. Observers had the opportunity to take the tutorial quiz up to three times to score 90% or above. Overall, 40% of observers met criterion on the tutorial quiz on the first attempt and 60% met criterion on the second attempt.

Interrater Reliability of Practice Scoring

The accuracy of graduate student administration of the QIASD was assessed using a practice classroom video with 30% of observers. The master scores were set based on the average practice video scores obtained by this researcher and one of the QIASD developers with expertise using the QIASD. Graduate students in one course (n = 36) were required to score the QIASD on a 20-minute sample classroom video of a middle school special education classroom serving students with ASD. Practice scores were uploaded in a course quiz and responses were compared to the master scores. Graduate students had three opportunities to score the practice video to meet the 80% criterion for acceptable QIASD administration. Specific feedback was automatically provided on correct and incorrect items to further inform observers on the scoring process. The mean interrater agreement between these graduate student scores and the master score for the practice video was 81% exact agreement and 100% agreement within two points of the master score.

Upon completion of all steps of the training protocol, observers were provided a link to the online Qualtrics version of the QIASD to conduct on-site classroom observations. Written procedures were available in the course modules to support observers in setting up and completing the classroom observation (see Appendix E).

Data Collection

Prior to data collection, the researcher received permission to conduct the study from the University of Central Florida's Institutional Review Board (see Appendix A). Once the proposed

study was approved by the IRB, the researcher began gathering the data. The sample of classroom observations across two semesters (fall 2017 and spring 2018) were completed by graduate students in the special education program at a central Florida university. As part of their coursework, the graduate students conducted a one-hour classroom observation using an online Qualtrics version of the QIASD instrument created by the researcher with permission from the instrument developers, Pearl et al. (2017). Relevant special education course modules were modified to reflect the process of using the QIASD for a field-based assignment. Course module development included: (a) outlining the steps for graduate students to complete a classroom observation using the QIASD; (b) creating a training to ensure observers understood the QIASD components and scoring system; (c) building in a measure of score reliability through a practice training video to assess observer accuracy using the rating scale compared to a master score; and (d) transferring the QIASD to Qualtrics for a secure database of observation responses.

As part of field-based assignments in graduate courses, observers completed one-hour observations using the QIASD in K-12 classrooms serving students with ASD. Observer scores were collected using the online Qualtrics version of the QIASD. Faculty provided the researcher access to the raw data in Qualtrics after removal of observer-identifying information (i.e., observer names). Data were stored on a secure university server. Participant consent was not required due to the IRB exempt status for non-human research.

Research Design

A confirmatory factor analysis (CFA) was used to investigate the construct validity of the QIASD measure. The fit of the hypothesized QIASD factor structure with the sample data was compared to an alternative model (AERA et al., 2014; Kline, 2016). A CFA offers evidence of construct validity "if the factor structure of the scale is consistent with the constructs the instrument purports to measure" (Floyd & Widaman, 1995, p. 287). The design of this study allows for an iterative analysis of the internal structure of the QIASD based on the CFA results.

Data Analysis

The QIASD scoring system is an interval scale from N/A and 0-4. For the purpose of this study, the researcher recoded the ratings to give "N/A" a numerical value of 0 because it is a meaningful score indicating no opportunity to observe the item. The scale was recoded as 0-5 for the analyses in this study to represent the full continuous range of scores. Data were screened for missing values and outliers using IBM SPSS version 24. Descriptive statistics were run to provide observer and setting demographics and measures of central tendency. Statistical assumptions for conducting a CFA were examined, including sample size, factorability, normality, linearity, and multicollinearity. Cronbach's alpha was calculated to measure internal consistency reliability and a CFA was run in SPSS AMOS version 23 to assess the construct validity of the QIASD scale. Further analyses and methodological adjustments may be implemented based on the results of initial data analyses.

Internal Consistency Reliability

Research Question 1a.: To what extent does the QIASD produce reliable scores as measured by internal consistency reliability?

The measurement scores should first be analyzed for good reliability to conduct a confirmatory factor analysis and make inferences from the data (Kline, 2016). The reliability of an instrument is a gauge of how well the item scores measure what the test is intended to measure. Based on classical test theory, the reliability of a measure depends on the degree the scores are free of measurement error (Gall et al., 2007). The internal consistency reliability of a measurement refers to the degree the indicators produce consistent scores to reflect the same construct (Field, 2013; Gall et al., 2007). The internal consistency reliability of the QIASD was assessed in this study to determine how well the quality indicator scores provide a consistent measure of the seven special education teaching dimensions identified for classrooms serving students with ASD.

Cronbach's alpha (Cronbach, 1951) was calculated to measure the correlations between scores on the individual indicators, subsets, and the overall QIASD scale. Reliability coefficients range from values of .00 to 1.00 and are estimated as "one minus the proportion of total observed variance due to random error" (Kline, 2016, p. 90). If the alpha coefficient value is low, then the indicator scores are not highly correlated, signifying large measurement error and low reliability. A threshold value of .70 was used to indicate good internal consistency reliability (Kline, 2016).

Construct Validity

Research Question 1b.: To what extent is construct validity of the QIASD scores achieved as measured by a confirmatory factor analysis?

The hypothesized relationship between the seven proposed constructs and 51 quality indicators of the QIASD was tested with a CFA. The hypothesized CFA model was based on prior research and theory supporting the CEC standards (CEC, 2014, 2017) and the QIASD indicators recently validated by expert review and feedback (Pearl et al., 2017). The preferred model is guided by the theoretical framework of the CEC standards with the 51 indicators conceptually grouped under the seven CEC standards. Based on the results of the seven-factor model CFA, an appropriate alternative model was specified for comparison to investigate the construct validity of the QIASD. The AMOS software version 24 (Arbuckle, 2010) was used to complete the initial confirmatory factor analyses.

The measurement structure of the QIASD instrument was represented by the relationships between the latent variables (CEC standards) and the underlying observed variables (quality indicators). The CFA was used to verify the pattern of loadings on the proposed factors within the preferred seven-factor structure of the QIASD Instrument (Brown, 2014; Field, 2013). A construct refers to a concept or an attribute that is not operationally defined and is based on theory and/or prior research (Cronbach & Meele, 1955; Harrington, 2009). A construct may have one or several dimensions, which are called factors or latent variables (Brown, 2014). Observed variables are the measured variables representing evidence of the theoretical factor (Brown, 2014; Tabachnick & Fidell, 2013). The a priori hypothesized CFA model of the QIASD consisted of 51 observed variables (quality classroom indicators) loading onto seven latent

variables (the subgroups aligned with the CEC special educator preparation standards). The model pathways diagram (see Figure 2) visually represents the hypothesized factor structure of the QIASD.

Model Fit Indices

Multiple model fit indices and cutoff values were selected based on recommendations found in the literature (Brown, 2006; Hu & Bentler, 1999; Kline, 2016; Kock, 2015). First, the chi-square goodness-of-fit (X^2) was reported with a cutoff value of $p > .05$. The root mean square error of approximation (RMSEA; < = .08) was reported although it may be less accurate with small samples (Hu & Bentler, 1999). The Tucker-Lewis fit index (TLE; > = .95) also was included as it is preferable for smaller sample sizes (Brown, 2006). The goodness-of-fit (GFI; >= .95) and the adjusted goodness-of-fit (AGFI; > = .90) indices were used to provide a different conceptualization of model fit with adjustments for the number of parameter estimates (Tabachnick & Fidell, 2013). The standardized root mean square residual (SRMR; < = .08) was included with the alternative model, which is a fit index based on the residuals (Brown, 2006; Hu & Bentler, 1999).

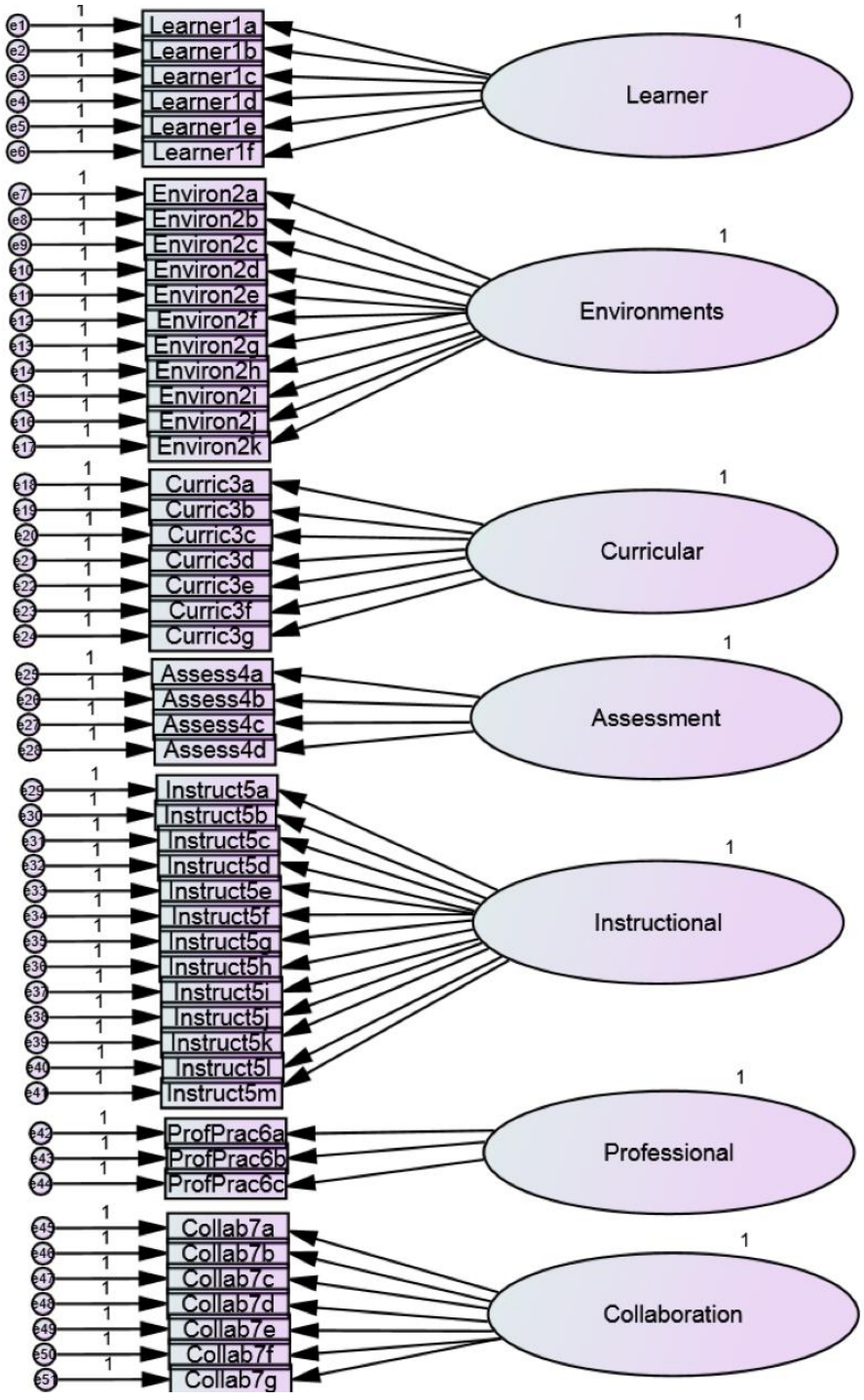


Figure 2: Preferred Seven-Factor Structure for CFA

The results were examined to verify the hypothesized pattern of loadings onto the seven factors (CEC standards) of the QIASD (Brown, 2014). The criteria for statistical significance of variable loadings onto factors were checked with consideration of sample size (Field, 2013). Factor loadings are the regression coefficients for predicting the indicators from the latent factors (Harrington, 2009). Tabachnick and Fidell (2007) suggested a general rule of thumb for interpreting loadings as excellent (> 0.71), very good (> 0.63), good (> 0.55), fair (> 0.45), and poor (> 0.32). Stevens (2002) suggested statistically significant loadings of greater than .512 for a sample size of 100 and larger loadings of .772 for a smaller sample size of 50. Kline (2015) recommended a moderately high magnitude of .50 or greater for significant factor loadings.

Factor Models.

Multiple CFA's were conducted to analyze the fit of the data to the preferred seven-factor model and the relative fit of the data compared to the alternative models (Kline, 2015; Tabachnick & Fidell, 2001). Alternative models may be determined a priori based on theory and supporting literature, or after testing the preferred model factor loadings and examining the fit indices for amount of correlation between variables (Kline, 2015; Reddy et al., 2013; Tabchnick & Fidell, 2007). The next section details the preferred seven-factor model and two proposed alternative models based on the literature. The results of the initial CFA analysis ultimately informed the researcher on the appropriate alternate model(s) for comparison.

The seven-factor first-order model was the preferred structure for the QIASD. The theoretical construct defined as effective teaching practices of students with ASD was supported through alignment with the seven CEC standards and corresponding quality classroom indicators

(see Figure 2). The seven-factor model was specified as 51 observed variables loading onto seven latent variables: Learner Development and Individual Learning Differences (Learner; 6 indicators), Learning Environments (Environ; 11 indicators), Curricular Content Knowledge (Curric; 7 indicators), Assessment (Assess; 4 indicators), Instructional Planning and Strategies (Instruct; 13 indicators), Professional Learning and Practice (ProfPrac; 3 indicators), and Collaboration (Collab; 7 indicators).

The proposed fourteen-factor alternative model was conceptualized from the Autism Program Quality Indicators (APQI; Crimmins et al., 2001). The APQI consists of 81 autism program quality indicators clustered under 14 categories: individual evaluation (8 indicators), development of the Individualized Education Program (8), curriculum (7), instructional activities (5), instructional methods (6), instructional environments (4), review and monitoring of progress and outcomes (4), family involvement and support (7), inclusion (4), planning the move from one setting to another (5), challenging behavior (9), community collaboration (3), personnel (6), and program evaluation (5). Similar to the QIASD, the APQI dimensions were founded on professional literature, field testing, and feedback from national experts and other stakeholders (Crimmins et al., 2001). The main difference is the QIASD alignment with the seven CEC standards. The fourteen-factor alternative QIASD model would hypothesize the 51 observed variables realigned with the fourteen APQI categories.

The alternative one-factor model combined all seven CEC standards into one over-arching construct to represent quality teaching of students with ASD. The one-factor model consisted of the 51 observed variables (quality indicators) loading onto one latent variable to assess the model fit to one overarching construct.

Model Identification

Identification of the measurement model for the CFA was established as follows. (Kline, 2015; Tabachnick & Fidell, 2013). The latent variables (factors) in the model are hypothetical and require a scale. The factors were scaled by setting their variances to 1.0. The error variance in the model is the variance not due to the factor. Tabachnick & Fidell's (2013) steps were used to identify the model parameters. The number of data points were compared to the number of parameters in the model. The hypothesized model for the QIASD had $(51(51 + 1))/2 = 1326$ data points. The number of parameters of the model was calculated by adding the number of regression coefficients, variances, and covariances estimated in the model. The hypothesized QIASD model specified 51 regression coefficients (observed variables), 51 error terms, and 0 covariances (between factors), totaling 102 parameters. The QIASD model had more data points (1362) than parameters (102), which is required for model identification. Finally, the model degrees of freedom were calculated by subtracting the parameters from the number of data points, equaling 1224. Thus, the hypothesized QIASD model specified 102 parameter estimates with 1224 degrees of freedom. The model was considered overidentified and ready to proceed with the CFA (Kline, 2015; Tabachnick & Fidell, 2013).

Model Estimation

Maximum likelihood (ML) is considered the standard method to estimate a factor model (Brown, 2014; Hoyle, 2000). According to Brown (2006), the ML estimator "aims to find the parameter values that make the observed data most likely" (p. 73). The ML method is desirable because it provides standard errors for each parameter estimate that can be used to calculate p-

values and confidence intervals (Brown, 2014). Three main assumptions for ML estimation are a large enough sample size, observed variables with continuous measurement scales, and multivariate normality. (Brown, 2014; Tabachnick & Fidell, 2013). The use of the ML estimation method or other more appropriate estimation methods was determined upon results of the data.

Assumptions

The following assumptions were considered to conduct a confirmatory factor analysis (Lomax & Hahs-Vaughn, 2012; Kline, 2015):

1. Independence of Observations – The study was designed to maximize independent observations using data from a sample of classroom observations conducted in different geographical locations. Classroom observations using the QIASD were conducted by 102 graduate students across 19 districts and 73 different schools across Florida.
2. Adequate Sample Size – A target sample size of 100-150 classroom observations was based on a combination of recommendations from the literature, an a priori power analysis, and available resources for practically acquiring the sample for this study (Kline, 2015). A large sample size is recommended because correlations are less stable with smaller samples (Tabachnick & Fidell, 2013). A post hoc power analysis was conducted to confirm an adequate sample size.
3. Continuous Measurement Scale – The QIASD has a numerical interval measurement scale with six rating options and was considered a continuous measure.

4. Normality – The researcher examined skewness and kurtosis to check for normal distribution of data. The standard ML estimator is reasonably robust to violations of multivariate normality (Hoyle, 2000). In the case of non-normal data, more appropriate estimation methods were considered, such as the robust Satorra-Bentler method (Satorra & Bentler, 1988) and the Bollen-Stine method (Bollen-Stine, 1992).
5. Linearity – Linear relationships between variables were checked through correlations and visual inspection of bivariate scatterplots. Variables with very low correlations below .3 should be reconsidered or transformed (Field, 2013; Tabachnick & Fidell, 2013).
6. Multicollinearity – Pearson correlation coefficients were inspected for variables too highly related. Extremely high multicollinearity obscures any unique contributions of those variables to a factor (Field, 2013; Tabachnick & Fidell, 2013).
7. Outliers – Outliers have greater impact on the overall variable than other scores and can lead to Type I or Type II errors (Tabachnick & Fidell, 2013). The data was checked for multivariate outliers through Cook's distance to detect cases with high discrepancy from others.
8. Factorability – Tabachnick and Fidell (2013) suggested a correlation matrix should include "several sizable correlations" (p. 619) as evidence to support factors. Correlations were checked for those exceeding .30 as evidence of relationships between the variables underlying each factor (Tabachnick & Fidel, 2013). The Kaiser-Meyer-Olkin measure of sampling adequacy (KMO; Kaiser, 1970) was also used to verify an adequate sample size for separating observed variables into factors (Field, 2013). A KMO value of .5 or .6 and

above are recommended for factorability of observed variables (Kaiser, 1974; Tabachnick & Fidell, 2013).

Potential Limitations

In classical test theory, measurement error refers to the discrepancy between a hypothetical true score and the score actually obtained (Gall et al., 2007). All measures of hypothetical constructs will have measurement error to some extent. Measurement error may occur from conditions in the environment during assessment, variability in how raters feel, and raters not consistently implementing or follow scoring procedures (Gall et al., 2007). Internal consistency reliability was assessed to determine the level of measurement error. Observers completed a thorough training protocol and were assessed on the QIASD rating system prior to conducting observations in order to minimize potential scoring error.

Demographic information about the observers and setting was collected to reduce sampling and selection effects (MacCallum & Austin, 2000). The QIASD items were operationally defined for observers in a tutorial to minimize selection effects and maximize valid interpretations of the results. Selection of QIASD quality indicators was supported by the literature on CEC special educator standards and evidence-based practices for effectively teaching students with ASD (CEC, 2008; Little et al., 1999).

Observer effects (Gall et al., 2007) were minimized as follows:

(a) Observer bias was minimized by operationally defining and providing training on quality indicators and scoring. An interrater reliability check was conducted for a 20% sample of

observes using a practice scoring video compared to master scores. The training emphasized the QIASD was not a teacher evaluation, rather a positive tool for helping special educators grow and strengthen the environment for teaching students with autism.

(b) Observer omission was addressed within the QIASD scoring system, which allowed for teacher interviews and artifact reviews when there was no opportunity to observe certain items. A reminder for observers to review and check all indicators were scored before submitting was included on the last page of the Qualtrics QIASD.

(c) Observer drift is a potential decline in the observer's skill to collect data as prescribed. To minimize this effect, observers completed the on-site QIASD observation within two weeks of completing a training tutorial and passing a quiz to criterion.

Finally, caution must be taken not to overly favor the preferred model being examined (MacCallum & Austin, 2000). This confirmation bias was mitigated by considering and discussing alternative models of fit for the data.

CHAPTER FOUR: RESULTS

Overview of Data Analysis

The purpose of this study was to establish internal structure validity evidence of the Quality Indicators for Classrooms Serving Students with Autism Spectrum Disorder (QIASD) ratings for the intended use in evaluating teaching performance for students with ASD. Chapter Four features the iterative data analyses procedures conducted to gain the most valid conclusions with the final population sample. Results are presented in this chapter for (a) sampling, (b) descriptive statistics, (b) and data analyses per research question. The study was designed to answer the following research questions:

Research Question 1a: To what extent does the QIASD produce reliable scores as measured by internal consistency reliability?

Research Question 1b: To what extent is construct validity of the QIASD scores achieved as measured by a confirmatory factor analysis?

Research Hypothesis: The researcher hypothesized the 51 observed variables (quality indicators) of the QIASD will fit the proposed reflective model by loading onto the seven latent variables (factors) with these data.

Data Screening Results

A total of 121 QIASD responses were collected through Qualtrics. First, these data were inspected for missing variables within cases (rows) and variables (columns). Nineteen cases

(rows) were removed due to fully incomplete responses, thus with more than 20% missing data points. These 19 cases were likely respondents who accessed Qualtrics to preview the QIASD and submitted the form without inputting any data. Of the remaining 102 responses (N = 102), there were seven variables with missing values all less than 5% missing. An inspection of the data showed both variables (columns) and cases (rows) with missing data had no more than 2 missing data points. The data set had only 8 total missing data points, which is less than 1% of the data. The missing data points were imputed using regression due to the very small number of missing variables (Tabachnick & Fidell, 2013).

The data were checked for multivariate outliers through Cook's distance to detect cases with high discrepancy from others. Cook's distance greater than 1.0 may suggest a case has undue influence on the overall model (Field, 2013). The maximum value of Cook's distance in this sample was .360, which indicated no problematic outliers in these data.

Descriptive Statistics Results

Observer Demographics

Ninety-nine of the total 102 graduate student observers responded to the initial demographic questions (see Table 5). These respondents (n = 99) consisted of 88 females (89%) and 11 males (11%). The reported ethnicity of observers (n = 99) was 76 (77%) white, 13 (13%) African-American, and 10 (10%) other ethnicity. Eighty-four (85%) were certified in special education and fifteen (15%) in general education. A majority of observers, 76 (77%), were currently pursuing state endorsement for teaching students with autism spectrum disorder, 5%

already had ASD endorsement, and 18% were not seeking state endorsement. Current job roles included 74 (75%) special education teachers, 11 (11%) general education teachers, 2 (2%) administrators, and 12 (12%) others (i.e., Title-1 teachers, ESE specialist, School counselor, instructional assistant, and staffing specialist). The number of years observers had been teaching were 10 (10%) less than one year, 19 (19%) between 1-2 years, 26 (27%) between 3-5 years, 25 (25%) between 6-9 years, and 19 (19%) had been teaching 10 or more years.

Table 5

Observer Characteristics

		Percentage
Gender	Female	89%
	Male	11%
Ethnicity	African American/Black	13%
	White / Caucasian	77%
	Other	10%
Teaching Certification	General Education	15%
	Special Education	85%
ASD Endorsement	Yes	5%
	Currently Pursuing	77%
	No and Not Pursuing	18%
Current Job / Position	Special Education Teacher	75%
	General Education Teacher	11%
	Administrator	2%
	Other	12%
Number of Years Teaching	<1	10%
	1 - 2	19%
	3 - 5	27%
	6 - 9	25%
	10 +	19%

Setting Demographics

All observers (N = 102) responded to the demographic questions on setting (see Table 6). The special education classroom observations were conducted across 19 districts in a total of 73 different schools, comprised of 47 (64%) elementary schools, 16 (22%) middle schools, and 10 (14%) high schools across Florida. Of these schools, 25 (34%) were located in urban areas, 43 (59%) in suburban areas, and 5 (7%) in rural areas. Sixty-eight graduate students observed in elementary classrooms (grades kindergarten to 6), twenty-two observed in middle school classrooms (grades 7 to 8), and twelve in high school classrooms (grade 9 to age 22). During the time of their observations, graduate students reported 69 (68%) classrooms had 10 or less students, 20 (20%) classrooms had between 11-15 students, 11 (10%) had between 16-20 students, and 2 (2%) classrooms had more than 20 students present. The number of staff in observed classrooms ranged from 1 to 6, with 83 (81%) classrooms having 2-3 staff present, 5 (5%) having only 1 staff member, 12 (12%) having 4 staff members, and the remaining 2 (2%) classrooms with 5 and 6 staff members.

Table 6

Setting Characteristics

		Percentage
Location	Rural	7%
	Suburban	59%
	Urban	34%
School Type / Grade Level	Elementary (1 – 6)	64%
	Middle (7 – 8)	22%
	High (9 – Age 22)	14%
Number of Students in Class	< = 10	68%
	11 - 15	20%
	16 - 20	10%
	> 20	2%
Number of Staff	1	5%
	2 - 3	81%
	4	12%
	5 - 6	2%

Measures of Central Tendency

Tables 7 and 8 show the results of descriptive statistics run to examine the central tendencies and range of ratings scored by observers. Reporting measures of central tendency can assist future researchers with replication and verification of results (Kline, 2015). The mean ratings for observed variables with these data ranged from 1.47 to 4.59, with standard deviations between .598 and 2.124. Observers used the full range of five ratings in scoring the majority of quality indicators (42 out of 51). Six indicators had smaller ranges of ratings, with more responses towards the higher end of the scale.

Table 7

QIASD Observed Variables Descriptive Statistics (*N*=102)

Observed Variable	Mean	Std. Deviation	Range	Skewness	Kurtosis
Learner1a	4.07	.693	3	-0.456	0.358
Learner1b	4.20	.985	4	-1.672	3.188
Learner1c	4.02	.995	5	-1.886	5.085
Learner1d	3.75	1.467	5	-1.619	1.824
Learner1e	3.75	1.318	5	-1.131	0.868
Learner1f	3.89	1.033	5	-1.648	4.21
Environ2a	4.39	.810	4	-1.75	4.391
Environ2b	4.59	.635	2	-1.287	0.529
Environ2c	4.02	1.386	5	-1.584	1.861
Environ2d	4.02	1.177	5	-1.376	1.577
Environ2e	3.09	1.642	5	-0.459	-1.032
Environ2f	4.08	.972	5	-1.612	4.479
Environ2g	3.92	1.232	5	-1.597	2.63
Environ2h	4.62	.598	3	-1.603	2.942
Environ2i	4.12	.836	3	-0.433	-0.899
Environ2j	2.40	1.809	5	-0.222	-1.485
Environ2k	3.29	1.870	5	-0.847	-0.837
Curric3a	3.91	1.187	5	-2.111	4.801
Curric3b	3.83	1.343	5	-1.79	2.84
Curric3c	3.85	1.238	5	-1.79	3.453
Curric3d	3.52	1.419	5	-1.556	1.646
Curric3e	3.16	1.756	5	-0.927	-0.582
Curric3f	3.33	1.563	5	-1.001	0.011
Curric3g	3.35	1.398	5	-1.033	0.509
Assess4a	3.80	1.298	5	-1.457	1.916
Assess4b	3.80	1.235	5	-1.643	2.497
Assess4c	3.70	1.434	5	-1.401	1.267
Assess4d	3.48	1.876	5	-1.102	-0.378
Instruct5a	3.80	1.275	5	-1.993	3.842
Instruct5b	4.02	.901	3	-0.536	-0.588
Instruct5c	4.28	.883	5	-1.65	4.514
Instruct5d	4.24	.823	5	-1.661	5.837
Instruct5e	3.66	1.368	5	-1.627	2.179
Instruct5f	4.41	.680	2	-0.733	-0.579
Instruct5g	3.77	1.052	5	-1.666	4.464
Instruct5h	4.09	.785	4	-0.784	1.267
Instruct5i	3.89	1.142	5	-1.614	3.331
Instruct5j	3.96	1.125	5	-2.262	5.833
Instruct5k	3.62	1.379	5	-1.382	1.52
Instruct5l	1.97	2.046	5	0.302	-1.634
Instruct5m	3.73	1.394	5	-1.775	2.515
ProfPrac6a	3.98	1.442	5	-2.007	3.169

Observed Variable	Mean	Std. Deviation	Range	Skewness	Kurtosis
ProfPrac6b	4.19	1.241	5	-2.298	5.345
ProfPrac6c	2.64	2.124	5	-0.346	-1.732
Collab7a	3.28	1.445	5	-0.754	-0.34
Collab7b	4.19	1.132	5	-2.213	5.701
Collab7c	4.37	.994	5	-2.721	9.334
Collab7d	3.99	1.472	5	-1.9	2.736
Collab7e	3.16	1.773	5	-0.851	-0.624
Collab7f	1.47	1.974	5	0.73	-1.276
Collab7g	3.09	1.904	5	-0.769	-0.966

Table 8

QIASD Latent Variables and Total Descriptive Statistics (N = 102)

Scale	N of items	Mean (M)	SD	Skewness	Kurtosis
Learner	6	3.94	.726	-1.026	.711
Environment	11	3.55	.564	-.326	-.756
Curriculum	7	3.57	.937	-.884	1.091
Assessment	4	3.70	1.157	-1.286	1.410
Instruction	13	3.80	.592	-.253	-.347
Professional Practice	3	3.60	1.09	-.783	.193
Collaboration	7	3.36	.917	-.787	.918
Total Score	51	3.72	.579		

Statistical Assumptions Results

The statistical assumptions for confirmatory factor analysis (CFA) were tested using IBM SPSS version 23. The following section provides the results of these analyses of statistical assumptions including sample size, factorability, normality, linearity, and multicollinearity (Kline, 2016; Tabachnick & Fidell, 2013).

Sample Size

This investigation examined the internal structure of the QIASD with a purposive sample of 102 observations (N = 102) conducted over two course semesters by graduate students in K-12 special education classrooms serving students with ASD. The researcher aimed for a sample size of 100-150 classroom observations based on a combination of recommendations from the literature, an a priori power analysis, and available resources for practically acquiring the sample for this study (Kline, 2015).

The final sample size was unexpectedly impacted by variables outside the researcher's control. The researcher previously confirmed with faculty the QIASD classroom observation would be completed as an assignment within two courses across the fall and spring semesters. However, one of the courses in the spring semester was taught by a different faculty member who had already incorporated a different classroom observation assignment not using the QIASD. The faculty member agreed to offer the QIASD assignment as an optional replacement for two quiz grades. Only 2 graduate students in that course opted to complete the QIASD observation. This impacted the overall observation sample size the researcher was able to obtain for this study. The final sample (N=102) of classroom observations was less than the close to 150 observations originally anticipated.

Factorability

The R-matrix was inspected for Pearson correlation coefficients above .30. Multiple variables with higher correlations suggest feasibility of factoring the variables (Tabachnick &

Fidell, 2013). The Kaiser-Meyer-Olkin test of Sampling Adequacy (Kaiser, 1974) was calculated to account for any effects from all of the other variables on the pairwise correlations. A minimum KMO value of .60 or above was used as the recommended criteria for conducting factor analyses (Tabachnick & Fidell, 2013). The KMO for these data was .643, indicating the variables are factorable. The Bartlett's Test of Sphericity was also significant (p = .000), which confirmed there was some capacity with these data to reduce variables into factors.

Normality

Parametric statistical tests assume normal distribution of the data (Field, 2013). In covariance-based CFA, multivariate nonnormality can influence results and bias goodness-of-fit test statistics (Kaplan, 2000; Satorra & Bentler, 1994). First, all observed variables were examined for a normal distribution of scores by checking the skewness and kurtosis values. SPSS uses 0 to indicate normal kurtosis, thus anything greater than 0 was considered as excess kurtosis (Field, 2013). The researcher used the general rule of thumb for skewness and kurtosis within an absolute value range of +/- 2 as considered normal (Field, 2013; Hahs-Vaughn & Lomax, 2012). Table 7 shows the skewness and kurtosis values of the 51 observed variables. Five variables were highly negatively skewed: Curric3a (-2.11), Instruc5j (-2.26), ProfPrac6b (2.30), Collab7b (-2.21) and Collab7c (-2.72). Kurtosis values ranged between -1.732 and 9.334 (see Table 7). Twenty-four of the observed variables were highly leptokurtic with values > 2. The leptokurtic distributions indicated there was not a lot of variance in responses with observers rating those indicators very similarly. A visual inspection of histograms for the observed variables verified the shape of the distributions were not normal with these data.

Next, the researcher further checked the data distribution using the Kolmogorov-Smirnov (K-S) and Shapiro-Wilk (S-W) statistical tests of normality. The K-S and S-W tests of normality examine whether scores deviate from a normal distribution (Field, 2013) and may be used with small to medium samples ($N < 300$; West et al., 1996). A significant K-S test ($p < .05$) means the distribution of scores is significantly different from normal. The K-S test was highly significant ($p = .000$) for all variables, indicating nonnormal distribution of these data.

Linearity

Confirmatory factor analysis and Pearson correlations assume linear relationships among variables (Lomax & Hahs-Vaughn, 2012; Tabachnick & Fidell, 2013). The researcher inspected bivariate scatterplots for linearity among pairs of variables. Linearity assumes a straight-line relationship between pairs of variables, meaning an increase or decrease in one variable leads to either an increase or decrease in the other variable (Tabachnick & Fidell, 2013). When data are skewed, meaning they are not linear or are curvilinear, "the mean is not a good indicator of the central tendency of the scores in the distribution" (Tabachnick & Fidell, 2013, p. 87). The bivariate scatterplots and skewness values suggested multiple pairs of variables had nonlinear relationships. The researcher chose not to transform the data because it was reasonable to expect variables to be skewed in this population. Previous research suggests observers conducting classroom observation measures often rate teachers similarly with higher effectiveness ratings (Kane & Staiger, 2012; Lash, Tran, & Huang, 2016). Thus, the researcher made provisions in the analysis to take the nonnormality of the data into account by using the Bollen-Stine bootstrapping method in AMOS.

Multicollinearity

The hypothesized QIASD reflective measurement model for the CFA assumed the observed variable scores were caused by the latent variables (factors). The reflective measurement model in this study assumed observed variables grouped under a factor had moderately high correlations with each other and low correlations with observed variables not grouped in that factor (Kock, 2015). Problems with multicollinearity may occur when variables are extremely highly correlated (Tabachnick & Fidell, 2013). The R-matrix was inspected for Pearson correlation values above .08, which would signal variables were too highly correlated and potentially redundant (Field, 2013). No variables on the R-matrix were above .80 indicating no multicollinearity issues with these data.

Research Question Analysis Results

The QIASD was designed to rate the educational quality of special education classrooms serving students with ASD. The QIASD instrument consists of 51 items (quality indicators) grouped under seven professional practice standards (Pearl et al., 2017). The intended use of the QIASD is to inform teachers of the presence of specific quality indicators supported by evidence-based practices and professional practice standards (Pearl et al., 2017). The hypothetical basis for the QIASD was special education classrooms with high levels of these quality indicators possessed the teaching practices and educational components deemed necessary for successfully teaching students with ASD. The purpose of this study was to assess

the internal consistency reliability and construct validity of the QIASD scores in measuring special education teaching effectiveness of students with ASD.

Internal Consistency Reliability

Research Question 1a: To what extent does the QIASD produce reliable scores as measured by internal consistency reliability?

The reflective measurement model for the QIASD was examined for internal consistency reliability with these data (N = 102). The researcher calculated Cronbach's alpha (Cronbach, 1951) with SPSS version 23 to measure the correlations between individual indicator scores, subsets, and the overall scale. Covariance is a measure of bivariate correlations that indicates how much two variables vary together in a linear association (Field, 2013). The coefficient alpha ranges from 0 to 1, with higher numbers approaching 1 meaning more items have shared covariance and likely measure the same underlying construct (Field, 2013). A threshold value of .70 was used to indicate good internal consistency reliability (Kline, 2016).

The Cronbach's alpha coefficient for the overall QIASD scale (51 items) was α = .913, which indicated high overall internal consistency reliability with these data (Kline, 2016). Researchers suggest when a scale has several factors, then the Cronbach's alpha formula should be applied separately to each subgroup of variables (Cronbach, 1951; Field, 2013; Osborne, 2013). The QIASD has seven subgroups (standards) delineated within the measure, so the Cronbach's alpha was run for each set of observed variables specified in those subgroups. The Cronbach's alpha scores for the seven subgroups are reported in Table 9.

Table 9

Cronbach's alpha for Subgroups

Factor	Cronbach's alpha	N of items	Cronbach's alpha if Item Deleted
Learner Development and Individual Learning Differences	.732	6	Range .616 to .727 No improvement if variable removed
Learning Environments	**.683**	11	Range .611 to .751 Improve to .751 if variable removed (Environ2k)
Curricular Content Knowledge	.781	7	Range .727 to .787 Improve to .787 if variable removed (Curric3b)
Assessment	.787	4	Range .686 to .760 No improvement
Instructional Planning and Strategies	.744	13	Range .702 to .768 Improve to .768 if variable removed (Instruc5l)
Professional Learning and Practice	**.362**	3	Range .183 to .302 No improvement
Collaboration	**.680**	7	Range .579 to .691 Improve to .691 if variable removed (Collab7b)

Four of the QIASD subgroups (factors) had Cronbach's alpha values above .70: Learner Development and Individual Learning Differences, Curricular Content Knowledge, Assessment, and Instructional Planning and Strategies. The Cronbach's alpha scores for Learning Environments and Collaboration were slightly low, at .683 and .680 respectively. An examination of the Cronbach's alpha if Item Deleted suggested the Cronbach's alpha of the Learning Environments subgroup would improve from .683 to .751 if variable Environ2k was removed. The Cronbach's alpha for Collaboration would improve from .680 to .691 if variable Collab7b was removed. Lastly, Professional Learning and Practice had an extremely low

Cronbach's alpha value of .362, and no improvements would be made through variable deletion, which suggested poor internal consistency of this subgroup.

Construct Validity Results

Research Question 1b: To what extent is construct validity of the QIASD scores achieved as measured by a confirmatory factor analysis?

A confirmatory factor analysis (CFA) was chosen as the main method of statistical analysis to establish construct validity evidence by verifying the internal structure of the QIASD model with these data (AERA et al., 2014; Kline, 2016). The researcher used CFA to test the hypothesized relationship between the observed variables and their underlying latent variables (factors). The researcher provided empirical and theoretical evidence in the literature review for the CFA model to support a reflective relationship between the 51 quality indicators and the seven factors aligned with the CEC professional practice standards (CEC, 2014; Pearl et al., 2017; Wong et al., 2015). This reflective model assumed the observed variables grouped under each latent variable were internally consistent, meaning they had positive, moderately high inter-correlations. Observed variables of the same latent variable were presumed to measure a common construct and could potentially be removed or interchanged without affecting the construct (Kline, 2015). The following section describes the results of the data analysis and the adjustments made based on the results of these data.

Covariance-based Model CFA Results

The covariance-based CFA (Joreskog, 1978) is the most popular method of CFA (Hair et al., 2014). The researcher used SPSS AMOS version 23 to conduct a covariance-based CFA analysis to determine how well the QIASD model estimated a covariance matrix for this data sample (Hair et al., 2014; Kline, 2015). A reflective measurement model for the QIASD was identified with a causal direction from the latent variables to the respective observed variables. It was theorized effective teaching practices for students with ASD would predict the classroom quality indicator rating scores. This hypothesized QIASD reflective model was specified with 51 observed variables (indicators) loading onto seven latent variables (factors), with no specified directional relationship between the latent variables (see Chapter 3, Figure 2).

The researcher had access to the AMOS software and planned to use the maximum likelihood estimator (ML) to examine factor loadings and the model fit to these data. However, the analyses of statistical assumptions showed nonlinearity and nonnormal data distributions. In addition, the sample size was smaller than anticipated (N = 102). This limitation could emphasize the effects of multivariate nonnormality (Lei & Lomax, 2005). The use of regular ML estimation with multivariate nonnormal data can lead to biased conclusions about parameter estimates and model adequacy (Bentler & Yuan, 1999). The ML statistic may deviate significantly from the chi-square distribution when used with small samples and nonnormal data (Hu, Bentler & Kano, 1992). Best practice recommends robust estimation methods with nonnormal data in order to produce corrected chi-square values (Field, 2013; Yuan & Bentler, 1999). However, the robust estimation methods, such as maximum likelihood robust (MLR) or Satorra-Bentler, were not available in AMOS (Arbuckle, 2008). Instead, the ML estimation with Bollen-Stine

bootstrapping in AMOS was used to estimate model fit in place of the traditional chi-square statistic (Bollen-Stine, 1993). Bootstrap samples were set to 250 (Nevitt & Hancock, 2001) and the cutoff significance level used was $p > .05$ to indicate model fit.

The covariance-based CFA analysis revealed a poor model fit for the proposed reflective measurement model (see Figure 3). The observed model chi-square was $X^2 = 2481.2$, $df = 1224$, $p = .000$, indicated poor fit to the model. The Bollen-Stine bootstrapped sample produced a p value of .263, which was non-significant at alpha level .05 and indicated the bootstrapped sample had a better model fit to these data than the observed sample. However, other goodness-of-fit indices for the observed data, the RMSEA (.101), the TLI (.374), the GIF (.535), and the AGFI (.496) did not meet acceptable thresholds for good model fit.

The factor loadings are reported in Table 10. Factor loadings were analyzed at an alpha of .05 and a factor loading threshold of .55 to identify whether the observed variables loaded significantly on the factors. Hair et al. (1998) recommended factor loadings with an absolute value of .55 for statistical significance in sample sizes of 100. The results from these data showed 27 variables with nonsignificant factor loadings less than .55.

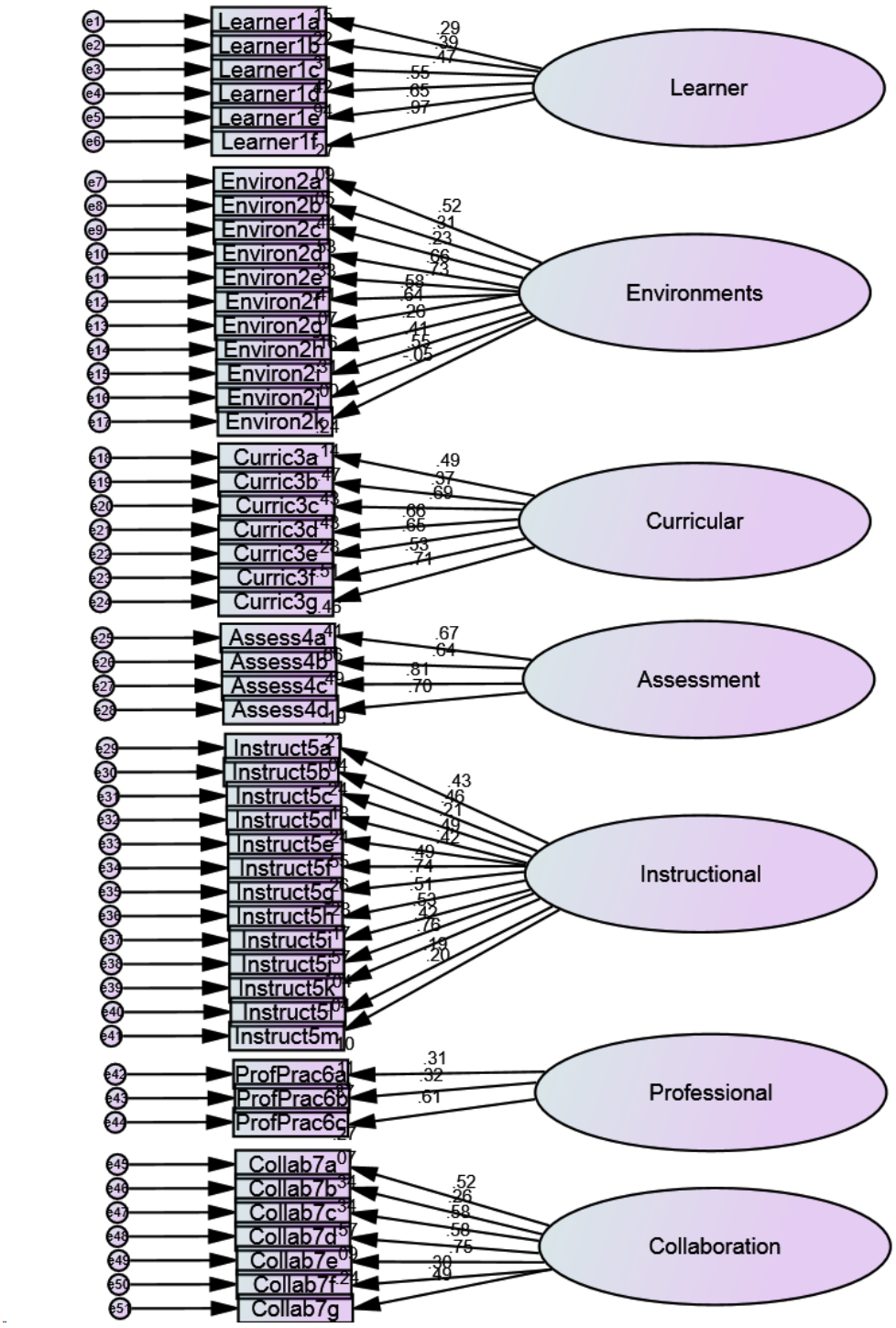


Figure 3: Covariance-based CFA Model with factor loadings

Variables with loadings < .55 are bolded in Table 10. The Learner factor had three variables with loadings < .55; Environ had five variable loadings < .55; Curric had two variable loadings < .55; Assess had all significant variable loadings > .55; Instruct had eleven variables load < .55; ProfPrac had two variables load < .55; and Collab had 4 variables load < .55.

Next, the squared multiple correlations, which are the communality estimates, were inspected to determine how much variance in the indicator variables were accounted for by the latent factors (see Table 10). For example, the Learner factor only accounted for about 9% of the variance in observed indicator Learner1a, but accounted for about 94% of variance in Learner 1f. When communalities are lower than .40, the observed variable may not load significantly onto a factor (Arbuckle, 2008). The output showed 35 out of the 51 observed variables had squared multiple correlations lower than .40, indicating 69% of the observed variables were unlikely to significantly load onto their respective latent variables with these data.

Collectively, the goodness-of-fit indices, factor loadings, and communality estimates suggested the hypothesized QIASD seven-factor reflective measurement model was not a good fit to these data.

Table 10

Covariance-based CFA Factor Loadings and Communalities

Factor	Observed Variable	Factor Loading	Communalities (R^2)
Learner: Learner Development and Individual Learning Differences	1a. Instruction is individualized and based on learner characteristics, interests, and ongoing assessment.	**0.293**	**0.086**
	1b. Schedules reflect a variety of learning formats, including 1:1 instruction, small group, large group, independent work, and social interaction/leisure options.	**0.389**	**0.151**
	1c. Instruction incorporates natural and individualized reinforcers.	**0.467**	**0.218**
	1d. Students with slow rates of learning are provided intensive levels of instruction, including daily one-on-one instruction sessions.	0.553	**0.306**
	1e. All adults have knowledge/access to IEP objectives being worked on for each student.	0.646	0.417
	1f. IEP goals and objectives are embedded within daily activities and routines throughout the day to promote maintenance and generalization.	0.969	0.939
Environ: Learning Environments	2a. Room arrangement has clearly defined visual boundaries for specific activities.	0.523	**0.274**
	2b. Room arrangement allows for supervision of all students at all times; and prevents or minimizes problem behaviors.	**0.305**	**0.093**
	2c. Staff ratio of 1 adult for every 3 students is maintained during (at least 75%) observation. Allow greater ratio if students included part of the day and not on access points.	**0.233**	**0.054**
	2d. A daily classroom schedule is posted at student level, is visible and appropriate for students' level of symbolic functioning, and is used throughout the day. Schedule indicates what activity is current.	0.66	0.436
	2e. Individual schedules are posted at child level and are being used correctly. Schedule is referred to for each activity, sequence of activities is adhered to unless change is noted. Student is engaged in using schedule.	0.728	0.53
	2f. Transitions are supported by routines, environmental arrangement and scheduling.	0.578	**0.334**

Factor	Observed Variable	Factor Loading	Communalities (R^2)
	2g. Visual supports are at the correct level of symbolic functioning, and are used to enhance predictability, facilitate transitions, and help convey expectations.	0.638	0.407
	2h. Instructional materials and furniture are age appropriate.	**0.26**	**0.068**
	2i. Classroom materials are well organized (i.e. labeled, conveniently located, and stored when not in use).	**0.406**	**0.165**
	2j. Individual workstations, when present, are arranged left-right or top-bottom, and tell how much work, what work, when finished, and what's next. Workstation materials are varied from day to day and are educationally/functionally relevant.	0.552	**0.305**
	2k. The teacher can provide examples of opportunities for meaningful interaction and friendships with peers without disabilities.	**-0.051**	**0.003**
Curric: Curricular Content Knowledge	3a. Schedule and activities reflect distribution of curriculum across multiple domains appropriate for the age, level and individual needs of students in classroom.	**0.495**	**0.245**
	3b. Curriculum/activities address and are aligned with appropriate grade level general education curriculum and standards.	**0.371**	**0.138**
	3c. Curriculum/activities address social communication skills (i.e. pragmatics, conversation, perspective taking) with adults and peers	0.687	0.472
	3d. Curriculum/activities address functional communication for all students	0.658	0.433
	3e. Curriculum/activities address functional life skills and adaptive behavior to maximize independent functioning in school, home, vocational, and community settings.	0.655	0.429
	3f. Specialized instruction to enhance social participation across environments is provided. If social skills instruction is infused, there is evidence of planning and evaluation.	0.528	**0.278**
	3g. Curriculum/activities address self-regulation and self-monitoring.	0.712	0.507
Assess: Assessment	4a. Written data are gathered consistently and frequently (daily or weekly) to track progress on IEP goals and objectives.	0.675	0.455
	4b. Assessment tools and methods are selected, adapted and used to accommodate the abilities and needs of individuals with developmental disabilities/autism spectrum disorders.	0.642	0.412

Factor	Observed Variable	Factor Loading	Communalities (R^2)
	4c. Data are collected for monitoring and analyzing challenging behavior and its communicative intent.	0.81	0.655
	4d. Students displaying behavioral difficulties have an individualized behavior plan being implemented or have been referred for Functional Behavior Assessment (FBA).	0.698	0.488
Insruct: Instructional Planning and Strategies	5a. Instruction is systematic and based on learner characteristics, interests, and ongoing assessment.	**0.433**	**0.187**
	5b. Students remain actively engaged in learning opportunities throughout observation, with no more than 2 minutes down time.	**0.46**	**0.212**
	5c. During five minute observation, staff interacts with each student at least once to teach or promote learning. Excluding students who are engaged in independent work.	**0.209**	**0.044**
	5d. Instructional pace promotes high rates of correct responding, correct responses are reinforced or prompting/error correction is provided as needed.	**0.488**	**0.238**
	5e. Skills are taught in the context of naturally occurring activities and daily routines. There is no down time for teaching.	**0.422**	**0.178**
	5f. Communication directed to students is clear, relevant, appropriate to language ability, and grammatically correct.	**0.486**	**0.236**
	5g. Communication directed to students presents opportunities for dialogue (rather than being largely directive).	0.74	0.548
	5h. Communication directed to students consists of largely instructive/positive comments in comparison to corrective comments.	**0.513**	**0.264**
	5i. Behavior problems are minimized by using proactive strategies including choices, clear expectations and positive reinforcement.	**0.526**	**0.277**
	5j. Instructional methods are grounded in evidence-based practices.	**0.418**	**0.174**
	5k. Staff create opportunities for spontaneous use of communication skills including student-to-student interactions.	0.756	0.571
	5l. Students without verbal communication have AAC and actively use across activities.	**0.192**	**0.037**
	5m. Technologies are employed to support instructional assessment, planning, and delivery for individuals with exceptionalities.	**0.202**	**0.041**

Factor	Observed Variable	Factor Loading	Communalities (R^2)
ProfPrac: Professional Learning and Practice	6a. "Hands-on" contact with students promotes independence and preserves dignity.	**0.313**	**0.098**
	6b. Inter-staff communication is respectful of students and limited in content to classroom issues and instruction. Confidentiality of students is preserved.	**0.325**	**0.105**
	6c. Restrictive procedures employed are supported by a Functional Behavior Assessment and Behavior Intervention Plan.	0.612	**0.375**
Collab: Collaboration	7a. A staff schedule showing staff and student assignments, locations, and activities, is prominently posted and being followed.	**0.517**	**0.267**
	7b. All classroom staff is involved in delivering instruction, including during out-of-classroom activities (lunch, recess, CBI).	**0.258**	**0.066**
	7c. There is a consistent system in place for regular (daily/weekly), informative and positive communication with families regarding student participation, progress and concerns.	0.582	**0.339**
	7d. Two-way communication is encouraged by soliciting information and questions from families.	0.582	**0.339**
	7e. A variety of opportunities for family involvement are provided (classroom activities, information sharing, and parent training).	0.754	0.568
	7f. Teacher collaborates with team members to plan transition to adulthood that encourages full community participation.	**0.303**	**0.092**
	7g. Teacher collaborates with school personnel and community members in integrating students with ASD in various settings.	**0.492**	**0.242**

Exploratory Factor Analysis Results

A follow-up exploratory factor analysis (EFA) in IBM SPSS version 23 was completed to further examine the problems with variable loadings discovered in the covariance-based CFA. Principal axis factors was used as the method of factor extraction given the assumption of normality is violated (Costello & Osborne, 2005).

In EFA, rotation is used to help interpret the data by maximizing the highly correlated variables and minimizing those with low correlations (Tabachnick & Fidell, 2013). After rotation, the researcher can see the factor structure that has the best fit to the data. There are two types of rotation methods. Orthogonal rotation methods do not allow factors to correlate, while oblique methods allow the factors to correlate (Costello & Osborne, 2005). Although no directional relationships between factors were specified, some correlation may be expected due to the nature of the factors all theoretically measuring the same construct. The researcher used the Promax oblique rotation method, which allows latent variables to correlate (Tabachnick & Fidell, 2013).

The Kaiser-Meyer-Olkin measure was checked to verify sampling adequacy (Tabachnick & Fidell, 2013). When a sample size is too small, the correlations may not stabilize and could influence the validity of the factor analysis (Floyd & Widaman, 1995). Kaiser (1974) suggested KMO values at .5 were barely acceptable. The researcher used Hutcheson and Sofroniou's (1999) guidelines on acceptable KMO values, with values below .5 as unacceptable, values of .6 as "mediocre", and values above .7 as acceptable (Field, 2013, p. 685).

These data produced a KMO value of .662 that may be considered barley acceptable for factorability. The KMO indicated about 66% of the observed variables can be explained by some factors. The Bartlett's Test of Sphericity significant at .000, suggesting some interrelated variables could group under factors, and thus rejecting the null hypothesis that there is no correlation among the 51 observed variables (Field, 2013).

Kaiser-Guttman criterion was used to identify factors with eigenvalues greater than 1.00 (Floyd & Widaman, 1995; Kline, 2015). The results of the EFA are displayed in Table 11. These findings indicated the 51 observed variables loaded onto 16 factors, which explained 62.04% of the variance in the model. Visual examination of the scree plot (see Figure 4) verified observed variable loadings onto 16 factors.

Table 11

EFA Total Variance Explained

Factor	Total Eigenvalue	Cumulative % Sums of Squared Loadings
1	10.583	20.062
2	3.478	26.201
3	2.893	31.134
4	2.612	35.567
5	2.232	39.249
6	2.037	42.457
7	1.845	45.388
8	1.758	48.074
9	1.626	50.515
10	1.553	52.793
11	1.330	54.628
12	1.265	56.385
13	1.215	58.014
14	1.153	59.546
15	1.047	60.835
16	**1.022**	**62.042**
17	.906	
18	.895	

Factor	Total Eigenvalue	Cumulative % Sums of Squared Loadings
19	.819	
20	.771	
21	.737	
22	.695	
23	.665	
24	.629	
25	.575	
26	.544	
27	.516	
28	.490	
29	.453	
30	.428	
31	.398	
32	.380	
33	.338	
34	.319	
35	.303	
36	.276	
37	.265	
38	.252	
39	.234	
40	.209	
41	.195	
42	.171	
43	.156	
44	.147	
45	.108	
46	.106	
47	.099	
48	.078	
49	.074	
50	.070	
51	.048	

Note. Extraction Method: Principal Axis Factoring.

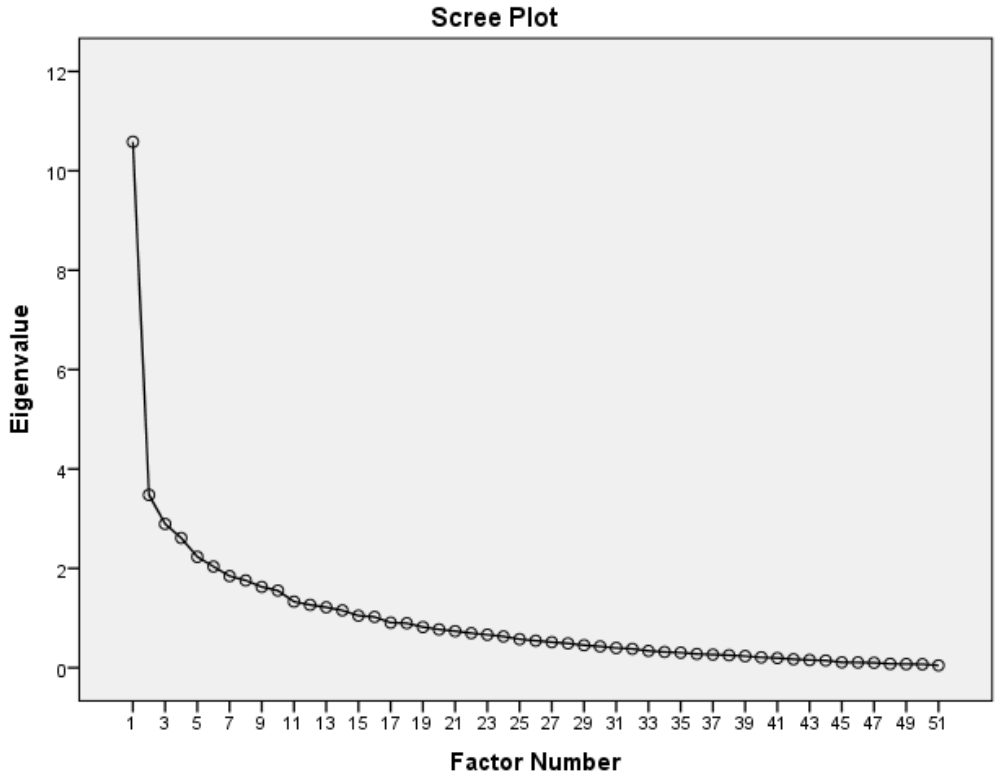


Figure 4: EFA Scree Plot

The initial EFA did not produce a pattern or structure matrix for the sixteen factors because the rotation failed to converge in 25 iterations, so the researcher increased the Maximum Iterations for Convergence to 50 (Field, 2013). An inspection of the pattern matrix after Promax oblique rotation provided information on the amount of unique contribution an observed variable had to a factor (Field, 2013). The pattern matrix for these data showed sixteen factors. Four observed variables (Learner1b, Learner1c, Environ2a, and Environ2g) loaded highly onto more than one of those sixteen factors. The structure matrix, which takes into consideration the shared variance between factors (Field, 2013), indicated at least 20 observed variables loaded highly onto more than one of the sixteen factors.

The researcher ran a second EFA with a forced seven-factor extraction to represent the hypothesized QIASD model. A summary of the results is displayed in Table 12. Inspection of the results indicated less than 50% (42.90%) of the cumulative variance was explained by the seven-factor model. As indicated by factor loadings equal to or above .40, thirteen of the observed variables did not load highly onto any of the seven specified factors. The nine additional eigenvalues greater than 1.00 indicated these data fit sixteen factors better than the hypothesized seven factors.

Table 12

Summary of EFA Results for the QIASD

Observed Variable	Rotated Factor Loadings						
	1	2	3	4	5	6	7
Learner1a	0.089	0.178	**0.492**	0.108	-0.336	0.116	0.273
Learner1b	0.343	-0.17	0.347	0.018	-0.087	0.089	0.351
Learner1c	**0.4**	0.288	0.051	-0.022	0.235	-0.073	-0.135
Learner1d	0.095	0.022	0.054	-0.147	0.126	0.343	**0.537**
Learner1e	**0.477**	-0.012	-0.062	-0.118	0.109	**0.436**	-0.082
Learner1f	**0.531**	0.077	0.012	-0.17	0.124	0.391	0.115
Environ2a	**0.458**	-0.089	0.266	-0.067	-0.236	0.087	0.188
Environ2b	0.279	-0.176	0.278	0.095	-0.203	-0.227	0.195
Environ2c	-0.036	0.092	0.019	0.16	-0.099	-0.025	**0.525**
Environ2d	**0.793**	0.026	-0.049	0.055	-0.16	-0.042	-0.015
Environ2e	**0.617**	0.223	-0.025	0.003	-0.18	0.062	0.058
Environ2f	**0.407**	0.305	0.126	-0.27	0.194	-0.15	0.043
Environ2g	**0.599**	0.145	0.134	0.193	-0.098	-0.218	-0.197
Environ2h	0.15	-0.101	**0.413**	-0.218	0.23	-0.011	0.056
Environ2i	0.224	-0.03	**0.43**	0.014	-0.08	0.041	0.071
Environ2j	**0.43**	0.149	-0.133	0.338	-0.082	0.073	0.119
Environ2k	-0.269	0.04	-0.015	0.12	**0.584**	0.005	-0.071
Curric3a	0.383	-0.097	0.124	0.302	0.02	0.112	-0.13
Curric3b	-0.054	-0.103	-0.11	0.156	**0.401**	0.105	0.162
Curric3c	0.06	-0.094	0.167	**0.463**	0.238	-0.143	0.109
Curric3d	0.121	0.01	0.014	**0.525**	0.146	-0.255	0.143
Curric3e	0.159	0.169	0.138	**0.518**	0.124	-0.029	-0.087
Curric3f	0.072	-0.115	0.256	0.393	0.177	0.229	-0.095

Observed Variable	Rotated Factor Loadings						
	1	2	3	4	5	6	7
Curric3g	0.085	-0.241	0.11	**0.466**	0.353	-0.042	0.047
Assess4a	0.195	**0.65**	0.058	-0.066	-0.151	0.285	-0.211
Assess4b	-0.288	**0.777**	0.324	0.102	-0.134	0.077	0.152
Assess4c	0.046	**0.643**	0.02	0.041	0.068	0.12	0.081
Assess4d	0.061	**0.573**	-0.266	-0.023	0.149	0.059	0.184
Instruct5a	-0.008	-0.083	0.025	0.278	0.295	0.164	0.271
Instruct5b	-0.017	-0.097	**0.693**	0.072	0.002	-0.025	0.068
Instruct5c	-0.228	0.065	0.292	0.065	-0.059	0.077	**0.416**
Instruct5d	-0.164	0.148	**0.42**	-0.088	0.222	0.155	0.067
Instruct5e	0.156	0.149	0.295	0.076	0.1	0.173	-0.175
Instruct5f	0.026	0.249	**0.475**	-0.051	0.132	-0.005	0.079
Instruct5g	0.031	0.047	0.191	0.084	**0.709**	-0.111	-0.048
Instruct5h	-0.065	0.079	**0.501**	0.008	0.163	0.061	-0.057
Instruct5i	0.141	**0.436**	0.227	-0.124	0.284	-0.249	-0.013
Instruct5j	-0.012	-0.015	0.322	0.25	0.059	0.302	-0.053
Instruct5k	-0.175	0.027	0.178	0.277	**0.77**	-0.015	-0.026
Instruct5l	0.121	0.351	-0.209	0.376	-0.096	0.006	0.167
Instruct5m	-0.081	0.014	-0.062	**0.493**	0.012	-0.005	0.165
ProfPrac6a	0.25	0.306	-0.019	0.103	-0.076	-0.059	-0.06
ProfPrac6b	-0.03	0.013	0.112	0.168	0.082	-0.217	**0.455**
ProfPrac6c	0.103	0.305	-0.343	0.202	0.106	-0.081	0.361
Collab7a	**0.624**	0.009	-0.176	0.363	-0.131	0.194	-0.169
Collab7b	0.333	-0.075	0.126	0.051	0.018	0.096	-0.039
Collab7c	0.076	0.103	0.128	-0.027	-0.156	**0.674**	0.023
Collab7d	-0.098	**0.451**	0.113	0.075	0.031	**0.476**	-0.048
Collab7e	0.223	-0.016	-0.085	0.338	0.143	0.331	0.071
Collab7f	-0.038	0.28	-0.12	**0.475**	0.02	0	0.091
Collab7g	-0.149	0.047	-0.061	0.334	0.223	0.275	-0.017
Eigenvalues	10.583	3.478	2.893	2.612	2.232	2.037	1.845
% of variance	20.751	6.820	5.673	5.121	4.376	3.995	3.618
α	.732	.683	.781	.787	.744	.362	.680

Note: Factor loadings equal to or over .40 appear in bold.

Next, the forced seven-factor EFA pattern matrix showed eight observed variables loaded highly onto more than one factor. In addition, the pattern of observed variable loadings on factors did not fit the hypothesized QIASD model of variable relationships. For example, the forced seven-factor EFA resulted in Factor 1 containing observed variables Learner1c,

Learner1e, Learner1f, Environ2a, Environ2d, Environ2e, Environ2f, Environ2g, Environ2j, and Collab7a. The researcher could potentially have improved the pattern matrix for the seven-factor model by removing the 13 variables with low loadings (Field, 2013). However, this would not be practical because Pearl and colleagues (2017) identified all 51 observed variables as important items in the QIASD.

The factor correlation matrix (see Table 13) presents the amount of correlation between the seven specified factors. The factor correlations are all fairly low (< .40) and do not suggest a strong relationship between any factors (Field, 2013). These findings are consistent with the QIASD model specifications with no direct relationships hypothesized between latent variables.

Table 13

Factor Correlation Matrix

Factor	1	2	3	4	5	6	7
1	1.000						
2	.349	1.000					
3	.312	.133	1.000				
4	.273	.086	.223	1.000			
5	.385	.212	.244	.169	1.000		
6	.179	.097	-.018	.169	.166	1.000	
7	.373	.187	.125	.115	.237	.218	1.000

Note. Extraction Method: Principal Axis Factoring. Rotation Method: Promax with Kaiser Normalization.

The Reproduced Correlations matrices for the original EFA and the EFA with seven specified factors displayed the differences between the observed correlation coefficients and those predicted from the model (Field, 2013). A good model should have most values smaller than .05 (Field, 2013). The sixteen-factor model produced 99 (7%) residuals with absolute values greater than .05, and the seven-factor model produced 475 (37%). The large percent of

residuals greater than .05 in the forced seven-factor model provided further evidence the model was not a good fit with these data.

Alternative Model CFA Analysis

Researchers recommend comparing a preferred hypothesized model with alternative models when examining goodness-of-fit to the data (Kline, 2015). The researcher planned to compare the seven-factor model to two alternative factor-models determined a priori based on support from the literature (i.e, Crimmins, 2001). However, the inconsistent results from the covariance-based CFA and the EFA analyses suggested a problem with the dimensions of the hypothesized model. A CFA relies on data assumptions and appropriate model specification to obtain meaningful results (Kline, 2015). The low sample size, inconsistent correlations, and lack of model fit with these data, made it impractical to compare the preferred model with the a priori chosen alternative models (Kline, 2015; MacCallum, Widaman, Zhang, & Hong, 1999).

Alternative models also may be determined post hoc, after testing the factor loadings and examining fit indices for correlation between variables in the preferred model (Kline, 2015; Tabchnick & Fidell, 2007). The researcher analyzed a new alternative model in response to the initial data analysis results and a reconceptualization of the theoretical structure of the QIASD measurement model.

Basis for Formative Model

The original reflective model of the QIASD in this study assumed the groups of quality indicators were defined by the theoretical concept of effective teaching represented by the seven

standards. The observed variables were presumed to be highly inter-correlated as "conceptually similar dimensions of their corresponding reflective latent variables" (Bollen & Bauldry, 2011, p. 270). The poor factor loading scores in the previous results led to reexamination of the relationship between variables in the QIASD model. The 51 quality indicators on the test were previously deemed important and have been validated for content to align with professional teaching practice standards (Pearl et al., 2017). Rather than removing items for the purpose of obtaining a better reflective model fit, an alternative model conceived as formative rather than reflective was explored.

A measurement model may be considered formative if the observed variables predict the latent variables and if removing an observed variable would change what the latent variable is measuring (Diamantopoulos & Winklhofer, 2001; Jarvis et al. 2003). Groups of formative observed variables may or may not be conceptually similar and as independent variables lend weight to the latent variable (Bollen & Bauldry, 2011; Fornell & Bookstein, 1982). A composite scale assumes formative latent variables are exact linear combinations of the observed variables and implies the composite error variance is zero (Bollen & Bauldry, 2011; Sarstedt et al., 2016).

The formative measurement model should have some theoretical and/or empirical basis (Kock, 2014; Sarstedt et al., 2016). The QIASD was designed to measure the theoretical construct of effective teaching practices as defined by alignment with seven professional practice standards (CEC, 2014) and 51 quality indicators stemming from empirical research in the field of special education for students with ASD (Pearl et al., 2017). The selected quality indicators may be conceived as separate components of the standards and drawn from unique perspectives (CEC, 2014; NCR, 2001; Wong et al., 2015). For example, Learning Environment indicator 2c

measures the "staff ratio" in the classroom, and indicator 2e measures whether "individual schedules are posted at child level and are being used correctly." Both of these quality indicators relate to the learning environment, yet they can be viewed as formative observed variables because they measure different features of that environment. The empirical literature offered sufficient evidence to conceptualize the QIASD with a formative measurement model structure.

Justification for Analysis

The tests of statistical assumptions showed these data had nonlinear and nonnormal distributions and the sample size (N = 102) was smaller than desired to obtain a power of .08. Traditional covariance-based CFA software programs and analyses rely on these assumptions to produce valid results (Kline, 2015). Widely used ML-based model fit indices are not robust to violations of statistical assumptions (Kline, 2015; Tabachnick & Fidell, 2013). With these data, the use of a nonparametric statistical tool was supported. A nonparametric technique for analyzing a theorized model is partial least squares (PLS; Kock, 2015). PLS algorithms calculate approximate latent variable scores through composites, not factors (Hair et al., 2014). Composite scores represent the theoretical latent constructs as formed by the sets of observed variables (Kock, 2014; Sarstedt, Ringle, Smith, Reams, & Hair Jr., 2014). The PLS-based CFA method is a more suitable analysis than covariance-based CFA when the variables are formative and the statistical assumptions are violated (Hair et al., 2014; Sarstedt et al., 2014).

WarpPLS version 5.0 was used to examine the alternative QIASD model (see Figure 5) due to the software's ability to handle formative variables and data that deviate from normal (Kock, 2015). The PLS algorithm generates composites based on linear combinations of

observed variables (Wold, 1980). The WarpPLS default method for calculating *p* values and related coefficients is Stable3, a resampling method similar to bootstrapping. The Stable3 method was "specifically aimed at increasing accuracy and statistical power" that was useful for the smaller sample size in this study (Kock, 2015, p. 10). In a Monte Carlo simulation (Kock, 2014), the standard errors estimated with the Stable3 method in WarpPLS were more accurate than with bootstrapping and lead to greater statistical power with small sample sizes.

Additional justification for using WarpPLS with these data was evident from the results of the unimodality and normality tests applied to the latent variables. An outcome of "No" on these tests indicates the latent variable distributions are not multivariate unimodal and not normal. Kock (2015) recommended if at least one latent variable resulted with no unimodality or normality, "the nonparametric methods used in this software are particularly appropriate" (p. 68). In this analysis, the results showed "No" on five out of the seven latent variables, signifying the WarpPLS nonparametric tests used for this CFA were appropriate.

PLS CFA Results

Kock's (2015) criteria was used for evaluating the formative model fit indices with the PLS CFA analysis. Only one classic fit statistic, the AFVIF, was provided in the results due to the type of analysis used (PLS algorithm) and the simple model specification between observed variables and latent variables, with no specified relationships between latent variables and no overall construct variable. The classic average full collinearity variance inflation factor (AFVIF) reflects the amount of multicollinearity between latent variables. High AFVIF values may indicate redundant latent variables that measure the same underlying construct. An acceptable

AFVIF should have a value equal to or less than 3.3. The alternative fit statistics employed were the standardized root mean square residual (SRMR), acceptable if less than or equal to .1, and the standardized mean absolute residual (SMAR), acceptable if less than .1.

The results of the PLS CFA indicated a better model fit with the formative measurement model to these data than the original reflective measurement model. The AFVIF = 1.833 (acceptable if < = 5, ideally < = 3.3), the SRMR = 0.139 (acceptable if < = 0.1), and the SMAR = 0.111 (acceptable if < = 0.1).

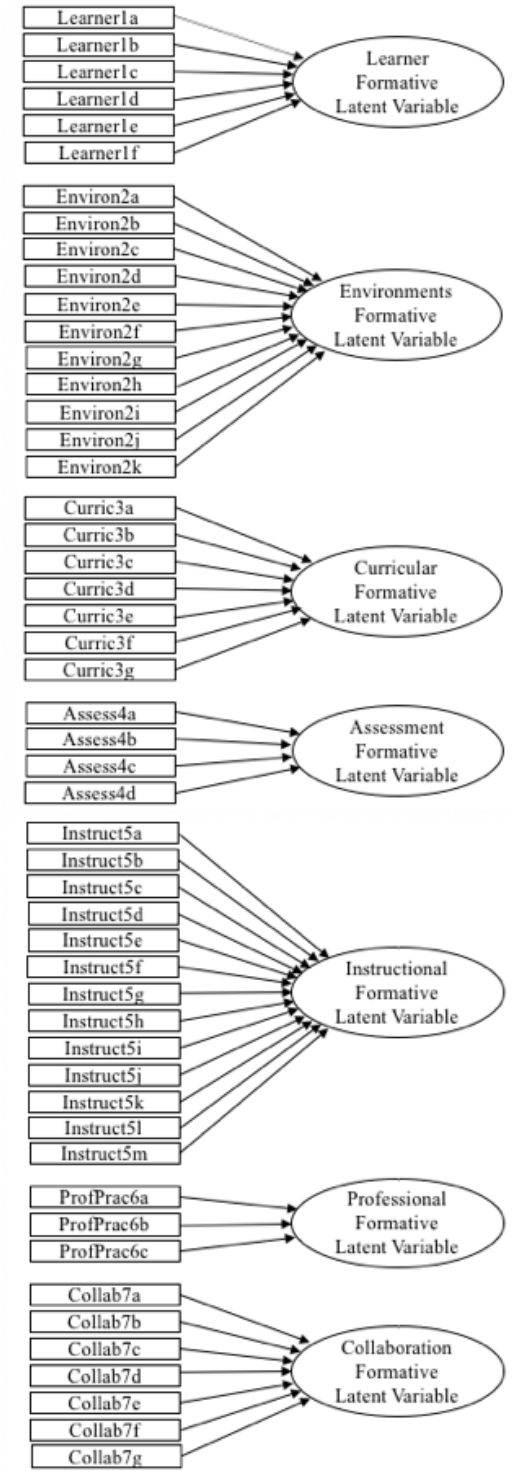


Figure 5: QIASD Formative Measurement Model

Kock and Mayfield (2015) recommend inspecting the observed variable weights and variance inflation factors (VIFs) to validate the specification of variables as formative. The ideal *p*-value for weights is below .05, which would indicate the observed variables were significantly associated with scores of their respective latent variables (Kock, 2015). Statistically non-significant weights (> .05), may signal issues with collinearity. High collinearity of observed variables assumed to measure different facets of a formative latent variable may suggest those observed variables actually measure the same thing. In formative models, the observed variables are expected to measure difference components of the latent variable, and thus should not be redundant. The researcher used a VIF value of below 3.3 to signify observed variables that were not redundant (Kock & Mayfield, 2015; Petter et al., 2007).

The results for observed variable weights, *p*-values, and VIFs are displayed in Table 15. Thirteen of the observed variable weights (noted in bold on Table 14) had *p*-values > .05, which indicated high collinearity of those variables. All six observed variables in the *Learner* latent variable were significant at = < .05. Four of the eleven *Environ* variables (2b, 2c, 2h, and 2k) were nonsignificant at > .05. One of the seven *Curric* variables (3b) was > .05. All four of the *Assess* variables were < .05. Seven of the 13 *Instruct* variables (5a, 5b, 5c, 5e, 5j, 5l and 5m) were > .05. All three *ProfPrac* variables, were < .05. And, one of the seven *Collab* variables (7b) was > .05.

The VIF values for the 51 observed variables were all below the recommended threshold of 3.3 for not being redundant. In terms of evidence for defining the QIASD variables as formative, the low VIF values offered good support for the variables as formative. Furthermore,

the majority of observed variable weights had significant p values, though some multicollinearity concerns exist for thirteen of the observed variables with nonsignificant p values.

Table 14

PLS CFA Formative Model Results

	Observed Variable Weights									
	Learner	Environ	Curric	Assess	Instruct	ProfPrac	Collab	*p*	VIF	ES
1a	(0.195)							0.020	1.220	0.102
1b	(0.244)							0.005	1.429	0.159
1c	(0.219)							0.010	1.312	0.128
1d	(0.258)							0.003	1.602	0.178
1e	(0.244)							0.005	1.692	0.159
1f	(0.320)							<0.001	2.587	0.274
2a		(0.191)						0.023	1.648	0.121
2b		(0.125)						**0.097**	1.351	0.052
2c		(0.086)						**0.187**	1.108	0.025
2d		(0.207)						0.015	1.703	0.143
2e		(0.222)						0.010	2.003	0.165
2f		(0.192)						0.022	1.559	0.123
2g		(0.206)						0.015	1.641	0.141
2h		(0.103)						**0.143**	1.171	0.035
2i		(0.159)						0.049	1.445	0.084
2j		(0.181)						0.029	1.570	0.109
2k		(-0.012)						**0.451**	1.145	**0.000**
3a			(0.194)					0.021	1.390	0.117
3b			(0.149)					**0.060**	1.251	0.069
3c			(0.238)					0.006	1.666	0.176
3d			(0.228)					0.008	1.591	0.161
3e			(0.233)					0.007	1.565	0.169
3f			(0.194)					0.021	1.388	0.116
3g			(0.248)					0.004	1.727	0.191
4a				(0.316)				<0.001	1.621	0.249
4b				(0.304)				<0.001	1.560	0.231
4c				(0.336)				<0.001	1.926	0.282
4d				(0.309)				<0.001	1.658	0.238
5a					(0.129)			**0.090**	1.347	0.063
5b					(0.149)			**0.061**	1.410	0.083
5c					(0.082)			**0.199**	1.194	0.025
5d					(0.160)			0.047	1.431	0.096
5e					(0.140)			**0.073**	1.372	0.073
5f					(0.162)			0.046	1.538	0.098
5g					(0.186)			0.026	2.424	0.129
5h					(0.161)			0.047	1.559	0.097

	Observed Variable Weights									
	Learner	Environ	Curric	Assess	Instruct	ProfPrac	Collab	*p*	VIF	ES
5i					(0.156)			0.052	1.642	0.091
5j					(0.143)			**0.069**	1.423	0.076
5k					(0.191)			0.022	2.489	0.137
5l					(0.066)			**0.248**	1.205	**0.016**
5m					(0.068)			**0.244**	1.309	**0.017**
6a						(0.467)		<0.001	1.043	0.290
6b						(0.476)		<0.001	1.046	0.301
6c						(0.553)		<0.001	1.074	0.408
7a							(0.239)	0.006	1.404	0.146
7b							(0.131)	**0.087**	1.133	0.044
7c							(0.270)	0.002	1.646	0.186
7d							(0.261)	0.003	1.682	0.174
7e							(0.307)	<0.001	1.720	0.241
7f							(0.165)	0.042	1.151	0.070
7g							(0.233)	0.007	1.322	0.139

Note. Desirable for formative indicators: $p = < 0.05$, VIF < 3.3, and Effect Size (ES) = > .02.

The researcher checked for positive weight-loading signs (WLS) for the observed variables of all latent variables. A negative WLS (-1) would suggest an instance of Simpson's paradox, meaning the hypothesized relationship between the observed variable and the latent variable is unlikely or reversed (Kock & Mayfield, 2015; Pearl, 2009; Wagner, 1982). The WLS for all observed variables in these data were positive (+1).

A recommended effect size of .02 (small), .15 (medium), and .35 (large) for the path coefficients (Cohen, 1988) was used to evaluate the level of significance of observed variable weights on the latent variables. Effect sizes were inspected for minimum values equal or greater than .02 (see Table 15). Three observed variables had effect sizes below .02: *Learner2k* (.000), *Instruct51* (.016), and *Instruct5m* (.017), indicating these observed variables did not weigh significantly on their respective latent variables.

Results support the QIASD formative model had better fit to the data than the hypothesized reflective model. Further research is needed to identify an acceptable model. The

findings shed light on the importance and the challenges of developing high-quality teacher evaluations and specifically on the iterative process of improving and checking the internal structure of the QIASD measure. The implications of these results are discussed in the next chapter.

CHAPTER FIVE: DISCUSSION

This study contributes to the existing research on examining the psychometric properties of teaching performance measures designed for special education teachers in classrooms serving students with autism spectrum disorder. Chapter 5 provides a review of the study including the purpose, research methodology used, and a discussion of the results from the data analysis. This chapter examines the resulting implications for the field of special education, including relevance to data analysis, instrument development, practitioners, and researchers. Additionally, the chapter reflects on the limitations of the study and recommends future research related to the development, use, and validity of the QIASD measure.

Review of Problem and Purpose

Autism spectrum disorder (ASD) affects more than 450,000 students aged 6 to 21 across the nation. These students receive special education services, with over 140,000 educated outside the general education classroom for 60% or more of the school day (DOE, OSERS, OSEP, 2016). Many students with more severe symptoms of ASD are taught by special education teachers within separate or self-contained classrooms (Hart & Whalon, 2011; White, Keonig, & Scahill, 2007). Yet many special educators lack sufficient preparation and support to implement evidence-based practices for students with ASD (Belfiore, Fritts, & Herman, 2008; Brock et al., 2014; Jennett, Harris, & Mesibov, 2003; Morrier, Hess, & Heflin, 2011; NCR, 2001). The unique learning profiles of students with ASD requires specialized and individualized instructional strategies and supports for meaningful education to occur (Anderson et al., 2014; National

Research Council, 2001; Spencer, Evmenova, Boon, & Hayes-Harris, 2014; Spooner, Knight, Browder, & Smith, 2012). Numerous evidence-based practices (Odom, et al., 2010; Simpson, 2005; Wong et al., 2015) and core components of effective teaching for students with ASD are recognized in the literature (e.g., Iovanonne et al., 2003; NRC, 2001). Despite this wealth of evidence on best practices for effectively teaching students with ASD, researchers suggest the educational, employment, and quality-of-life outcomes for individuals with ASD remain uncertain, often poor, and far below those of peers without disabilities (Bishop-Fitzpatrick et al., 2016; Light & McNaughton, 2015; Roux, Shattuck, Rast, Rava, & Anderson, 2015; U.S. Department of Labor, 2017). A disconnect exists between the research supporting best teaching practices and the outcomes for students with ASD.

Efforts to improve the effectiveness and quality of education for all students is driven by legislative and policy requirements (ESSA, 2015). Yet the appropriateness of current teacher evaluation systems for special education classrooms is disputed (Johnson & Semmelroth, 2014; Jones & Brownell, 2014; McCaffrey & Buzick, 2014; Woolf, 2015). Challenges associated with evaluating special education teachers and classrooms include varied instructional responsibilities, heterogeneous student populations, specialized knowledge, and a range of teaching conditions and environments (Buzick & Jones, 2015; Goe et al., 2008; Jones & Brownell, 2014). Classroom observation measures can provide evidence of teaching practices and student learning (Crowe et al., 2017; Kane, McCaffrey, Miller & Staiger, 2013). But using one universal quality measure for special education teaching does not account for the specialized teacher skill sets and classroom differences necessary to meet the unique needs of students with disabilities (Economic Policy Institute, 2010; Crowe et al., 2017; Johnson & Semmelroth, 2014).

Effective teaching and quality classroom practice positively influence student-learning outcomes (Goldhaber, 2010; Kane & Staiger, 2012; Linstead et al., 2017). Educators are required to teach students with disabilities, including the rising number of students with ASD (CDC, 2014), using the same academic standards as for their general education peers (ESSA, 2015; Hart & Whalon, 2011). Legislative emphasis on accountability to improve outcomes for all students (ESSA, 2015) means evaluation methods are needed to identify the quality and effectiveness of teaching within special education classrooms serving students with ASD. Many measures of effective teaching are available for general education classrooms and teachers of academic content areas (Holdheide, 2015; Goe et al., 2008). Few instruments have been developed for special education classrooms (Darling-Hammond, 2015; Goe et al., 2008; Jones & Brownell, 2014). Only two instruments rate the quality of classrooms specifically serving students with ASD: the APERS (Odom et al., 2013) and the QIASD (Pearl et al., 2017).

A systematic literature review in Chapter 2 revealed a clear gap in the existing research on developing high-quality observation measures to assess teaching practices in K-12 special education classrooms serving students with ASD and on investigating the psychometric properties of those measures (Crowe et al., 2017; Jacob et al., 2016; Semmelroth & Johnson, 2014). Several observation tools focused on special education classrooms (Semmelroth & Johnson, 2014; Tsai et al., 2013) and specific methodologies for teaching students with ASD (Leaf et al., 2016). However, review of the literature revealed only two empirical articles published on observation measures specific to teaching effectiveness of students with ASD in K-12 special education classrooms (Odom et al., 2013 and Pearl et al., 2017). A review of the literature did not yield any published psychometric data on the validity or reliability of APERS

(Odom et al., 2013) scores. Pearl et al. (2017) recently conducted a content validity study that supported alignment of the 51 quality indicators selected for inclusion in the QIASD with the seven CEC initial practice standards for special education of students with ASD (CEC, 2014).

The problem is very few psychometrically sound measures are available to identify the quality of special education classrooms serving students with ASD (Crowe et al, 2017; Johnson et al, 2016). Observation measures designed for special education classrooms should capture the unique context, student characteristics, and specialized teacher skills that represent quality teaching and influence student growth (Semmelroth & Johnson, 2014; Tandy et al., 2016). A current need is the creation of instruments to judge the presence of quality teaching and educational supports in special education classrooms serving students with ASD (Crowe et al., 2017; Holdheide, 2015). The QIASD is intended to meet this need as a measure of quality teaching practices necessary for special education classrooms to effectively serve students with ASD (Pearl et al, 2017).

As specified by current educational testing standards, examining the measurement validity of the QIASD is a step towards trusting the intended inferences made from the scores (AERA, APA, & NCME, 1999; 2014). Pearl et al. (2017) obtained psychometric evidence for the content of the QIASD through expert feedback on the selected indicators as appropriate measures of autism classroom quality. The next step in the process of validation evidence is to examine the internal structure of the QIASD instrument (AERA, APA, & NCME, 2014).

In this study, the researcher investigated the internal structure of the QIASD measure. The purpose of this research was to add to the validity evidence of the QIASD ratings by examining the internal consistency reliability and the construct validity of the scores from a

sample of K-12 special education classrooms serving students with ASD. The researcher provides implications to the field of special education, limitations, and future research recommendations related to the study objectives.

Implications of Literature Review

The results of the structured literature review in this study revealed two challenges associated with developing and validating measures of effective teaching in special education classrooms serving students with ASD. First, the construct of *effective teaching* has encompassed multiple concepts in the literature, including content knowledge, pedagogical skills, student characteristics, family support, school climate, and classroom learning environment (Cantrell, 2013; Connor, 2013; Little et al., 2009; Marshall et al., 2016). Educational policies over the past two decades also reflect changes in how effective teaching is evaluated through teacher qualifications, student achievement growth scores, and delineation of ineffective teaching (ESSA, 2015; IDEA, 2004; NCLB, 2002).

Second, the unique and heterogeneous learning characteristics of students with autism spectrum disorder, especially those students with more severe ASD symptoms, require teachers to use specialized knowledge, skills, and practices for effectively instructing students with ASD (Iovannone et al., 2003; Johnson & Semmelroth, 2014; Wong et al., 2015). Many teacher evaluation systems used in school districts across the nation are based on general education and do not adequately address the roles and professional practices unique to special education teachers (Crowe et al., 2017; Kane & Staiger, 2012; Semmelroth & Johnson, 2014; USDOE,

OPEPD, 2016). Researchers should identify aspects of effective teaching and quality educational practices to include in a special education classroom observation scale. Current literature supports several options for teaching students with autism spectrum disorder, including evidence-based interventions (i.e., Odom et al., 2010; Simpson, 2005; Wong et al., 2015), curriculums and programs (i.e., Mesibov, Howley, & Naftel, 2015; Odom et al., 2013; Turnbull & Knapp, 2017), standards (i.e., CEC, 2015), methodologies (i.e, Leaf et al., 2016; Mesibov, Shea, & Schopler, 2005), and quality indicators (i.e., Crimmins et al., 2001; NRC, 2001; Pearl et al., 2017). One single assessment may not be able to measure every component of effective teaching of students with ASD. The researcher's methodology and decision processes to establish construct validity evidence for the QIASD scores reflect challenges consistent with those in the literature and are discussed in the next section.

Implications for Methodology

A confirmatory factor analysis (CFA) was chosen as the main method of statistical analysis to establish construct validity evidence by verifying the internal structure of the QIASD model with these data (AERA, et al., 2014; Kline, 20). A CFA is a theoretical approach to test an a priori hypothesized model of the underlying structure of a set of variables (Tabachnick & Fidell, 2013). A CFA with QIASD scores tested the hypothesized relationship between the observed variables and their underlying latent variables (factors). The researcher based the QIASD model for this analysis on empirical research and theory supporting the relationship between the 51 quality indicators and the seven factors aligned with the CEC professional

practice standards (CEC, 2014). The results and the iterative methodological adjustments to the data analysis made in attempt to obtain the most valid results are discussed.

Implications of Analysis

Descriptive Statistics

One of the intended uses of the QIASD is to differentiate "teacher performance with students with ASD" (Pearl et al., 2017, p. 67) using a rating scale to describe the presence of quality indicators on a range of 0 (unsatisfactory) through 4 (highly effective). The researcher was interested in seeing if the range of data scores showed observers were only using a few of the rating options (i.e., effective and highly effective), as researchers have suggested can happen with classroom observation tools (Doherty & Jacobs, 2015). If observers were not using the full range of rating options, this QIASD scoring system may need to be modified or further research conducted on score distributions to determine if raters are not distinguishing effective teaching from ineffective teaching. Based on this sample (N = 102), 42 observed variables were rated using the entire range of scores (0-5). These results indicate the scoring system may have an appropriate range of options to rate the presence of quality indicators in this sample.

Statistical Assumptions

The results from these data showed violations of the statistical assumptions for conducting a confirmatory factor analysis. Traditional covariance-based CFA software programs and analyses rely on these assumptions to produce valid results (Kline, 2015). In addition, widely

used ML-based model fit indices are not robust to violations of statistical assumptions (Kline, 2015; Tabachnick & Fidell, 2013). Fortunately, nonparametric methods and related software have been developed to run confirmatory statistical analyses with these data.

The researcher initially estimated a minimum sample size greater than 100 ($N > 100$). The sample size of the actual data was 102 ($N = 102$). The post hoc power analysis through WarPLS indicated a sample size of $N = 146$ in order to obtain a power of .08. The researcher tested the reflective model using ML estimation with Bollen-Stine bootstrapping in AMOS because it is robust to smaller samples. However, with violations of the normality and linearity assumptions evident in the data, this covariance-based method may not have produced accurate results.

These violations of statistical assumptions, in combination with re-specifying variables as formative, led the researcher to re-run the analysis with the WarpPLS Stable3 method of analysis (Kock, 2014). The WarpPLS Stable3 is a nonparametric technique using a process similar to bootstrapping and is appropriate to use when data are multivariate nonnormal and nonlinear (Kock, 2014). In this study, the researcher's use of the WarpPLS nonparemtric method to analyze the formative model likely produced more accurate results that may be interpreted with more confidence than the AMOS results.

Internal Consistency Reliability

The internal consistency reliability of the QIASD scale was assessed in SPSS, version 23, using Cronbach's alpha (α). First, the researcher used Kline's (2015) recommendation for a Cronbach's alpha value of .7 to .8 to check the overall reliability of the QIASD scale. The high

total Chronbach's alpha of .913 indicated good internal consistency reliability for the QIASD scale (Kline, 2015). However, the researcher suggests caution when interpreting the overall alpha coefficient value, as it is dependent on the scale size. Specifically, as the number of items on the scale increases, the value of alpha also increases (Tabachnick & Fidell, 2013). As a result, a scale with a large number of items may obtain a large alpha value simply by having many items on the scale, not because the scale is reliable (Tabachnick & Fidell, 2013).

Next, the researcher reviewed the correlation values in the Corrected Item-Total Correlation column for values less than .3. This value would indicate the item does not correlate with the overall scale (Tabachnick & Fidell, 2013). The following eight quality indicators had correlations less than .3: Environ2b, Environ2h, Environ2k, Curric3b, Instruc5b, Instruc5c, Instruc5m, and Collab7g. The lack of correlation between so many individual indicators and the overall scale suggests the total internal consistency reliability score, α = .913, may be inflated, as described previously, due to the large number of items (51) on the scale. These low correlations may have posed a potential problem for obtaining accurate results with the confirmatory factor analysis on the reflective measurement model (Kline, 2015; Tabachnick & Fidell, 2013).

Finally, based on Cronbach's (1951) recommendation to apply alpha separately when several factors exist (Tabachnick & Fidell, 2013), the researcher inspected the Cronbach's alpha values for the seven subgroups. As displayed in Table 10 (Chapter 4), three out of the seven subgroups had Cronbach's alpha scores less than .7, suggesting possible problems with the internal consistency reliability of those subgroup scores. The main concern was with the *ProfPrac* subgroup, with a Cronbach's alpha of .362 that did not meet the rules for good internal consistency reliability (Kline, 2015). Two solutions to improve the internal consistency of

ProfPrac may be to add some indicators that are highly related to the three items already in the subgroup, or to eliminate poorly correlated items.

One reason for the low alpha value may be due to *ProfPrac* having only three indicators to measure professional practice, which is one of the smallest subgroups on the QIASD. Researchers may review the CEC Initial Preparation Standards (CECE, 2015), which include seven items of knowledge under the Professional Learning and Ethical Practice standard, to identify one or two more indicators to add to the QIASD. But the other small subgroup, *Assess*, with only four indicators, displayed good reliability (α = .787), which suggests the low reliability of *ProfPrac* may not be caused simply by a low number of quality indicators (Field, 2013). In this case, adding more items to the ProfPrac subgroup in attempt to improve the internal consistency would seem to mask an underlying problem with the indicators.

Instead, the poor internal consistency reliability for the *ProfPrac* subgroup scores may be explained by the diverse themes represented by the three individual indicators. The *ProfPrac* indicators, related to the CEC standard on Professional Learning and Practice, are as follows:

(a) "Hands-on" contact with students promotes independence and preserves dignity.

(b) Inter-staff communication is respectful of students and limited in content to classroom issues and instruction. Confidentiality of students is preserved.

(c) Restrictive procedures employed are supported by a Functional Behavior Assessment and Behavior Intervention Plan.

It appears the low correlation between these three indicators stems from being measures of very different aspects of professional learning and practice. The solution of removing indicators may not make sense because the three *ProfPrac* items were deemed important to the practical

application of the QIASD (Pearl et al., 2017). With this in mind, researchers would need to make a decision about what to do with the *ProfPrac* subgroup (i.e., reviewing the literature to explore potential changes to the *PrafPrac* indicators or possibly removing this subgroup).

The two other subgroups with Cronbach's alpha's below .7 were *Environ* (α = .683) and *Collab* (α = .680). The internal consistency reliability of *Environ* (Learning Environments) could be improved to .751 with the removal of variable Environ2k, which would meet expectations for good reliability. The quality indicator Environ2k specifies "the teacher can provide examples of opportunities for meaningful interaction and friendships with peers without disabilities" (Pearl et al., 2017). This indicator may be problematic because the QIASD was designed for self-contained special education classrooms and observations in this setting may not offer an opportunity to observe interactions with typically developing peers. Also, this indicator is very different from the other *Environ* indicators related to easily observable items such as room arrangement, visual schedules, and classroom materials. The low internal consistency of indiator Environ2k may be addressed by moving it to a different subgroup. The exploratory factor analysis results suggested Environ2k loaded significantly onto the Instruct (Instructional Planning and Strategies) indicator.

The low alpha value for the *Collab* (Collaboration) subgroup would improve a small amount from .680 to .691 with the removal of variable Collab7b. The quality indicator Collab7b states "all classroom staff is involved in delivering instruction, including during out-of-classroom activities (lunch, recess, CBI)." The CFA results confirmed Collab7b as a problematic indicator that did not load significantly onto the Collab factor. Similar to Environ2k, this indicator may be difficult to observe because it is based on activities outside the special education classroom.

Also, activities such as lunch and recess are often times when either teachers or paraeducator staff are taking their break or lunch time, which could be considered a form of staff collaboration. So, the suggestion in this indicator that "all" staff be involved in these activities may not accurately represent the collaboration occurring. The concept and wording of Collab7b should be reviewed.

While the Cronbach alpha values for subgroups *Envrion* and *Collab* are less than .7, researchers suggest values as low as .5 could fall into the range of acceptable reliability in early stages of research and when diversity within constructs is expected (Field, 2013; Nunnally, 1978; Tabachnick & Fidell, 2013). Research on the QIASD is in the early stages and the poor internal consistency of three subgroups indicates further research is needed to explore the scoring procedures and structure of the QIASD.

Construct Validity

The primary purpose of examining the construct validity of the QIASD scale was to provide evidence supporting the accuracy and interpretability of scores. The internal structure of the QIASD measure was evaluated in this study using three analyses: a covariance-based confirmatory factor analysis, an exploratory factor analysis, and a partial least squares confirmatory factor analysis. A comparison between the hypothesized seven-factor reflective model and the formative model assisted with identifying the best model fit with these data. The next section offers implications founded on the results of these analyses.

The reflective measurement model of the QIASD was based on prior research and grounded in the widely recognized Council for Exceptional Children (CEC) professional practice

standards. In this model, the researcher interpreted the 51 observed variables (quality indicators) as manifestations of the seven latent variables (standards). The researcher attempted to account for small sample size and violations of statistical assumptions using the Bollen-Stine bootstrapping method in AMOS.

The covariance-based CFA used to examine the hypothesized seven-factor reflective measurement model resulted in inconsistent findings. The Bollen-Stine *p*-value (.263) was nonsignificant ($\alpha = .05$) for the adjusted chi-square indicating an acceptable model fit based on the transformed bootstrap sample. In reality, the other fit indices and the factor loadings reflected how the actual observers ($N = 102$) responded on the QIASD. The RMSEA (.101), TLI (.374), GIF (.535) and AGFI (.496) model fit indices did not meet acceptable thresholds for good model fit. Similarly, the factor loadings (see Table 10, Chapter Four) showed over half (53%) of the observed variables did not load significantly onto their respective factors and the communalities indicated 69% of the observed variables accounted for only minimal amounts of variance (< .40) in their factors.

These results suggest a problem with the hypothesized seven factor model. The low factor loadings and communalities from this sample indicate the patterns of responses for multiple observed variables were not similar enough to be highly associated with the same factor. For example, indicators 1a, 1b, and 1c do not load significantly on the *Learner* factor. Learner1b and 1c are focused on individual learner needs within instruction, which may have posed a challenge for observers to rate these indicators accurately and consistently. Observers conducted the QIASD in classrooms where they were either familiar or unfamiliar with the students, which may have caused inconsistencies in the pattern of responses. Learner1a is based

on staff interacting with each student within a five-minute period. Fluctuations in ratings of this indicator may be due to the varied number of students, number of staff, and type of instruction occurring in the different classrooms used in this study. Future researchers should consider how differences in observers and settings may influence ratings on the QIASD and should plan to account for these differences within research studies.

The covariance-based CFA did not result in an acceptable fit of the reflective measurement model with these data as evidenced by the global fit indices and factor loadings. The hypothesized seven-factors were not well-defined by the 51 quality indicators as specified. The researcher may have removed problematic observed variables and retested the model fit, but far too many variables (27) had low factor loadings. Hair, Babin, & Krey (2017) advised against removal of more than 20% (in this case 10) of the observed variables, which would indicate a flawed measurement theory. In addition, two items should be considered when interpreting these findings. The sample size may have been too small to obtain valid bootstrap results (Arbuckle, 2008), and AMOS discarded 55 unused samples in the bootstrapping process, which indicated a potential problem with model specification (Arbuckle, 2008). The researcher conducted further analyses of the factor structure to identify the source of the problem.

The researcher conducted an exploratory factor analysis with these data allowing the observed variables to load freely on factors. The findings of the EFA (see Table 11) revealed the 51 observed variables loaded onto 16 factors, which was inconsistent with the specified seven-factor model. Less than 50% of the cumulative variance was explained by seven factors. These findings suggest the structure of the QIASD with 51 quality indicators grouped to measure seven factors may not make sense. The goal of aligning the QIASD with the seven CEC preparation

standards was to base the assessment on best practices and also to connect classroom practices with accredited teacher preparation content (Pearl et al., 2017). A 16-factor model indicated by the EFA may result in a more complicated assessment that may not align with theoretical or empirically-based dimensions of effective teaching.

Specification of a model to analyze the structural validity of an assessment's scores requires grouping variables based on theoretical or empirical support for the construct being measured. A QIASD content validation study (Pearl et al., 2017) supported the inclusion of the 51 quality indicators based on alignment with the seven CEC specialty set preparation standards for developmental disabilities and ASD. These CEC preparation standards were developed by professionals in the field of special education and stem from research-based practices for students with ASD (CEC, 2015). Woolf (2014) reported both special education teachers and administrators perceived all of the skills identified within the CEC standards as important to teacher effectiveness. Based on the study results, Woolf (2014) suggested some of the skills represented in the CEC standards groups may overlap and may not clearly define the behavior they intend to measure. The results of the current research suggest the 51 QIASD indicators do not actually measure the seven dimensions of teaching as hypothesized. The quality indicators selected to represent the seven domains on the QIASD may need to be more clearly operationally defined or reassessed as reliable measures of those domains. Future researchers should further explore the empirical validation of teaching evaluation instruments based on the CEC preparation standards and related skill sets.

The problematic patterns of factor loadings resulting from the CFA and the EFA led the researcher to question the theoretical dimensions of the hypothesized model. As described in

Chapter Four, the researcher reevaluated the empirical support for the reflective measurement model and specified an alternative formative measurement model. Many researchers have unintentionally specified a reflective model for a scale when the construct really calls for a formative model (MacKenzie, Podsakoff, & Jarvis, 2005). One main difficulty in correctly specifying a measurement model stems from interpreting the multitude of research and theoretical perspectives underlying a hypothesized construct. In this study, the researcher specified a reflective model based on empirical content validation (Pearl et al., 2017) of the quality indicators reflecting the CEC standards as presented in the QIASD measure. The data analysis results, with poor internal consistency and low factor loadings, implied this interpretation may not be accurate. Thus, the researcher specified an alternative formative measurement model established from the perspective of quality indicators as independent components grouped together to define the concepts represented by the seven CEC standards.

The formative measurement model had a better fit to these data than the reflective measurement model as evidenced by the PLS CFA results. The AFVIF (1.833) was significantly low and the SRMR (0.139) was very close to the acceptable value of 0.1. The findings showed low VIF values (< 3.3), significant observed variable weights ($p > .05$), and satisfactory effect sizes (> .02) for the majority of observed variables (see Table 15). Yet, the results also revealed some problems with the specified formative measurement model. The fit indices were not sufficient to accept a final model and the high collinearity among thirteen observed variables indicated potentially redundant items. The researcher chose not to remove these problematic observed variables since doing so may alter the theorized latent variable construct. The original construct domain in a formative model may be easily distorted if the number or type of observed

variables that define the latent variable are changed (Coltman, Devinney, Midgley, & Venaik, 2008). Researchers should further investigate the quality indicators deemed necessary to sufficiently represent the QIASD domains (standards) of quality special education classrooms for effectively teaching students with ASD.

The factor analyses in this study began the process of establishing construct validity of the QIASD measure and offered preliminary insights into refining the instrument. The main problem consistent across the three different factor analyses was the consequence of the small sample size. The small sample may have emphasized the nonnormality of these data (Field, 2013; Hahs-Vaughn & Lomax, 2012). The sample (N = 102) was too small to obtain a desired power of .80, which increased the probability of a Type II error (failing to reject the null hypothesis when it actually is false). Also, the small sample may have biased the results of the confirmatory factor analyses making the interpretations of model fit and factor loadings inaccurate. A larger sample size would offer more robust results to examine the internal structure of the QIASD measure.

Implications for Instrument Development

Review of QIASD Development

The present study served as a next step to establishing validity evidence for the QIASD ratings as used with a sample of observations from K-12 special education classrooms serving students with ASD in Florida. An initial study conducted by the authors of the QIASD lent validity evidence of the QIASD content with experts supporting the quality indicators as aligning

with the CEC standards (Pearl et al., 2017). Exploring the validity of classroom observation measures should include how well the instrument reflects professional standards of teaching practice (Goe et al., 2008). The QIASD is intended for use as a measure of teaching proficiency (Pearl et al., 2017), thus this study is a step towards examining how well the QIASD design supports the validity of score interpretations as intended.

The QIASD was designed to directly align with the CEC initial preparation standards and specialty set for students with ASD (CEC, 2015; Pearl et al., 2017). The QIASD may be a useful tool for measuring whether crucial skills and supports teachers have learned are actually in place. This study extended research by Pearl et al. (2017) by focusing on validity evidence important to the iterative development and future dissemination of the QIASD tool. Prior to developers actively using and disseminating an educational evaluation instrument, the Standards for Educational and Psychological Testing recommend all of the standards should be considered, as appropriate to the test and its intended use (AERA et al., 2014). The results of this study offer important information relevant to the development of an instrument designed to measure effective teaching practices in special education classrooms serving students with ASD. Researchers should further examine the empirical basis for the QIASD structure and consider modifications to ensure alignment with evidence-based practices and critical domains of effective instruction.

Designing an educational test "begins with consideration of expected interpretations for intended uses of the scores to be generated by the test (AERA et al., 2014, p. 75). Pearl et al. (2017) noted the purpose of the QIASD ratings are to provide "discrete and actionable feedback" (p. 3) and to assess the performance of special education teachers serving students with ASD.

Once the purpose is defined, test developers should generate a theoretical framework specifying aspects of the construct to be measured with the test content (AERA et al., 2014). The construct must be sufficiently defined in order to assess whether the indicators on the QIASD accurately measure the construct (AERA et al., 2014). Pearl and colleagues (2017) reported the theoretical framework of the QIASD aligned with the CEC professional practice standards and literature on evidence-based practices for teaching students with ASD. The results of this study highlighted the challenges translating this framework into a correctly specified model of the QIASD.

Model Specification

One of the greatest challenges to developing a high-quality educational assessment is to describe the hypothetical construct to be measured and to specify the model correctly based on the construct (AERA et al., 2014). Effective teaching is a broad construct that continues to be shaped by the field. The researcher in this study assumed the QIASD had a strong enough base of theoretical and empirical support to define a reflective measurement model for the 51 quality indicators to load on to the seven identified latent variables. As the researcher began analyzing the data, some issues became apparent (i.e., low internal consistency of subgroups and many indicators with low or cross factor loadings) prompted the researcher to re-examine the nature of the variables.

While the content of the QIASD has been endorsed by experts in the field (Pearl et al., 2017), supporting the alignment of the 51 quality indicators with the seven CEC professional practice standards, the directional relationship between the observed variables (quality indicators) and latent variables (standards) in a statistical model may be hypothesized in different

ways (Baxter, 2009). In practice, “the researcher has the flexibility to conceptualize a measurement model based on the construct definition the researcher specifies” (Sarstedt, Hair, Ringle, Thiele, & Gudergan, 2016, p. 4000). The researcher based the initial QIASD model in this study on the quality indicators as reflective variables, meaning the factors (standards) caused the related observed variables (quality indicators). The reflective model assumed the indicators would load onto respective factors because they were specified to measure the same property of the construct (Lowry & Gaskin, 2014).

The poor model fit and low factor loadings for so many indicators signified a problem with the hypothesized relationship between the variables. A closer look at the quality indicators showed they measure different aspects of the standard with which they align. For example, under Learning Environment, there are quality indicators measuring room arrangement, posting of classroom schedules, and opportunities to interact with peers without disabilities. These items all relate to the learning environment domain but they measure unique components of the construct and are not interchangeable (Jarvis, MacKenzie, & Podsakoff, 2003). The quality indicators may be more appropriately specified as formative observed variables that measure separate features of the seven QIASD domains (Kock, 2014; Lowry & Gaskin, 2014). This reasoning led to an interpretation of the QIASD model as formative, which was reinforced by the literature on individual evidence-based practices (Wong et al., 2015) and components of effective teaching of students with ASD (NRC, 2001).

The results of the PLS-based CFA supported the specification of a formative model. Thirty-eight observed variable weights had significant *p*-values (= < .05), which indicated a majority of observed variables were significantly associated with scores of the respective latent

variables (Kock & Mayfield, 2015). The VIF values for the observed variables were all below the recommended threshold of 3.3, indicating the observed variables were not redundant (Kock & Mayfield, 2015). These results suggest the data fit the alternative formative measurement model better than the original hypothesized seven-factor reflective model.

Procedures for Interpreting Ratings

The main goal of teacher evaluation is continuous improvement for teachers and students (Darling-Hammond, 2014). A fair and valid teacher evaluation system should offer actionable feedback in a way that can lead to improved practices, increased use of relevant evidence-based intervention, and better outcomes for all students (Darling-Hammond, 2015; Goe et al., 2008; Johnson & Semmelroth, 2014). The intended use of the QIASD is to measure teaching performance of students with ASD and to allow for actionable feedback to guide professional development and teacher improvement efforts (Pearl et al., 2017). The design of the instrument content delineated seven subgroups, based on professional practice standards, as primary dimensions of teaching performance. Currently no items in the QIASD exist for rating the overall construct of effective teaching or for rating the seven subgroups. The instrument rates only for the 51 individual quality indicators. Thus, determining the extent individual indicator scores relate to overall or subgroup scores is not possible at this time. The QIASD scoring system could be modified to enhance the ability of raters to identify overall teaching effectiveness and to distinguish between teaching performance within the seven subgroups.

Implications for Practice

Researchers suggest teachers are frequently rated as effective or highly effective (Doherty & Jacobs, 2015) and higher student achievement is positively correlated with effective teaching practices (Aaronson et al., 2007; Cantrell, 2013; Kane & Staiger, 2012; Goldhaber, 2010; Leigh, 2010). The above outcomes indicate most teachers are effectively implementing quality practices and student outcome measures must be overwhelmingly positive. Other data contradict this conclusion (Howlin, 2013; Shattuck, Narendorf, Cooper, & Sterzing, 2012). For example, many teachers are not prepared to implement evidence-based practices (Brock et al., 2014; Iovannone et al., 2003) and many students with special needs continue to have inadequate educational outcomes. In particular, the literature suggests students with ASD have some of the poorest employment, post-secondary, and quality-of-life outcomes compared to typically developing peers (Anderson, Liang, & Lord, 2014; Bishop-Fitzpatrick et al., 2016; Light & McNaughton, 2015; Roux et al., 2015).

Three main recommendations emerge from this research related to effective teaching and teaching evaluation practices in the field of special education. First, psychometrically sound measures are needed for special education teachers of students with ASD to identify accurately areas for improvement in their teaching and classrooms. The ESSA (2015) continued a focus on accountability for all student outcomes and placed responsibility on state and local education agencies to develop appropriate systems to evaluate teacher effectiveness. Researchers and practitioners have increasingly been concerned about the appropriateness and validity of widely used evaluation methods such as value-added models and generic teacher effectiveness

frameworks (Gansle et al., 2015; Johnson & Semmelroth, 2014; Jones & Brownell, 2013) to appropriately address the expertise and specialized practices of special education teachers (Darling-Hammond, 2015; Pearl et al., 2017).

Recent prevalence rates up to 1 in 59 children diagnosed with ASD (Baio et al., 2018) indicate more special educators will likely be teaching in classrooms serving students with ASD. Improving special education teaching practices and learning environments to effectively serve students with ASD is essential. Thus, classroom observation instruments for teachers of students with ASD should align with focused knowledge and skills required to meet the unique and diverse needs of learners with ASD. The researcher's review of the literature showed few such instruments are available. Thus, the development and dissemination of the QIASD and other similar instruments is important to improve quality instruction and retain effective teachers, which may lead to better outcomes for students with ASD.

Second, states and districts responsible for developing and/or implementing measures of special education teaching effectiveness should examine the validity of score interpretations according to the Standards for Educational and Psychological Testing (AERA et al., 2014). Most available observation tools for measuring teaching practices in special education classrooms have little or no published empirical evidence to support the reliability and validity of their scores (Crowe et al., 2015). Teaching evaluation instruments are used in today's educational system to hold states, districts, and teachers accountable for student learning. As such, these instruments should be of high quality so stakeholders can have confidence in the accuracy of the ratings to distinguish between levels of teaching performance. The Standards (AERA et al.,

2014) outline necessary steps to gather evidence for the validity and reliability of an instrument's scores to measure the knowledge and skills it purports to measure.

Third, the QIASD may be useful to a range of practitioners and professionals in the field of special education because it was designed "specifically to provide special education teachers serving students with ASD with discrete and actionable feedback" (Pearl et al., 2017, p. 60). The QIASD may be an option for pre-service teachers to self-assess pedagogical knowledge learned during their coursework by rating the presence of quality indicators in either their own field placement or another special education classroom. Teacher preparation program faculty may also use the QIASD to inform coursework geared towards teaching students with ASD and to determine areas that need improvement or increased attention in the course. The QIASD may also be a practical measure for in-service special education teachers to self-evaluate the quality of their educational programming for students with ASD and to identify topics for professional development. In addition, mentors and coaches for early career special education teachers of students with ASD may find the QIASD particularly useful as a guide for delivering meaningful, proactive feedback to support teacher development. Finally, the QIASD it may be used by school and district administration to determine the quality of education students with ASD are receiving, to identify any gaps in resources or supports to be addressed, to develop appropriate professional development, and to inform continuous program improvement.

Implications for Future Research

Statistical Assumptions

The researcher accounted for violations of statistical assumptions by implementing the nonparametric partial least squares (PLS) method of analysis. Nonetheless, the sample size was small. Future researchers may attempt to replicate this CFA on the QIASD with a larger, random sample in attempt to obtain more accurate and powerful results.

In addition, one component of checking the factor structure of a model is to test the measurement invariance across samples (AERA et al., 2016; Floyd & Widaman, 1995). Researchers may explore whether the QIASD measures the same construct across samples. The sample of classroom observations for this study were all in Florida and were conducted by a purposeful sample of observers that may not be a reliable representation of the true population. Future researchers should attempt to compare random samples of classroom observations to reflect better the diversity of the population when testing for measurement invariance. A potential research question might be: To what extent is measurement invariance of the QIASD scores established as measured by goodness of fit indices across multiple samples?

QIASD Development

The formative model makes sense for the intended use of the QIASD to identify specific areas of need and guide professional development and improvement plans to meet those specific needs (Pearl et al., 2017). Although results suggest the formative model was a better fit than the

reflective model, an inspection of the indicator weights suggested some of the variables may be reflective. Future researchers may conduct a qualitative review of the QIASD quality indicators and further examine the empirical and theoretical support for the most appropriate measurment model. Further information may guide continued structural analysis of the QIASD dimensions based on reflective or formative variables, or perhaps specifying a mixed model with both reflective and formative variables, which are common in behavioral research (Lowry & Gaskin, 2014). A research question might be: Is there new research (such as the CEC High-Leverage Practices and current research on evidence-based practices) that could inform the use or dimensions of the QIASD?

Researchers may also attempt to improve the QIASD structure by adding total score and subgroup score items to the test. These overall and group scores would allow the rater to distinguish accurately between different levels of teaching performance by identifying relative areas of strengths and weaknesses. To be part of a teacher evaluation system, as Pearl and colleagues (2017) suggested, would be a beneficial use of the QIASD; administrators need to be able to summarize areas for professional development and determine a score of overall teacher effectiveness. Researchers could use these scores in analyses to gather more evidence in support of the internal structure of the QIASD measure as it was intended to be used. A strong internal structure is necessary to have confidence in the accuracy of the scores to measure the intended construct (AERA et al., 2014).

Test development is an iterative process that involves multiple facets of implementing and evaluating the use and interpretations of the test (AERA et al., 2014). The following research

questions offer some further avenues for researchers to explore the psychometric properties and the functionality of the QIASD:

- Are there characteristics of raters (i.e., ethnicity, job type, experience, familiarity with teacher/classroom) that influence the scoring of the QIASD results?
- What is the convergent/divergent relationship between the QIASD scores and data from a similar assessment (such as the APERS or the QPQI)?
- To what degree do QIASD ratings of the 51 quality indicators reflect the seven hypothesized factors across different samples?
- Does the QIASD scale predict teacher evaluation scores as measured by a districts current teacher evaluation?
- Does the QIASD scale predict student performance as measured by achievement scores?

Measure Reliability

Future researchers should evaluate measures of effective teaching of students with ASD on whether raters can reliably score the items and whether the ratings discriminate among teachers as intended. In one study example of reliability, Lash, Tran, and Huang (2016) examined the distributions of teacher ratings in a Nevada school district using a teacher evaluation instrument adapted from the Framework for Teaching (Danielson, 2007). The district intended to use the instrument to distinguish between higher- and lower-performing teachers for the purpose of tenure, retention, and pay for performance decisions (Lash et al., 2016). The results of distribution of scores showed 85% of teachers rated as effective or highly effective,

indicating the measure did not discriminate well between effective and ineffective teachers (Lash et al., 2016).

Researchers may explore reliability of scores of the QIASD and similar instruments in multiple ways to support the consistency and accuracy of ratings, such as through the inter-rater reliability of observers, internal consistency reliability, or rating distributions. Potential research questions concerning score reliability of the QIASD may include: To what extent do observers consistently apply QIASD ratings as measured by inter-rater reliability agreement scores? Do the QIASD scores reflect test-retest reliability as measured by consistent ratings across administration on separate occasions? To what extent to QIASD scores reliably differentiate among effective teaching practices measured by the distribution of teacher ratings?

Professional Standards

Professional practice standards are important to the development of high quality, effective special education teachers (CEC, 2015). The CEC professional practice standards are currently endorsed by the Council for the Accreditaion of Educator Preparation (CAEP) as a means to demonstrate educators are prepared with the necessary knowledge and skills for effective teaching in the classroom (CEC, 2015). Researchers should support the continuous review and development of professional teaching standards to ensure they reflect current and appropriate knowledge and skills to meet the needs of stakeholders (CEC, 2015; McDonald, M., Kazemi, E., & Kavanagh, 2013). A teacher performance assessment based on practice standards could enhance special education teacher evaluation by focusing on the specific pedagogical skills and areas of expertise deemed crucial for effective teaching of students with disabilities

(Holdeheide et al., 2012; Jones & Brownell, 2014; Woolf, 2015). Quality educational assessments should undergo a validation process to ensure trustworthy inferences and decisions can be made with the scores (AERA et al., 2014).

The CEC Initial Specialty Set: Developmental Disabilities and Autism Spectrum Disorder preparation standards are regularly reviewed and updated with feedback from professionals and stakeholders as a means to validate the content of the standards (CEC, 2015). The QIASD was validated by subject matter experts for content aligning with the CEC specialty set of standards for teachers of students with ASD (Pearl et al., 2017). In the current study, the researcher explored the internal structure validity of the QIASD scores and found the framework of quality indicators based on the seven CEC specialty standards did not hold up as a measure of the construct as intended. The results showed redundancy and overlap of several quality indicators within and between subdomains, suggesting the items on the QIASD do not align under the seven quality teaching subdomains as hypothesized. These results reveal the need for further research to develop a strong, empirical basis for the QIASD as a framework for measuring quality teaching practices of special education teachers serving students with ASD. Perhaps the misalignment between the QIASD indicators and the CEC standards stem from a lack of validation evidence for this context. Researchers may continue to explore the validity of the CEC professional preparation standards as potential measures of effective teaching performance in evaluation instruments.

One area of research that may guide the refinement of the QIASD framework is high-leverage practices for special educators. The concept of high-leverage practices in education emerged with the current need for "learner-ready teachers with the necessary skills to

demonstrably improve achievement outcomes for all students" including those with disabilities (McLeskey & Brownell, 2015, p. 6). High-leverage practices have been defined as "a set of practices designed that are fundamental to support k-12 student learning and that can be taught, learned, and implemented by those entering the profession" (Windschitl, Thompson, Braaten, & Stroupe, 2012, p. 880). The Council for Exceptional Children and the Collaboration for Effective Educator Development, Accountability, and Reform (CEEDAR) developed a set of high-leverage practices for special education teachers that are grouped into four categories: collaboration, assessment, social/emotional/behavioral, and instructional (McLeskey, CEC, & CEEDAR, 2017). These high-leverage practices are based on current research and are meant to provide a focused set of key teaching practices critical for effectively educating students with disabilities in k-12 classrooms (CEC, 2017). Future researchers should investigate the use of these high-leverage practices to inform and improve the framework of evidence-based practices on the QIASD to measure the effectiveness of special education teachers of students with ASD.

The need for more qualified and effective special educators skilled in teaching students with ASD is supported by this study. Approximately 82% of the sample in this study, made up of graduate students in a Master of Special Education program, had or were currently pursuing state endorsement in teaching students with autism spectrum disorders. The QIASD is currently intended for teachers of students with ASD being educated within special education classrooms (Pearl et al., 2017) and includes items that may or may not be apparent or appropriate for general education classrooms. Researchers should consider assessing the validity of the QIASD as a tool for measuring high-leverage practices for teaching students with ASD in a variety of settings.

esearchers may examine the possible differences between priority practices in a self-contained special education classroom versus an inclusive general education classroom.

Social Validity

States and school districts invest large amounts of time and resources into developing teacher evaluation systems (Danielson, 2011; Hallinger, Heck, & Murphy, 2014) in the hopes they will ultimately improve teaching and thus increase student achievement (Kane, Kerr, & Pianta, 2014; Whitehurst, Chingos, & Lindquist, 2014). As states and districts have the flexibility to design teacher evaluation systems with their particular needs and priorities in mind (ESSA, 2015), it follows the instruments selected should match those needs. In addition, with the increased emphasis on family involvement (Garbacz, McIntyre, & Santiago, 2016; Webster, Cumming, & Rowland, 2017), parents of children with ASD likely have important views on the instruments used to measure best teaching practices for their children. Future researchers may consider collecting some social validity data by surveying teachers, administrators, and parents on how they feel about using the QIASD to measure effective teaching and as a component of teacher evaluation.

Study Limitations

Threats to Validity

A potential limitation of the study was the use of a purposive sample of special education classroom observations that could impact the external validity of the results (Gall, Gall, & Borg, 2007). The sample of observers (graduate students) and observed classrooms in this study may not be representative of the population of observers who would use the QIASD and of special education classrooms serving students with ASD. The results may not generalize to other potential observers (i.e., coaches, faculty, parents) or to special education classrooms in other geographical areas. The researcher attempted to minimize threats to external validity by identifying a sample of graduate student observers across two semesters who conducted special education classroom observations across a range of grade levels, schools, and districts.

A second potential limitation related to possible uncontrolled extraneous variables that may impact the internal validity of the study (Gall et al., 2007). The researcher attempted to control for possible extraneous variables, such as observer characteristics, that may have influenced the results by designing a thorough training and scoring reliability check with a set criterion all observers were required to meet before conducting the classroom observation using the QIASD.

Covariance-Based CFA Limitations

The covariance-based CFA examines constructs through the indicator loadings on factors and is advantageous in confirming overall fit of a hypothesized causal model (Lowry & Gaskin, 2014). This method estimates model parameters to minimize the discrepancy between the observed and proposed covariance matrices (Lowry & Gaskin, 2014). A main concern of covariance-based analyses is factor indeterminacy, which means it produces many models that fit the data and makes the argument for causality more difficult (Lowry & Gaskin, 2014). Researchers recommend covariance-based CFAs with models supported by "well-established theories that are empirically validated" in order to rule out competing models (Lowry & Gaskin, 2014, p. 130). The traditional covariance-based CFA method also assumes a larger sample size, normal data, and reflective observed variables (Kline, 2016). The limitations in this study for the covariance-based CFA analysis included a small sample size, nonnormal data, and possible issues with a reflective model specification.

The researcher addressed these limitations by using bootstrapping in AMOS, by re-examining the QIASD variables, and by using a different method of analysis (PLS-based CFA). The Bollen-Stine modified bootstrap method transforms the observed data to produce an artificial sample for which the null hypothesis is true (Bollen & Stine, 1992). The standard errors provided are approximate because the bootstrap sample is discrete, not continuous as the original population distribution (Arbuckle, 2008). Although the Bollen-Stine method produced a corrected chi-square value that indicated a good model fit, this must be interpreted with caution as the small sample size (N=102) may not have been large enough to correctly transform the sample (Arbuckle, 2008).

PLS-based CFA Limitations

The partial-least squares (PLS) CFA method differs from the covariance-based CFA as it tests the weights of components on composite scores, rather than loadings on factors. The PLS technique examines a model by explaining the variance in latent variables through iterative estimation of partial model relationships (Sarstedt et al., 2014). Unlike covariance-based CFA that is held to stringent statistical assumptions, the PLS CFA is a nonparametric approach robust to nonnormal data and it often achieves higher power with small samples sizes (Hair, Ringle, & Sarstedt, 2013; Kock, 2015; Sarstedt et al., 2014). In addition, the PLS method can analyze a model specified as formative, in which the latent variables are considered to be formed by the observed variables (Kock, 2015).

The main limitation of the PLS method in this study was the small sample size, which may bias results. The final sample size (N = 102) was lower than the recommended N = 146 to obtain a desired power of .08. Although the PLS CFA methods handle small sample sizes better than covariance-based CFAs, small samples are prone to biased indicator weights and affect the stability of the parameter estimates of a model (Hair, Hult, Ringle, Sarstedt, & Thiele, 2017; Kock, 2015). Rigdon (2016, p. 600) recommended when sample sizes are small, "the best course is to get more data," Hair et al. (2017, p. 629) noted "when populations are small and/or data is difficult to obtain, the application of PLS with smaller samples denotes a viable attempt to advance knowledge in these areas."

Conclusions

The Standards for Educational and Psychological Testing (2014) emphasize validation as "a process of constructing and evaluating arguments for and against the intended interpretation of the test scores and their relevance to the proposed use" (p. 11). The results of this study offer important insights into the continued development of the QIASD as a measure of quality teaching practices of students with ASD. A need exists for future researchers to conduct factor analyses using a larger, random sample of observers to increase the validity and power of the results (Kline, 2015). In addition, modifying the structure of the QIASD measure to include an overall construct score and subgroup scores may improve score interpretation and lead to stronger model comparison research.

The validation process can inform revisions of the instrument and the conceptual framework (AERA et al., 2014). Through this study, the researcher realized the need to further examine the framework of the QIASD as a representation of the intended construct. The internal consistency reliability and initial CFA results indicated the need to view the structure of the QIASD measure from a different perspective. The poor internal consistency reliability and model fit of the hypothesized reflective measurement model offered preliminary support for specifying a formative measurement model of the QIASD. While the formative measurement model obtained better fit to these data, further research is needed to find an acceptable model. An examination of the domains, the overall construct, and the scoring procedures of the QIASD may lead to improved validity and reliability of score interpretations.

Overall, this research is progress towards validation of the QIASD, an instrument that is designed to provide meaningful feedback to improve teaching and guide professional development to support successful outcomes for students with autism spectrum disorder. The Standards for Educational and Psychological Testing (AERA et al., 2014) highlight the need for multiple means of validation to occur within the development and implementation phases before widely disseminating an assessment for practical use. Future researchers should continue to examine current advancements in special educator preparation standards and evidence-based practices for teaching students with ASD to refine and improve the QIASD instrument to meet the needs of today's educational system.

APPENDIX A: UNIVERSITY OF CENTRAL FLORIDA INSTITUTIONAL REVIEW BOARD (IRB) APPROVAL FORM

University of Central Florida Institutional Review Board
Office of Research & Commercialization
12201 Research Parkway, Suite 501
Orlando, Florida 32826-3246
Telephone: 407-823-2901 or 407-882-2276
www.research.ucf.edu/compliance/irb.html

Determination of Exempt Human Research

From: **UCF Institutional Review Board #1**
FWA00000351, IRB00001138

To: **Rebecca K. Hopkins** and Co-PIs: **Eleazar Vasquez, Lisa A. Dieker, Matthew Todd Marino**

Date: **February 08, 2018**

Dear Researcher:

On 02/08/2018, the IRB reviewed the following activity as human participant research that is exempt from regulation:

Type of Review:	Exempt Determination Category #4
Project Title:	Establishing construct validity of the Quality Indicators for Classrooms Serving Students with Autism Spectrum Disorder (QIASD).
Investigator:	Rebecca K. Hopkins
IRB Number:	SBE-18-13716
Funding Agency:	
Grant Title:	
Research ID:	N/A

This determination applies only to the activities described in the IRB submission and does not apply should any changes be made. If changes are made and there are questions about whether these changes affect the exempt status of the human research, please contact the IRB. When you have completed your research, please submit a Study Closure request in iRIS so that IRB records will be accurate.

In the conduct of this research, you are responsible to follow the requirements of the Investigator Manual.

This letter is signed by:

Signature applied by Gillian Morien on 02/08/2018 05:02:04 PM EST

Designated Reviewer

APPENDIX B: QUALITY INDICATORS FOR CLASSROOMS SERVING STUDENTS WITH AUTISM (QIASD) FORM

Quality Indicators for Classroom Serving Students with ASD (QIASD)

The Quality Indicators for Classrooms Serving Students with Autism Spectrum Disorders (QIASD) was developed with the support of Project ASD, funded through the Office of Special Education Programs (OSEP). This instrument is designed to guide a classroom observer in evaluating the strength and consistency of specific indicators of quality educational programming for students with ASD. It includes quality indicators from the Observation Assessment for Teachers Providing Services to Students with Autism Spectrum Disorders (OAASD), the product of a Pepsa (Partnership for Effective Programs for Students with Autism) project by Dr. Teresa Daly (Director for the University of Central Florida Center for Autism and Related Disabilities (UCFCard) and Regina DeCatrel (Program Specialist in Autism, Seminole County School District); and subsequently revised and adopted by Florida Card Centers.

The QIASD reflects revisions and additions to quality indicators based on field testing of the OAASD and alignment with the Council for Exceptional Children (CEC) Initial Special Education Developmental Disabilities and Autism Specialty Set Standards. Seven CEC Preparation Standards to assure that professionals have mastered the specialized skills for safe and effective practice are addressed. The specialty sets capture the professional knowledge base, including empirical research, disciplined inquiry, informed theory, and the wisdom of practice for their area of expertise for each proposed knowledge and skill (CEC, 2010).

The QIASD consists of 52 quality indicators aligned with the seven CEC standards: (a) learner development and individual learning differences (b) learning environments (c) instruction curricular content knowledge, (d) assessment, (e) instructional planning and strategies, (f) professional learning and practice, and (g) collaboration. Each indicator is given a score of 0-4 or NA. Quality indicators received a 0 if absent; 1 if present, but not being used; 2 if present, but partially achieved; 3 if present and being actively used; 4 if present and being used consistently; and NA if there was not an opportunity to observe quality indicator during the one hour observation.

A column has been included for data collection method (DCM) for observers to indicate whether the data obtained was via direct observation (as indicated by the “DO” on the measure), and/or interview of the teacher or classroom staff (“I”), and/or artifact/example (“A”).

Observation Assessment for Teachers Providing Services to Students with Autism Spectrum Disorders Revised

Classroom/Teacher: _____ Administrator/Observer:_____

Date/Time:_____ School District: _____

School Name: _____ Grade Level of Students: _____

Activities Observed: _____ Service Delivery Model: _____

Number of Students Present: _____Number of Staff Present: _____

Scoring: On scale of 0-4, to what degree is this indicator present?

4: Highly Effective (Very Much Present)
3: Effective (Present)
2: Needs Improvement (Somewhat Present)
1: Developing (Very Limited Presence)
0: Unsatisfactory (Not Present)
NA: Unrated

Data Collection Method(s)

DO: Direct Observation
I: Interview
A: Artifact

LEARNER DEVELOPMENT AND INDIVIDUAL LEARNING DIFFERENCES		
CEC 1.0- Beginning special education professionals understand how exceptionalities may interact with development and learning and use this knowledge to provide meaningful and challenging learning experiences for individuals with exceptionalities.		
Quality Classroom Indicator:	**Rating**	**Comments**
a. Instruction is individualized and based on learner characteristics, interests, and ongoing assessment. DDA5 S1; DDA5 S4		
b. Schedules reflect a variety of learning formats for each student, including 1:1 instruction, small group, large group, independent work, and social interaction/leisure options. DDA5 S4		
c. Instruction incorporates natural and individualized reinforcers.		

d. Students with slow rates of learning are provided intensive levels of instruction, including daily one-on-one instruction sessions. DDA5 S1		
e. All adults have knowledge/access to IEP objectives being worked on for each student. Staff can respond with specifics to the question, "What is student working on?"		Interview/Artifact
f. IEP goals and objectives are embedded within daily activities and routines throughout the day to promote maintenance and generalization.		Interview/Artifact
LEARNING ENVIRONMENTS		
CEC 2.0- Beginning special education professionals create safe, inclusive, culturally responsive learning environments so that individuals with exceptionalities become active and effective learners and develop emotional well-being, positive social interactions, and self-determination.		
Quality Classroom Indicator:	**Rating**	**Comments**
a. Room arrangement has clearly defined visual boundaries for specific activities. DDA5 S10		
b. Room arrangement allows for supervision of all students at all times; and prevents or minimizes problem behaviors. DDA5 S10		
c. Staff ratio of 1 adult for every 3 students is maintained during (at least 75%) observation. Allow greater ratio if the students are included for part of the day and are not on access points.		
d. A daily classroom schedule is posted at student level, is visible and appropriate for students' level of symbolic functioning, and is used throughout the day. Schedule indicates what activity is current.		
e. Individual schedules are posted at child level and are being used correctly. Schedule is referred to for each activity, sequence of activities is adhered to unless change is noted. Student is engaged in using schedule.		
f. Transitions are supported by routines, environmental arrangement and scheduling.		
g. Visual supports are at the correct level of symbolic functioning, and are used to enhance predictability, facilitate transitions, and help convey expectations.		
h. Instructional materials and furniture are age appropriate. DDA2 S2		
i. Classroom materials are well organized (i.e. labeled, conveniently located, and stored when not in use).		
j. Individual workstations, when present, are arranged left-right or top-bottom, and tell how much work, what work, when finished, and what's next. Workstation materials are varied from day to day and are educationally/functionally relevant.		
k. The teacher can provide examples of opportunities for meaningful interaction and friendships with peers without disabilities.		Interview/Artifact

CURRICULAR CONTENT KNOWLEDGE		
CEC 3.0- Beginning special education professionals use knowledge of general and specialized curricula to individualize learning for individuals with exceptionalities.		
Quality Classroom Indicator:	**Rating**	**Comments**
a. Schedule and activities reflect distribution of curriculum across multiple domains that is appropriate for the age, level and individual needs of students in the classroom. DDA3 S4		
b. Curriculum/activities address and are aligned with appropriate grade level general education curriculum and standards. DDA2 S2; DDA3 S4; DDA5 S14		
c. Curriculum/activities address social communication skills (i.e. pragmatics, conversation, perspective taking) with adults and peers. DDA3 S1		
d. Curriculum/activities address functional communication (avoid/repair miscommunications) for all students. DDA3 S2		
e. Curriculum/activities address functional life skills and adaptive behavior to maximize independent functioning in school, home, vocational, and community settings. DDA1 S1; DDA3 S3; DDA5 S13		
f. Specialized instruction to enhance social participation across environments is provided. If social skills instruction is infused, there is evidence of planning and evaluation. DDA3 S5; DDA5 S12; DDA5 S15		
g. Curriculum/activities address self-regulation and self-monitoring. DDA5 S11		
ASSESSMENT		
4.0- Beginning special education professionals use multiple methods of assessment and data-sources in making educational decisions.		
Quality Classroom Indicator:	**Rating**	**Comments**
a. Written data are gathered consistently and frequently (daily or weekly) to track progress on IEP goals and objectives. DDA4 S1		Interview/Artifact
b. Assessment tools and methods are selected, adapted and used to accommodate the abilities and needs of individuals with developmental disabilities/autism spectrum disorders. DDA4 S1		Interview/Artifact
c. Data are collected for monitoring and analyzing challenging behavior and its communicative intent. DDA4 S2		Interview/Artifact
d. Students displaying behavioral difficulties have an individualized behavior plan that is being implemented or have been referred for a Functional Behavior Assessment (FBA). DDA4 S3		Interview/Artifact
INSTRUCTIONAL PLANNING AND STRATEGIES		

5.0- Beginning special education professionals select, adapt, and use a repertoire of evidence-based instructional strategies to advance learning of individuals with exceptionalities.		
Quality Classroom Indicator:	**Rating**	**Comments**
a. Instruction is systematic and based on learner characteristics, interests, and ongoing assessment. DDA2; S4; DDA3 S6; DDA5 S16		
b. Students remain actively engaged in learning opportunities throughout observation, with no more than 2 minutes down time.		
c. During five minute observation, staff interacts with each student at least once to teach or promote learning. Excluding students who are engaged in independent work.		
d. Instructional pace promotes high rates of correct responding, correct responses are reinforced or prompting/error correction is provided as needed.		
e. Skills are taught in the context of naturally occurring activities and daily routines. There is no down time for teaching.		
f. Communication directed to students is clear, relevant, appropriate to language ability, and grammatically correct.		
g. Communication directed to students presents opportunities for dialogue (rather than being largely directive).		
h. Communication directed to students consists of largely instructive/positive comments in comparison to corrective comments.		
i. Behavior problems are minimized by using proactive strategies including choices, clear expectations and positive reinforcement. DDA5 S5		
j. Instructional methods are grounded in evidence-based practices. DDA5 S3		
k. Staff create opportunities for spontaneous use of communication skills including student-to-student interactions.		
l. Students without verbal communication have AAC and actively use across activities. DDA5 S2		
m. Technologies are employed to support instructional assessment, planning, and delivery for individuals with exceptionalities.		
PROFESSIONAL LEARNING AND PRACTICE		
6.0- Beginning special education professionals use foundational knowledge of the field and the their professional Ethical Principles and Practice Standards to inform special education practice, to engage in lifelong learning, and to advance the profession.		
Quality Classroom Indicator:	**Rating**	**Comments**
a. "Hands-on" contact with students promotes independence and preserves dignity.		
b. Inter-staff communication is respectful of students and limited in content to classroom issues and instruction. Confidentiality of students is preserved.		
c. Restrictive procedures employed are supported by a Functional Behavior Assessment and Behavior Intervention Plan.		Interview/Artifact

COLLABORATION		
7.0- Beginning special education professionals collaborate with families, other educators, related service providers, individuals with exceptionalities, and personnel from community agencies in culturally responsive ways to address the needs of individuals with exceptionalities across a range of learning experiences.		
Quality Classroom Indicator:	**Rating**	**Comments**
a. A staff schedule showing staff and student assignments, locations, and activities, is prominently posted and being followed.		
b. All classroom staff is involved in delivering instruction, including during out-of-classroom activities (lunch, recess, CBI).		
c. There is a consistent system in place for regular (daily/weekly), informative and positive communication with families regarding student participation, progress and concerns.		Interview/Artifact
d. Two-way communication is encouraged by soliciting information and questions from families.		Interview/Artifact
e. A variety of opportunities for family involvement are provided (classroom activities, information sharing, and parent training).		Interview/Artifact
f. Teacher collaborates with team members to plan transition to adulthood that encourages full community participation. DDA5 S6; DDA5 S7; DDA7 S1		Interview/Artifact
g. Teacher collaborates with school personnel and community members in integrating students with ASD in various settings.		Interview/Artifact

Notes:

1. QIASD is based on the original OAASD, developed by Dr. Teresa Daly (UCFCARD) and Regina DeCatrel (Program Specialist in Autism, Seminole County School District). It was field tested and revised by Dr. Cynthia Pearl (Co-principal Investigator for Project ASD, University of Central Florida) and Jillian Gourwitz (Doctoral Candidate, Exceptional Student Education)
2. CEC Special Educator Preparation Standards- NCATE approved November 2012
3. DDA_S_ = CEC Special Education Developmental Disabilities and Autism Specialty Skill Set
4. Content Validity (Pearl et al., 2017)

APPENDIX C: SAMPLE OF QUALTRICS VERSION OF THE QIASD FROM

Quality Classroom Indicator (QI) Scoring

On a scale of N/A and 0-4, to what degree is this indicator present

4: Highly Effective (Very Much Present)
3: Effective (Present)
2: Needs Improvement (Somewhat Present)
1: Developing (Very Limited Presence)
0: Unsatisfactory (Not Present)
N/A: Unrated (No opportunity to observe during 1-hour observation)

Data Collection Method: (Select only one.)

Direct Observation
Interview
Artifact

<< >>

Powered by Qualtrics

CEC Standard 2.0 - *Learning Environments* focuses on creating safe, inclusive, culturally responsive learning environments so that individuals with exceptionalities become active and effective learners and develop emotional well-being, positive social interactions, and self-determination.

QI 2.0 a. Room arrangement has clearly defined visual boundaries for specific activities.

4 - Highly Effective

3 - Effective

2 - Needs Improvement

1 - Developing

0 - Unsatisfactory

N/A - Unrated

APPENDIX D: DEMOGRAPHIC PORTION OF QUALTRICS VERSION OF QIASD FORM

Date of your observation:

Observation time:

Start Time

End Time

School Name: (where you will be conducting your classroom observation)

School District: (where you will be conducting your observation)

Location of school in which you are conducting this observation?

- Rural
- Suburban
- Urban

Do you have State Endorsement for teaching students with autism spectrum disorder?

- Yes
- Currently pursuing ASD endorsement
- No and not pursuing

Current position/job title:

- School Administrator
- Special Education Teacher
- General Education Teacher
- Other

APPENDIX E: SAMPLES OF QIASD SCORING TRAINING MATERIALS

Guidelines for Completing the QIASD Observation Instrument

1. Complete the QIASD Tutorial – The link to access the Adobe Connect pre-recorded QIASD Tutorial is provided below:

   (QIASD link)

2. Take the Quiz – Once you have completed the QIASD tutorial, take the Quiz. The quiz is designed to assess your understanding of how to rate the quality indicators. You must pass this Quiz at 90% or above.

3. Practice Scoring - Download the QIASD and practice scoring the QIASD while watching the 20-minute sample classroom video provided in the course module. Once finished, take the practice video quiz by entering in your responses from your practice QIASD. You must score a 90% on this quiz. Once you have completed all of the training components and quizzes to criterion, you will be provided a link to the on-line Qualtrics version of the QIASD for your on-site observation.

4. You will have a two-week period from when you complete the training module to complete the QIASD classroom observation.

5. Scoring the QIASD – The Qualtrics version of the QIASD is designed so you can take a laptop or tablet into the classroom and directly enter the ratings onto the electronic form. You may scroll back and forth through the form and/or save the form as needed.

Tips for Setting up a Classroom Observation

- Identify the classroom for students with autism you intend to visit. The observation must be conducted in a K-12 special education classroom serving at least two students with autism spectrum disorder.

- If you are unable to find a classroom in your district, –

    - you may set up a visit to one of the Mentor Demonstration Classrooms. You may go to the MDC website (see below) and check out the MDC teachers by clicking on *Meet Our Teachers*. You can click on the *Protocol for Visiting Mentor Demonstration Classrooms* to find out more about the visitation requirements for the different school districts.

      (MDC website)

    - You may set up a visit at _____ school near the campus. Go to the link below for information about the school and to sign-up as a visitor/volunteer.

    - If you have difficulty finding a classroom to observe, please contact your professor.

- If you are not visiting a classroom in the district where you currently work, you will need to go to the school district website and/or contact the school to complete the necessary paperwork that allows you to enter the school for observation purposes.

- Send an e-mail to the teacher introducing yourself and describing the purpose of your observation. Let the teacher know you will need a total of 1 hour and that about 10-15 minutes of that time will be to ask some questions and review some artifacts. It often helps to provide 2-3 specific dates and times that you are available to observe and ask the teacher if any of those times work to observe an instructional time in the classroom.

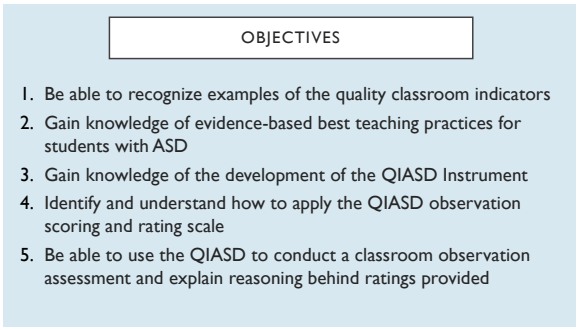
- Once you have scheduled an agreeable time, ask the teacher if she would prepare samples of the following documents for you to review during your observation:

    - Lesson Plan
    - Behavior Intervention Plan (if relevant)
    - Sample Individualized Education Plan (IEP)
    - Sample of IEP Data
    - Parent or Family and Teacher Communication

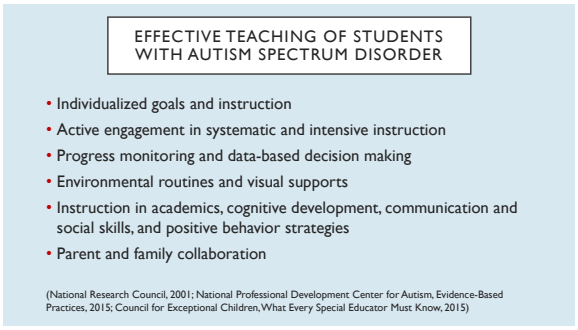
- Remember to be respectful and unobtrusive while in the classroom.

If you have any questions about conducting the observation, please contact your professor.

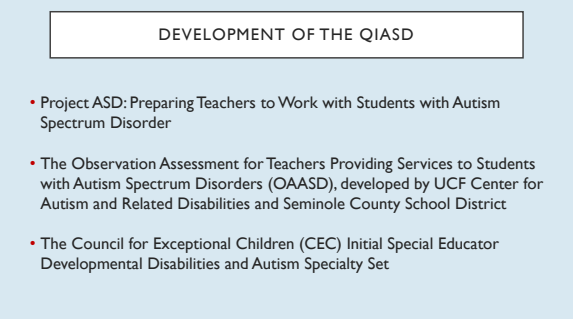
1. Quality Indicators for Classrooms Serving Students with Autism Spectrum Disorder



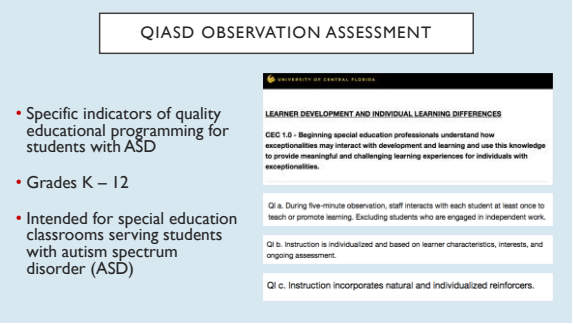
2. Objectives for Observers



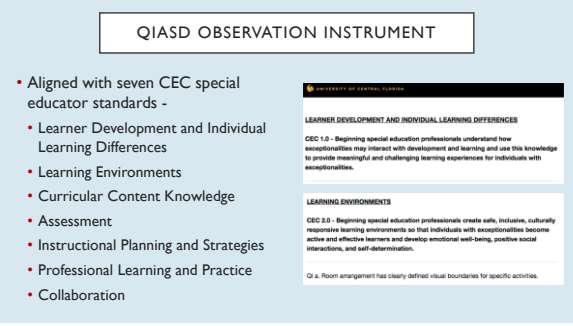
3. The QIASD is designed to assess the presence of quality indicators in special education classrooms serving students with ASD. Effective classrooms for students with ASD reflect a foundation of evidence-based best practices that demonstrate student learning outcomes are consistently achieved and well documented.



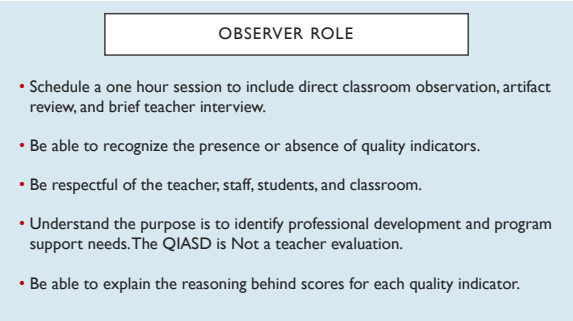
4. Development of the QIASD



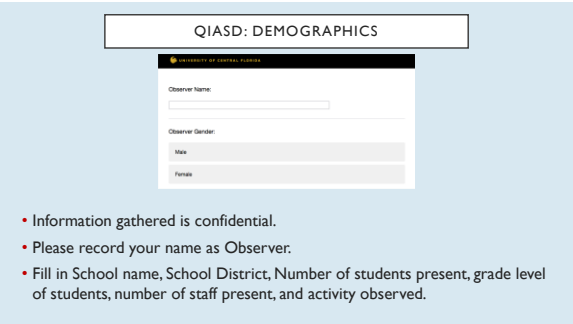
5. QIASD Observation Assessment



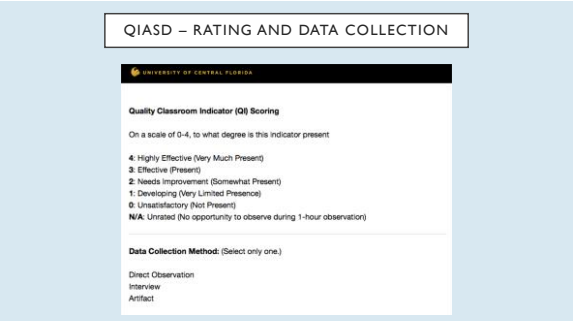
6. The QIASD is aligned with seven CEC special educator preparation standards from the Developmental Disabilities and Autism Specialty Set.



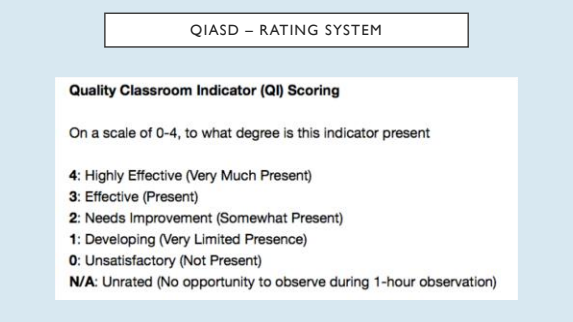
7. As an observer using the QIASD instrument, you should –
    a. Schedule a 1-hour observation session to include -
        i. direct classroom observation
        ii. artifact review
        iii. brief follow-up teacher interview.
    b. This tutorial will familiarize you with the quality indicators, so you will be able to recognize the presence or absence of them in the classroom during your observation period.
    c. Be respectful of the teacher, staff, students, and classroom. Develop a friendly, positive rapport with the teacher. And Do not be judgmental or disruptive to instruction.
    d. The purpose of the QIASD is to identify professional development and classroom support needs. It is Not an evaluation of the teacher.
    e. Be able to explain the reasoning behind scores for each quality indicator. You will use the comments section to note some specific teaching strategies or classroom activities that support your rating for each indicator based on your observation period.



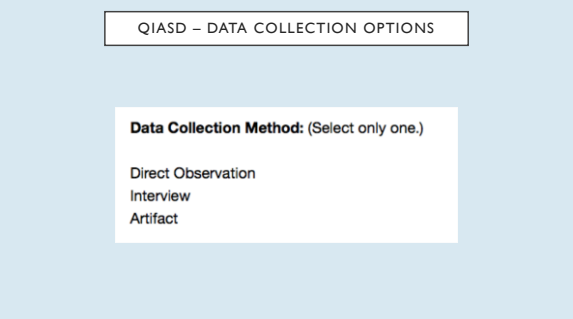
8. Demographic information.



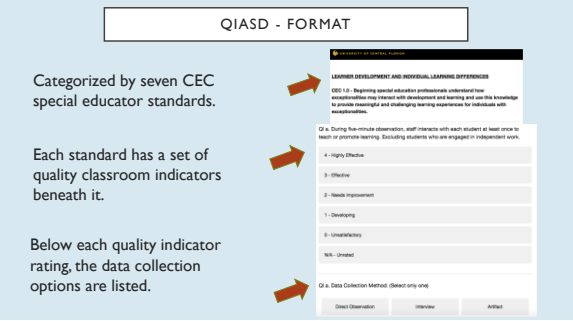
9. Let's take a closer look at these.



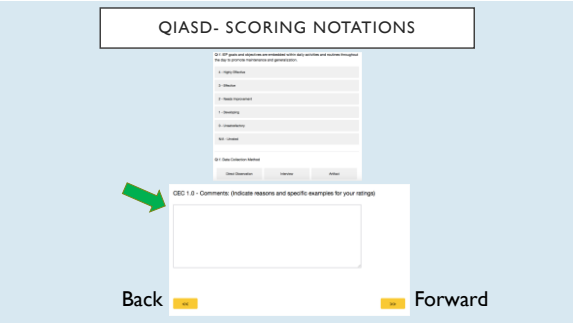
10. QIASD Rating System
    a. **4:** Highly Effective - the indicator is very much present and is consistently demonstrated at a high level of precision and expertise
    b. **3:** Effective - the indicator is present and actively demonstrated
    c. **2:** Needs Improvement - the indicator is somewhat present, but may not be demonstrated consistently
    d. **1:** Developing – there is very little presence of the indicator, it may be incomplete or ineffectively demonstrated
    e. **0:** Unsatisfactory - the indicator is not present at all, despite it being appropriate for the instructional context
    f. **N/A:** Unrated - there is no opportunity to observe the indicator during the scheduled 1-hour observation period; follow-up with a teacher interview and/or artifact review for further information.



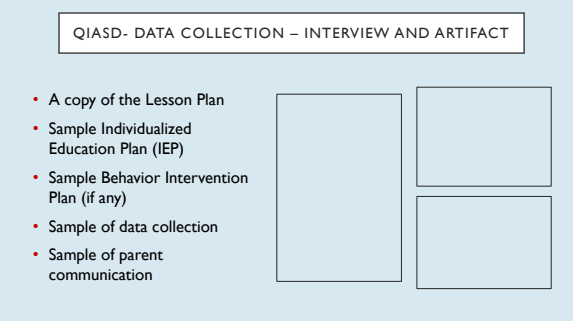
11. There are three options for gathering data during your observation period.
    a. Data on most indicators will be collected through Direct Observation – this is when the Observer is in the classroom recording scores on the rating scale based on real time observation of the quality indicators.
    b. If there is no opportunity to observe an indicator during the direct observation, then follow up with a brief interview with the teacher and/or a review of relevant artifacts in order to gather enough information to rate that item.



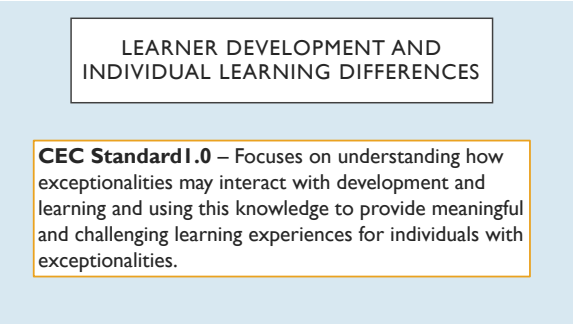
12. The QIASD identifies the seven CEC standards along with corresponding quality classroom indicators listed beneath each standard.
    a. The observer should select the relevant rating, based on the scale ranging from highly effective to unsatisfactory, and N/A for unrated items. Select one rating for each indicator.
    b. Under the rating scale, you will see the data collection method options. You may select only ONE data collection method. Please choose the method that is most representative for the indicator you are rating.
    c. Please be sure to select a rating for every indicator and a data collection method before submitting your completed form.



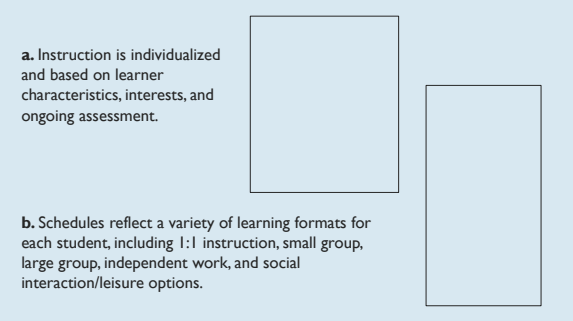
13. The seven sections of the QIASD will have a Comment Box. This space is to provide support for the ratings given to assess the presence of the quality classroom indicators. Include examples of specific evidence-based practices, activities and strategies from your observation.



14. Certain quality indicators may require an interview and/or an artifact review to determine if the indicator is present. Ask the classroom teacher for the following artifacts to be ready:
    a. A Lesson Plan
    b. An Individualized Education Plan (IEP)
    c. Behavior Intervention Plan (if any)
    d. A Sample of data collection
    e. An example of parent communication

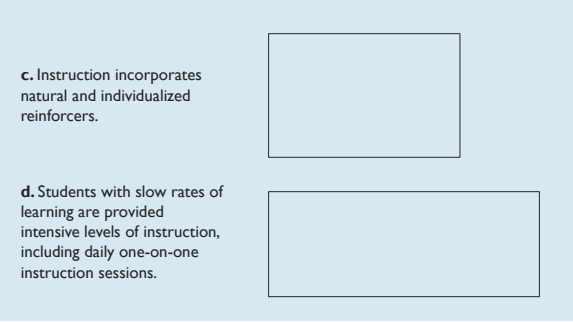


15. CEC standard 1 is Learner Development and Individual Learning Differences
    a. There are 6 quality indicators to support standard one.



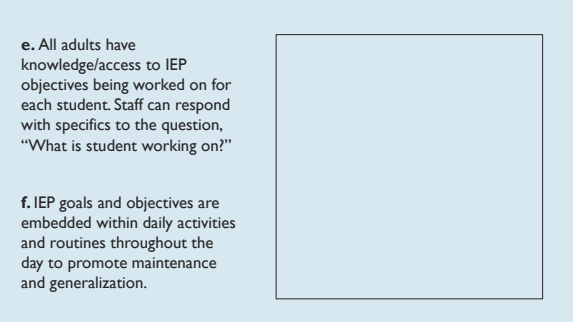
16. **a.**Instruction is individualized and based on learner characteristics, interests, and ongoing assessment.
    a. Look for evidence the teacher differentiates instruction and curriculum materials for individual learning levels, and if materials and/or rewards reflect student interests and preferences. Also, does the teacher assess student responding and adjust instruction accordingly to meet individual needs.

b.Schedules reflect a variety of learning formats for each student, including 1:1 instruction, small group, large group, independent work, and social interaction/leisure options.



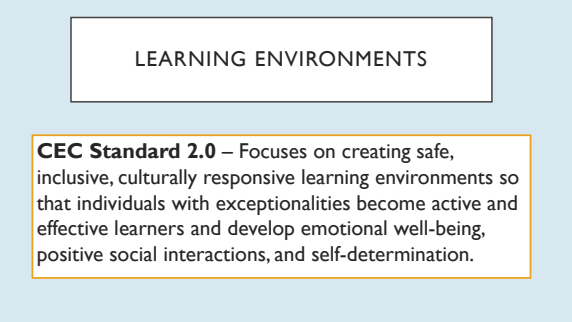
17. **c.** Instruction incorporates natural and individualized reinforcers.
    a. Look for positive feedback, short breaks, classroom reward systems, etc. And, is there evidence of choice or preference assessments to identify individualized student reinforcers.

d. Students with slow rates of learning are provided intensive levels of instruction, including daily one-on-one instruction sessions.
    b. Look for he teacher or other staff member working directly with an individual student on IEP goals and/or target curriculum content.



18. **e.** All adults have knowledge and access to IEP objectives being worked on for each student. Staff can respond with specifics to the question, "What is student working on?"
    a. Evidence for this indicator may be observed, especially if there is a written lesson plan or an iep goal sheet readily available. If not, the oberserver may need to ask staff directly what the student is working on. Do not interrupt instruction to do this. Wait until a natural break or speak to staff during a non-instructional time.

    f. IEP goals and objectives are embedded within daily activities and routines throughout the day to promote maintenance and generalization.
    b. Evidence may be found on a lesson plan, daily schedule, or during observed instruction (such a positive behavior chart being used during an academic lesson).



Adobe training continued with specific examples for all 51 indicators.

QIASD Tutorial Quiz Spring 2018

Directions:
Several scenarios are presented within the Quiz. Answer **True** or **False** for whether the Rating provided correctly represents the presence of the Quality Classroom Indicator. Once you click on the Submit button at the bottom, your Quiz will be submitted. You must get an 80% or above on this Quiz. You may re-access the Quiz multiple times with the link provided. If you have questions, please contact your professor.

1.
You are conducting an observation in a classroom with eight students, one teacher, and two paraprofessionals. The teacher and one paraprofessional are each working with small groups of three students. One paraprofessional is monitoring the other two students who are doing independent work at their desks. The three staff are in the classroom actively participating in instructional activities during your whole observation.

Based on the Quality Indicator: Staff ratio of 1 adult for every 3 students is maintained during at least 75% of the observation. Allow greater ratio if the students are included for part of the day and are not on access points.

Answer True/False: **Rating is 3: Effective**

- ○ True (1)
- ○ False (2)

REFERENCES

Aaronson, D., Barrow, L., & Sander, W. (2007). Teachers and student achievement in the Chicago public high schools. *Journal of Labor Economics*, *25*(1), 95-135.

Alexander, J. L., Ayres, K. M., & Smith, K. A. (2015). Training teachers in evidence-based practice for individuals with autism spectrum disorder: A review of the literature. *Teacher Education and Special Education*, *38*(1), 13-27.

American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME). (1999; 2014). *Standards for educational and psychological testing.* Washington DC: AERA.

American Institute for Research. (2013, March). *Improving college and career readiness for students with disabilities* (Issue Brief No. 1525). Washington, DC: Author. Retrieved from http://www.ccrscenter.org/products-resources/improving-college-and-career-readiness-students-disabilities

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). Arlington, VA: American Psychiatric Pub.

American Statistical Association (2014). *ASA statement on using value-added models for educational assessment.* Retrieved from http://www.amstat.org/asa/files/pdfs/POL-ASAVAM-Statement.pdf

Anderson, D. K., Liang, J. W., & Lord, C. (2014). Predicting young adult outcome among more and less cognitively able individuals with autism spectrum disorders. *Journal of Child Psychology and Psychiatry*, *55*(5), 485–494.

Arbuckle, J. (2008). *Amos 17.0 user's guide*. Amos Development Corporation. Chicago, IL: SPSS Inc.

Arbuckle, J. L. (2010). *IBM SPSS Amos 19 user's guide*. Crawfordville, FL: Amos Development Corporation.

Baio, J., Wiggins, L., Christensen, D.L., Maenner, M. J., Daniels, J., Warren, Z., … & Dowling, N. F. (2018). Prevalence of autism spectrum disorder among children aged 8 years — autism and developmental disabilities monitoring network, 11 sites, United States, 2014. *MMWR Surveillance Summaries, 67*(6),1–23. doi:10.15585/mmwr.ss6706a1

Barnes, G., Crowe, E., & Schaefer, B. (2007). *The cost of teacher turnover in five school districts: A pilot study*. National Commission on Teaching and America's Future.

Baxter, R. (2009). Reflective and formative metrics of relationship value: A commentary essay. *Journal of Business Research, 62*(12), 1370-1377.

Bear, G. G., Yang, C., Pell, M., & Gaskins, C. (2014). Validation of a brief measure of teachers' perceptions of school climate: Relations to student achievement and suspensions. *Learning Environments Research, 17*(3), 339-354.

Begoli, E., DeFalco, J., & Ogle, C. (2016). The promise and relevance of emerging technologies in the education of children with autism spectrum disorder. In Y. Kats (Ed.), *Supporting the education of children with autism spectrum disorders* (pp. 184-202). Chestnut Hill College: IGI Global.

Belfiore, P. J., Fritts, K. M., & Herman, B. C. (2008). The role of procedural integrity: Using self-monitoring to enhance discrete trial instruction (DTI). *Focus on Autism and Other Developmental Disabilities, 23*(2), 95-102.

Benedict, A. E., Thomas, R. A., Kimerling, J., & Leko, C. (2013). Trends in teacher evaluation: What every special education teacher should know. *Teaching Exceptional Children, 45*(5), 60-68.

Bentler, P. M., & Yuan, K. H. (1999). Structural equation modeling with small samples: Test statistics. *Multivariate Behavioral Research, 34*(2), 181-197.

Bireda, S., & Chait, R. (2011). Increasing teacher diversity: Strategies to improve the teacher workforce. *Center for American Progress*. . Retrieved from http://americanprogress.org

Bishop-Fitzpatrick, L., Hong, J., Smith, L. E., Makuch, R. A., Greenberg, J. S., & Mailick, M. R. (2016). Characterizing objective quality of life and normative outcomes in adults with autism spectrum disorder: An exploratory latent class analysis. *Journal of Autism and Developmental Disorders, 46*(8), 2707-2719.

Boe, E. E., Cook, L. H., & Sunderland, R. J. (2008). Teacher turnover: Examining exit attrition, teaching area transfer, and school migration. *Exceptional Children, 75*(1), 7-31.

Bollen, K. A., & Barb, K. H. (1981). Pearson's r and coarsely categorized measures. *American* Sociological Review, 46(2), 232-239.

Bollen, K, A, & Bauldry, S. (2011). Three Cs in measurement models: Causal indicators, composite indicators, and covariates. Psychological Methods, 16(3), 265-284.

Bollen, K. A., & Stine, R. A. (1992). Bootstrapping goodness-of-fit measures in structural equation models. *Sociological Methods & Research, 21*(2), 205-229.

Boyer, L., & Lee, C. (2001). Converting challenge to success: Supporting a new teacher of students with autism. *The Journal of Special Education, 35*(2), 75-83.

Brock, M. E., Huber, H. B., Carter, E. W., Juarez, A. P., & Warren, Z. E. (2014). Statewide assessment of professional development needs related to educating students with autism spectrum disorder. *Focus on Autism and Other Developmental Disabilities, 29*(2), 67-79.

Browder, D. M., & Cooper-Duffy, K. (2003). Evidence-based practices for students with severe disabilities and the requirement for accountability in "No Child Left Behind". *The Journal of Special Education, 37*(3), 157-163.

Brown, T. A. (2006). *Confirmatory factor analysis for applied research* (1st ed.). New York: Guilford Press.

Brown, T. A. (2014). *Confirmatory factor analysis for applied research* (2nd ed.). New York: Guilford Press.

Brownell, M. T., & Jones, N. D. (2015). Teacher evaluation in special education: Approaches, supporting research, and challenges confronted. *Journal of Special Education Leadership*, *28*(2), 63-73.

Buzick, H. M., & Jones, N. D. (2015). Using test scores from students with disabilities in teacher evaluation. *Educational Measurement, 34*(3), 28-38.

Camara, W. J. (2003). Professional testing standards: What educators need to know. In J. E. Wall & G. R. Walz (Eds.), *Measuring up: Assessment issues for teachers, counselors, and administrators* (pp. 417-425). Retrieved from https://files.eric.ed.gov/fulltext/ED480064.pdf

Cantrell, S. C., Almasi, J. F., Carter, J. C., & Rintamaa, M. (2013). Reading intervention in middle and high schools: Implementation fidelity, teacher efficacy, and student achievement. *Reading Psychology, 34*(1), 26-58.

Casabianca, J. M., McCaffrey, D. F., Gitomer, D. H., Bell, C. A., Hamre, B. K., & Pianta, R. C. (2013). Effect of observation mode on measures of secondary mathematics teaching. *Educational and Psychological Measurement, 73*(5), 757-783.

Christensen, D. L., Baio, J., Braun, K. V. N., Bilder, D., Charles, J., Constntino, J. N.,…Yeargin-Allsopp, M. (2016). Prevalence and characteristics of autism spectrum disorder among children aged 8 years—autism and developmental disabilities monitoring network, 11 sites, United States, 2012. *MMWR Surveillance Summaries, 65*(3), 1-23. Retrieved from https://www.cdc.gov/mmwr/volumes/65/ss/ss6503a1.htm#suggestedcitation

Cochran-Smith, M., & Fries, K. (2011). Teacher education for diversity. In A. F. Ball, & C. A. Tyson (Eds.), *Studying diversity in teacher education* (pp. 339-361). New York: Rowman & Littlefield Publishers.

Coggshall, J. G., Bivona, L., & Reschly, D. J. (2012). *Evaluating the effectiveness of teacher preparation programs for support and accountability*. (Research & Policy Brief No. 2483). Washington, DC: National Comprehensive Center for Teacher Quality. Retrieved from http://www.gtlcenter.org/products-resources/evaluating-effectiveness-teacher-preparation-programs-support-and-accountability

Cohen, J. (1988). *Statistical power analyses for the social sciences.* Hillsdale, NJ: Lawrence Erlbauni Associates.

Cohen, J., & Goldhaber, D. (2016). Observations on evaluating teacher performance. In J. A. Grissom, & P. Youngs (Eds.), *Improving teacher evaluation systems: Making the most of multiple measures* (pp. 8-21). New York: Teachers College Press.

Collie, R. J., Shapka, J. D., & Perry, N. E. (2012). School climate and social–emotional learning: Predicting teacher stress, job satisfaction, and teaching efficacy. *Journal of Educational Psychology, 104*(4), 1189-1204.

Coltman, T., Devinney, T. M., Midgley, D. F., & Venaik, S. (2008). Formative versus reflective measurement models: Two applications of formative measurement. *Journal of Business Research, 61*(12), 1250-1262.

Connelly, V., & Graham, S. (2009). Student teaching and teacher attrition in special education. *Teacher Education and Special Education, 32*(3), 257-269.

Conner, C. M. (2013). Commentary on two classroom observation systems: Moving toward a shared understanding of effective teaching. *School Psychology Quarterly, 28*(4), 342-346.

Cook, B. G., Carter, E. W., Cote, D. L., Kamman, M., McCarthy, T., Miller, M. L., ... & Travers, J. (2014). Evidence-based special education in the context of scarce evidence-based practices. *Teaching Exceptional Children, 47*(2), 81-84.

Cook, B. G., & Odom, S. L. (2013). Evidence-based practices and implementation science in special education. *Exceptional Children, 79*(2), 135-144.

Costello, A. B., & Osborne, J. W. (2005). Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. *Practical Assessment, Research & Evaluation, 10*(7), 1-9.

Council for Exceptional Children. (2017). *CEC specialty set validation resource manual*. Retrieved from

https://www.cec.sped.org/~/media/Files/Standards/CEC%20Initial%20and%20Advanced%20Specialty%20Sets/2017%20Specialty%20Set%20Validation%20Manual.pdf

Council for Exceptional Children. (2017). News from CEC: High-Leverage Practices in Special Education. *Teaching Excpetional Children*, 49(5), 355-360.

Council for Exceptional Children. (2012). *The Council for Exceptional Children's position on special education teacher evaluation*. Retrieved from https://www.cec.sped.org/~/media/Files/Policy/CEC%20Professional%20Policies%20and%20Positions/Position_on_Special_Education_Teacher_Evaluation_Background.pdf

Council for Exceptional Children. (2015). *What every special educator must know: Professional ethics and standards* (7th ed.). Arlington, VA: Author.

Crimmins, D. B., Durand, V. M., Theurer-Kaufman, K., & Everett, J. (2001). *Autism program quality indicators: A self-review and quality improvement guide for schools and programs serving students with autism spectrum disorders*. Retrieved from https://eric.ed.gov/?q=ED458767&id=ED458767

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*(3), 297-334.

Cronbach, L. J., & Meele, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52,* 281-302.

Crowe, C. C., Rivers, S. E., & Bertoli, M. C. (2017). Mind the gap: accountability, observation and special education. *Assessment in Education: Principles, Policy & Practice, 24*(1), 21-43.

Danielson, C. (2011). *Enhancing professional practice: A framework for teaching*. Alexandria, VA: ASCD.

Danielson, C. (2011). Evaluations that help teachers learn. *Educational Leadership, 68*(4), 35-39.

Danielson, C. (2013). *The framework for teaching: Evaluation instrument.* Princeton, NJ: Danielson Group.

Darling-Hammond, L. (2015). Can value added add value to teacher evaluation?. *Educational Researcher*, *44*(2), 132-137.

Darling-Hammond, L. (2010). Evaluating teacher effectiveness: How teacher performance assessments can measure and improve teaching. *Center for American Progress*. Retrieved from ERIC database. (ED535859).

Darling-Hammond, L. (2014). One piece of the whole: Teacher evaluation as part of a comprehensive system for teaching and learning. *American Educator*, *38*(1), 4.

Darling-Hammond, L. (2000). Solving the dilemmas of teacher supply, demand, and standards: How we can ensure a competent, caring, and qualified teacher for every child. *National Commission on Teaching & America's Future.* Retrieved from ERIC database. (ED463337).

Darling-Hammond, L., Amrein-Beardsley, A., Haertel, E., & Rothstein, J. (2012). Evaluating teacher evaluation. *Phi Delta Kappan*, *93*(6), 8-15.

Dee, T. S., & Wyckoff, J. (2015). Incentives, selection, and teacher performance: Evidence from IMPACT. *Journal of Policy Analysis and Management, 34*(2), 267-297.

Diamantopoulos, A., & Winklhofer, H. M. (2001). Index construction with formative indicators: An alternative to scale development. *Journal of Marketing Research, 38*(2), 269-277.

District of Columbia Public Schools. (2014). *IMPACT: The District of Columbia Public Schools Effectiveness Assessment System for School-Based Personnel*. Retrieved from http://dcps.dc.gov/node/979742

Doherty, K. M., & Jacobs, S. (2015) *State of the states 2015: Evaluating teaching, leading, and learning.* Washington, DC: National Council on Teacher Quality. Retrieved from https://www.nctq.org/dmsView/StateofStates2015

Eaves, L., & Ho, H. (2008). Young adult outcome of autism spectrum disorders. *Journal of Autism and Developmental Disorders, 38*(4), 739–747.

Economic Policy Institute. (2010, August). *Problems with the use of student test scores to evaluate teachers.* (Issue Brief No. 278). Washington, DC: Author. Retrieved from https://files.eric.ed.gov/fulltext/ED516803.pdf

Elliott, E. M., Isaacs, M. L., & Chugani, C. D. (2010). Promoting self-efficacy in early career teachers: A principal's guide for differentiated mentoring and supervision. *Florida Journal of Educational Administration & Policy*, *4*(1), 131-146.

Emery, D. W., & Vandenberg, B. (2010). Special education teacher burnout and ACT. *International Journal of Special Education, 25*(3), 119-131.

Every Student Succeeds Act of 2015, 20 U.S.C § 6301 *et seq*.

Farley, M., & McMahon, B. (2014). Range of outcomes and challenges in middle and later life. In F. Volkmar, B. Reichow, & J. McPartland (Eds.), *Adolescents and adults with autism spectrum disorders* (pp. 211-238). New York, NY: Springer.

Farley, M.A., McMahon, W.M., Fombonne, E., Jenson, W.R., Miller, J., Gardner, M., & Coon, H. (2009). Twenty-year outcome for individuals with autism and average or near average cognitive abilities. *Autism Research, 2*(2), 109–118.

Field, A. (2013). *Discovering statistics using IBM SPSS statistics*. Washington DC: Sage.

Fein, D., Barton, M., Eigsti, I., Kelley, E., Naigles, L., Schultz, R., . . . Tyson, K. (2013). Optimal outcome in individuals with a history of autism. *Journal of Child Psychology and Psychiatry, 54*(2), 195–205.

Figlio, D. N., & Kenny, L. W. (2007). Individual teacher incentives and student performance. *Journal of Public Economics, 91*(5), 901-914.

Floyd, F. J., & Widaman, K. F. (1995). Factor analysis in the development and refinement of clinical assessment instruments. *Psychological assessment, 7*(3), 286.

Fornell, C., & Bookstein, F. L. (1982). Two structural equation models: LISREL and PLS applied to consumer exit-voice theory. *Journal of Marketing Research*, *19*(4), 440-452.

Gall, M. D., Gall, J. P., & Borg, W. R. (2007). *Educational research: An introduction*. New York: Person Education.

Gansle, K. A., Noell, G. H., Grandstaff-Beckers, G., Stringer, A., Roberts, N., & Burns, J. M. (2015). Value-added assessment of teacher preparation: Implications for special education. *Intervention in School and Clinic, 51*(2), 106-111.

Garbacz, S. A., McIntyre, L. L., & Santiago, R. T. (2016). Family involvement and parent–teacher relationships for students with autism spectrum disorders. *School Psychology Quarterly*, *31*(4), 478.

Gehrke, R. S., & McCoy, K. (2007). Sustaining and retaining beginning special educators: It takes a village. *Teaching and Teacher Education, 23*(4), 490-500.

Goe, L. (2007). *The link between teacher quality and student outcomes.* Washington, DC: National Comprehensive Center for Teacher Quality. Retrieved from http://www.gtlcenter.org/products-resources/link-between-teacher-quality-and-student-outcomes-research-synthesis

Goe, L., Bell, C., & Little, O. (2008). *Approaches to evaluating teacher effectiveness: A research synthesis.* Washington, DC: National Comprehensive Center for Teacher Quality. Retrieved from http://www.gtlcenter.org/products-resources/approaches-evaluating-teacher-effectiveness-research-synthesis

Goe, L., & Croft, A. (2009). *Methods of evaluating teacher effectiveness. (Research-To-Practice Brief No. 3307).* Washington, DC: National Comprehensive Center for Teacher Quality.

Goldhaber, D. (2010). *When the stakes are high, can we rely on value-added? Exploring the use of value-added models to inform teacher workforce decisions.* Washington DC: Center for American Progress. Retrieved from https://www.americanprogress.org/issues/default/reports/2010/12/01/8720/when-the-stakes-are-high-can-we-rely-on-value-added/

Goldhaber, D. (2015). Exploring the potential of value-added performance measures to affect the quality of the teacher workforce. *Educational Researcher*, *44*(2), 87-95.

Grossman, P., Greenberg, S., Hammerness, K., Cohen, J., Alston, C., & Brown, M. (2009). *Development of the protocol for language arts teaching observation (PLATO).* In Annual meeting of the American Educational Research Association, San Diego, CA.

Grossman, P. & McDonald, M. (2008). Back to the future: Directions for research in teaching and teacher education. *American Educational Research Journal, 45*(1), 184-205.

Guadagnoli, E., & Velicer, W. F. (1988). Relation to sample size to the stability of component patterns. *Psychological Bulletin, 103*(2), 265.

Hair Jr, J. F., Babin, B. J., & Krey, N. (2017). Covariance-based structural equation modeling in the Journal of Advertising: Review and recommendations. *Journal of Advertising*, *46*(1), 163-177.

Hair Jr, J. F., Hult, G. T. M., Ringle, C., & Sarstedt, M. (2016). *A primer on partial least squares structural equation modeling (PLS-SEM)*. Thousand Oaks, CA: Sage Publications.

Hair, J. F., Hult, G. T. M., Ringle, C. M., Sarstedt, M., & Thiele, K. O. (2017). Mirror, mirror on the wall: A comparative evaluation of composite-based structural equation modeling methods. *Journal of the Academy of Marketing Science*, *45*(5), 616-632.

Hair, J. F., Ringle, C. M., & Sarstedt, M. (2013). Partial least squares structural equation modeling: Rigorous applications, better results and higher acceptance. *Long Range Planning*, *46*(1-2), 1-12.

Hair Jr, J. F., Sarstedt, M., Hopkins, L., & G. Kuppelwieser, V. (2014). Partial least squares structural equation modeling (PLS-SEM): An emerging tool in business research. *European Business Review*, *26*(2), 106-121.

Hallinger, P., Heck, R. H., & Murphy, J. (2014). Teacher evaluation and school improvement: An analysis of the evidence. *Educational Assessment, Evaluation and Accountability*, *26*(1), 5-28.

Hamre, B. K., Pianta, R. C., Mashburn, A. J., & Downer, J. T. (2007). Building a science of classrooms: Application of the CLASS framework in over 4,000 US early childhood and elementary classrooms. *Foundation for Childhood Development, 30*, 1-35.

Harrington, D. (2009). *Confirmatory factor analysis.* New York: Oxford University Press.

Harris, D. N., & Herrington, C. D. (2015). Value added meets the schools: The effects of using test-based teacher evaluation on the work of teachers and leaders [Special issue]. *Educational Researcher, 44*(2), 71-76.

Hart, J. E., & Whalon, K. J. (2011). Creating social opportunities for students with autism spectrum disorder in inclusive settings. *Intervention in School and Clinic, 46*(5), 273-279.

Ho, A. D., & Kane, T. J. (2013). *The reliability of classroom observations by school personnel*. Seattle, WA: Measures of Effective Teaching Project, Bill and Melinda Gates Foundation.

Holdheide, L. (2015). Same debate, new opportunity: Designing teacher evaluation systems that promote and support educators in practices that advance all students' learning. *Journal of Special Education Leadership, 28*(2), 74-81.

Hoyle, R. H. (2000). Confirmatory factor analysis. In H. E. A. Tinsley & S. D. Brown (Eds.), *Handbook of applied multivariate statistics and mathematical modeling* (pp. 465-497). New York: Academic Press.

Howlin, P., Goode, S., Hutton, J., & Rutter, M. (2004). Adult outcome for children with autism. *Journal of Child Psychology and Psychiatry, and Allied Disciplines, 45*(2), 212–229.

Hill, H. C., Blunk, M., Charalambous, C., Lewis, J., Phelps, G. C., Sleep, L., & Ball, D. L. (2008). Mathematical knowledge for teaching and the mathematical quality of instruction: An exploratory study. *Cognition and Instruction*, *26*(4), 430-511.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6,* 1–55.

Hu, L. T., Bentler, P. M., & Kano, Y. (1992). Can test statistics in covariance structure analysis be trusted?. *Psychological Bulletin, 112*(2), 351.

Hutcheson, G. D., & Sofroniou, N. (1999). *The multivariate social scientist: Introductory statistics using generalized linear models*. London: Sage.

Individuals with Disabilities Education Improvement Act of 2004, 20 U.S.C. § 1401 et seq.

Iovannone, R., Dunlap, G., Huber, H., & Kincaid, D. (2003). Effective educational practices for students with autism spectrum disorders. *Focus on Autism and Other Developmental Disabilities, 18*(3), 150-165.

Jackson, D. L. (2003). Revisiting sample size and number of parameter estimates: Some support for the N:q hypothesis. *Structural Equation Modeling, 10*(1), 128-141.

Jacob, B., Rockoff, J. E., Taylor, E. S., Lindy, B., & Rosen, R. (2016). *Teacher applicant hiring and teacher performance: Evidence from DC public schools* (No. w22054). National Bureau of Economic Research.

Jarvis, C. B., MacKenzie, S. B., & Podsakoff, P. M. (2003). A critical review of construct indicators and measurement model misspecification in marketing and consumer research. *Journal of Consumer Research 30*(2), 199-218.

Jennett, H. K., Harris, S. L., & Mesibov, G. B. (2003). Commitment to philosophy, teacher efficacy, and burnout among teachers of children with autism. *Journal of Autism and Developmental Disorders, 33*(6), 583-593.

Johnson, E. S., Crawford, A., Moylan, L. A., Ford, J. W. (2016). Issues in evaluating special education teachers: Challenges and current perspectives. *Texas Education Review, 4*(1), 71-83.

Johnson, E., & Semmelroth, C. L. (2014). Special education teacher evaluation: Why it matters, what makes it challenging, and how to address these challenges. *Assessment for Effective Intervention, 39*(2), 71-82.

Jones, N. D., & Brownell, M. T. (2014). Examining the use of classroom observations in the evaluation of special education teachers. *Assessment for Effective Intervention, 39*(2), 112-124.

Jones, N. D., Buzick, H. M., & Turkan, S. (2013). Including students with disabilities and English learners in measures of educator effectiveness. *Educational Researcher, 42*(4), 234-241.

Jöreskog, K. G. (1978). Structural analysis of covariance and correlation matrices. *Psychometrika*, *43*(4), 443-477.

Kaiser, H. F. (1970). A second generation little jiffy. *Psychometrika, 35*(4), 401-415.

Kaiser, H. F. (1974). An index of factorial simplicity. *Psychometrika*, *39*(1), 31-36.

Kane, M. (2006). Content-related validity evidence in test development. In T. M. Haladyna, & S. M. Downing (Eds.), *Handbook of test development* (pp. 131-153). Mahwah, NJ: Lawrence Erlbaum Assoc.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*(1), 1-73.

Kane, T. J., McCaffrey, D. F., Miller, T., & Staiger, D. O. (2013). Have we identified effective teachers? Validating measures of effective teaching using random assignment. Seattle, WA: Measures of Effective Teaching Project, Bill and Melinda Gates Foundation.

Kane, T., Kerr, K., & Pianta, R. (2014). *Designing teacher evaluation systems: New guidance from the measures of effective teaching project*. San Francisco, CA: John Wiley & Sons.

Kane, T. J., & Staiger, D. O. (2012). Gathering feedback for teaching: Combining high-quality observations with student surveys and achievement gains. Seattle, WA: Measures of Effective Teaching Project, Bill and Melinda Gates Foundation.

Kraft, M. A., & Gilmour, A. F. (2016). Can principals promote teacher development as evaluators? A case study of principals' views and experiences. *Educational Administration Quarterly*, *52*(5), 711-753.

Kraft, M. A., & Gilmour, A. F. (2017). Revisiting the widget effect: Teacher evaluation reforms and the distribution of teacher effectiveness. *Educational Researcher, 46*(5), 234-249.

Kearns, J. F., Kleinert, H. L., Thurlow, M. L., Gong, B., & Quenemoen, R. (2015). Alternate assessments as one measure of teacher effectiveness: Implications for our field. *Research and Practice for Persons with Severe Disabilities, 40*(1), 20-35.

Keigher, A. (2010). *Teacher attrition and mobility: Results from the 2008-09 teacher follow-up survey.* (NCES 2010-353). U.S. Department of Education. Washington, DC: National Center for Education Statistics. Retrieved from https://nces.ed.gov/pubs2010/2010353.pdf

Kersting, N. B., Chen, M. K., & Stigler, J. W. (2013). Value-added teacher estimates as part of teacher evaluations: Exploring the effects of data and model specifications on the stability of teacher value-added scores. *Education Policy Analysis Archives, 21*(7), 2-39.

Kiuhara, S. A., & Kratochwill, T. R. (2017). Research methods in special education. In J. M. Kauffman, D. P. Hallahan, P. C. Pullen (Eds.), *Handbook of special education* (pp. 105-107). New York: Routledge.

Kline, R. B. (2010). *Principles and practice of structural equation modeling* (3rd ed.). New York: Guilford Press.

Kline, R. B. (2015). *Principles and practice of structural equation modeling (*4th ed.). New York: Guilford Press.

Kline, T. (2005). *Psychological testing: A practical approach to design and evaluation*. Thousand Oaks: Sage.

Kock, N. (2014). Using data labels to discover moderating effects in PLS-based structural equation modeling. *International Journal of e-Collaboration, 10*(4), 1-16.

Kock, N. (2015). *WarpPLS 5.0 user manual*. Laredo, TX: ScriptWarp Systems.

Kock, N., & Mayfield, M. (2015). PLS-based SEM algorithms: The good neighbor assumption, collinearity, and nonlinearity. *Information Management and Business Review*, 7(2), 113-130.

Koegel, L. K., Koegel, R. L., Ashbaugh, K., & Bradshaw, J. (2014). The importance of early identification and intervention for children with or at risk for autism spectrum disorders. *International Journal of Speech-Language Pathology, 16*(1), 50-56.

Gall, M. D., Gall, J. P., & Borg, W. R. (2007). *Educational research: An introduction*. New York: Person Education.

Garbacz, S. A., McIntyre, L. L., & Santiago, R. T. (2016). Family involvement and parent–teacher relationships for students with autism spectrum disorders. *School Psychology Quarterly*, *31*(4), 478-490.

Goldhaber, D. (2010). When the stakes are high, can we rely on value-added? Exploring the use of value-added models to inform teacher workforce decisions. Washington, DC: Center for American Progress. Retrieved from ERIC database. (ED535870).

Goldring, R., Taie, S., and Riddles, M. (2014). *Teacher attrition and mobility: Results from the 2012–13 teacher follow-up survey* (NCES 2014-077). U.S. Department of Education. Washington, DC: National Center for Education Statistics. Retrieved from http://nces.ed.gov/pubsearch.

Lash, A., Tran, L., & Huang, M. (2016). *Examining the validity of ratings from a classroom observation instrument for use in a district's teacher evaluation system.* Washington, DC: Regional Educational Laboratory West.

Leaf, J. B., Leaf, R., McCRAY, C., Lamkns, C., Taubman, M., McEACHIN, J., & Cihon, J. H. (2017). A preliminary analysis of a behavioral classrooms needs assessment. *International Electronic Journal of Elementary Education, 9*(2), 385-404.

Lei, M., & Lomax, R. G. (2005). The effect of varying degrees of nonnormality in structural equation modeling. *Structural Equation Modeling*, *12*(1), 1-27.

Leigh, A. (2010). Estimating teacher effectiveness from two-year changes in students' test scores. *Economics of Education Review, 29*(3), 480-488.

Leigh, J. P., & Du, J. (2015). Brief report: Forecasting the economic burden of autism in 2015 and 2025 in the United States. *Journal of Autism and Developmental Disorders, 45*(12), 4135-4139.

Levy, A., & Perry, A. (2011). Outcomes in adolescents and adults with autism: A review of the literature. *Research in Autism Spectrum Disorders, 5*(4), 1271–1282.

Light, J., & Mcnaughton, D. (2015). Designing AAC research and intervention to improve outcomes for individuals with complex communication needs. *Augmentative and Alternative Communication, 31*(2), 85-96.

Linstead, E., Dixon, D. R., French, R., Granpeesheh, D., Adams, H., German, R., . . . Kornack, J. (2017). Intensity and learning outcomes in the treatment of children with autism spectrum disorder. *Behavior Modification, 41*(2), 229-252.

Little, O., Goe, L., & Bell, C. (2009). A practical guide to evaluating teacher effectiveness. *National Comprehensive Center for Teacher Quality*. Retrieved from http://www.gtlcenter.org/sites/default/files/docs/practicalGuide.pdf

Lowry, P. B., & Gaskin, J. (2014). Partial least squares (PLS) structural equation modeling (SEM) for building and testing behavioral causal theory: When to choose it and how to use it. *IEEE Transactions on Professional Communication, 57*(2), 123-146.

Lyons, J., Cappadocia, M. C., & Weiss, J. A. (2011). Brief report: Social characteristics of students with autism spectrum disorders across classroom settings. *Journal on Developmental Disabilities, 17*, 77–82.

MacCallum, R. C., & Austin, J. T. (2000). Applications of structural equation modeling in psychological research. *Annual Review of Psychology, 51*(1), 201-226.

MacCallum, R. C., Browne, M. W., & Sugawara, H. M. (1996). Power analysis and determination of sample size for covariance structure modeling. *Psychological methods, 1*(2), 130.

MacKenzie, S. B., Podsakoff, P. M., & Jarvis, C. B. (2005). The problem of measurement model misspecification in behavioral and organizational research and some recommended solutions. *Journal of Applied Psychology, 90*(4), 710-730.

McDonald, M., Kazemi, E., & Kavanagh, S. S. (2013). Core practices and pedagogies of teacher education: A call for a common language and collective activity. *Journal of Teacher Education*, *64*(5), 378-386.

McLeskey, J., Council for Exceptional Children, & Collaboration for Effective Educator Development, Accountability and Reform. (2017). *High-leverage practices in special education*. Arlington, VA: Council for Exceptional Children.

McLeskey, J., & Brownell, M. (2015). High-leverage practices and teacher preparation in special education (Document No. PR-1). Retrieved from University of Florida, Collaboration for Effective Educator, Development, Accountability, and Reform Center website: http://ceedar.education.ufl.edu/tools/best-practice-review/

Marder, M., & Walkington, C. (2012). *UTeach Teacher Observation Protocol*. Retrieved from https://uteach.utexas.edu/sites/default/files/UTOP_Paper_Non_Anonymous_4_3_2011.pdf

Marsh, H. W., & Hau, K. T. (1999). Confirmatory factor analysis: Strategies for small sample sizes. *Statistical Strategies for Small Sample Research, 1*, 251-284.

Marshall, J. C., Smart, J., & Alston, D. M. (2016). Development and validation of Teacher Intentionality of Practice Scale (TIPS): A measure to evaluate and scaffold teacher effectiveness. *Teaching and Teacher Education, 59,* 159-168.

McCaffrey, D. F., & Buzick, H. M. (2014). *Is value-added accurate for teachers of students with disabilities? What we know series: Value-added methods and applications* (Knowledge Brief No. 14). Stanford, CA: Carnegie Foundation for the Advancement of Teaching. Retrieved from http://www.carnegieknowledgenetwork.org/briefs/teacher_disabilities/

McLeskey, J., & Billingsley, B. S. (2008). How does the quality and stability of the teaching force influence the research-to-practice gap? A perspective on the teacher shortage in special education. *Remedial and Special Education, 29*(5), 293-305.

McIntosh, R., Vaughn, S., Schumm, J. S., Haager, D., & Lee, O. (1993). Observations of students with learning disabilities in general education classrooms. *Exceptional Children, 60*(3), 249-261.

Mesibov, G., Howley, M., & Naftel, S. (2015). *Accessing the Curriculum for Learners with Autism Spectrum Disorders: Using the TEACCH programme to help inclusion*. New York, NY: Routledge.

Messick, S. (1989). Validity. In R. L. Linn (ed.), *Educational Measurement* (pp. 13-103). New York: Macmillan.

Millett, C. M., Stickler, L. M., Payne, D. G., & Dwyer, C. A. (2007). A culture of evidence: Critical features of assessments for postsecondary student learning. *Educational Testing Service.* Retrieved from ERIC database. (ED499995).

Morrier, M. J., Hess, K. L., & Heflin, L. J. (2011). Teacher training for implementation of teaching strategies for students with autism spectrum disorders. *Teacher Education and Special Education, 34*(2), 119-132.

Nagy, C. J., & Wang, N. (2007). The alternate route teachers' transition to the classroom: Preparation, support, and retention. *NASSP Bulletin, 91*(1), 98-113.

National Commission on Teaching and America's Future. (2007). *Policy brief: The high cost of teacher turnover*. Washington, DC: NCTAF. Retrieved from https://nctaf.org/wp-content/uploads/2012/01/NCTAF-Cost-of-Teacher-Turnover-2007-policy-brief.pdf

National Research Council. (2001). Educating children with autism. Committee on educational interventions for children with autism. In C. Lord, & J. P. McGee (Eds.). *Division of behavioral and social sciences and education*. Washington, DC: National Academy Press.

Nevitt, J., & Hancock, G. R. (2001). Performance of bootstrapping approaches to model test statistics and parameter standard error estimation in structural equation modeling. *Structural Equation Modeling, 8*(3), 353-377.

No Child Left Behind Act of 2001, 20 U.S.C § 6301 *et seq.*

Nunnally, J. (1978). *Psychometric methods*. New York: McGraw-Hill.

Odom, S. L., Brantlinger, E., Gersten, R., Horner, R. H., Thompson, B., & Harris, K. R. (2005). Research in special education: Scientific methods and evidence-based practices. *Exceptional children, 71*(2), 137-148.

Odom, S. L., Collet-Klingenberg, L., Rogers, S. J., & Hatton, D. D. (2010). Evidence-based practices in interventions for children and youth with autism spectrum

disorders. *Preventing School Failure: Alternative Education for Children and Youth, 54*(4), 275-282.

Odom, S. L., Cox, A. W., Brock, M. E., & National Professional Development Center on ASD. (2013). Implementation science, professional development, and autism spectrum disorders. *Exceptional Children, 79*(2), 233-251.

Odom, S. L., & Wong, C. (2015). Connecting the dots: Supporting students with autism spectrum disorder. *American Educator, 39*(2), 12.

Pearl, C. E., Vasquez III, E., Marino, M. T., Wienke, W., Donehower, C., Gourwitz, J., . . . Duerr, S. R. (2017). Establishing content validity of the Quality Indicators for Classrooms Serving Students with Autism Spectrum Disorders instrument. *Teacher Education and Special Education.* Advance online publication. doi:10.1177/0888406416687814.

Pellicano, E. (2012). Do autistic symptoms persist across time? Evidence of substantial change in symptomatology over a 3-year period in cognitively able children with autism. American *Journal on Intellectual and Developmental Disabilities, 117*(2), 156–166.

Pennsylvania Department of Education. (2014). *Pennsylvania Teacher Effectiveness Evaluation System.* Retrieved from https://www.nctq.org/docs/Educator_Effectiveness_Administrative_Manual.pdf

Petter, S., Straub, D., & Rai, A. (2007). Specifying formative constructs in information systems research. *MIS Quarterly, 31*(4), 623-656.

Pianta, R. C., & Hamre, B. K. (2009). Conceptualization, measurement, and improvement of classroom processes: Standardized observation can leverage capacity. *Educational Researcher, 38*(2), 109-119.

Pianta, R. C., Hamre, B. K., & Mintz, S. L. (2012). *Classroom Assessment Scoring System: Secondary Class*. Charlotesville, VA: Teachstone.

Plash, S. H., & Piotrowski, C. (2006). Impact of the No Child Left Behind Act in Alabama. *Journal of Instructional Psychology, 33*(3), 223-227.

Reddy, L. A., Fabiano, G., Dudek, C. M., & Hsu, L. (2013). Development and construct validity of the Classroom Strategies Scale-Observer Form. *School Psychology Quarterly, 28*(4), 317-341.

Rigdon, E. E. (2016). Choosing PLS path modeling as analytical method in European management research: A realist perspective. *European Management Journal, 34*(6), 598-605.

Rockoff, J. E. (2004). The impact of individual teachers on student achievement: Evidence from panel data. *The American Economic Review, 94*(2), 247-252.

Rogers, S. J., & Vismara, L. A. (2008). Evidence-based comprehensive treatments for early autism. Journal of Clinical Child & Adolescent Psychology, 37(1), 8-38.

Roux, A. M., Shattuck, P. T., Cooper, B. P., Anderson, K. A., Wagner, M., & Narendorf, S. C. (2013). Postsecondary employment experiences among young adults with an autism spectrum disorder. *Journal of the American Academy of Child & Adolescent Psychiatry, 52*(9), 931-939.

Roux, A. M., Shattuck, P. T., Rast, J. E., Rava, J. A., & Anderson, K.A. (2015) *National autism indicators report: Transition into young adulthood.* Philadelphia, PA: Life Course Outcomes Research Program, A.J. Drexel Autism Institute, Drexel University.

Ryan, S. V., Nathaniel, P., Pendergast, L. L., Saeki, E., Segool, N., & Schwing, S. (2017). Leaving the teaching profession: The role of teacher stress and educational accountability policies on turnover intent. *Teaching and Teacher Education, 66*, 1-11.

Saeki, E., Pendergast, L., Segool, N. K., & Nathaniel, P. (2015). Potential psychosocial and instructional consequences of the common core state standards: Implications for research and practice. *Contemporary School Psychology, 19*(2), 89-97.

Sansosti, F. J., & Sansosti, J. M. (2013). Effective school-based service delivery for students with autism spectrum disorders: Where we are and where we need to go. *Psychology in the Schools, 50*(3), 229-244.

Sarstedt, M., Hair, J. F., Ringle, C. M., Thiele, K. O., & Gudergan, S. P. (2016). Estimation issues with PLS and CBSEM: Where the bias lies!. *Journal of Business Research, 69*(10), 3998-4010.

Sarstedt, M., Ringle, C. M., Smith, D., Reams, R., & Hair Jr, J. F. (2014). Partial least squares structural equation modeling (PLS-SEM): A useful tool for family business researchers. *Journal of Family Business Strategy, 5*(1), 105-115.

Sass, D. A., Seal, A. K., & Martin, N. K. (2011). Predicting teacher retention using stress and support variables. *Journal of Educational Administration, 49*(2), 200-215.

Satorra, A., & Bentler, P. M. (1988, August). Scaling corrections for chi-square statistics in covariance structure analysis. *Proceedings of the American Statistical Association, 1*, 308-313.

Semmelroth, C. L., & Johnson, E. (2014). Measuring rater reliability on a special education observation tool. *Assessment for Effective Intervention, 39*(3), 131-145.

Shattuck, P. T., Narendorf, S. C., Cooper, B., Sterzing, P. R., Wagner, M., & Taylor J. L. (2012). Postsecondary education and employment among youth with an autism spectrum disorder. *Pediatrics, 129*(6), 1-8.

Shogren, K. A., & Plotner, A. J. (2012). Transition planning for students with intellectual disability, autism, or other disabilities: Data from the National Longitudinal Transition Study-2. *Intellectual and Developmental Disabilities, 50*(1), 16-30.

Shyman, E. (2012). Teacher education in autism spectrum disorders: A potential blueprint. *Education and Training in Autism and Developmental Disabilities, 47*(2), 187-197.

Simpson, R. L. (2005). Evidence-based practices and students with autism spectrum disorders. *Focus on Autism and Other Developmental Disabilities*, *20*(3), 140-149.

Smith, T., & Iadarola, S. (2015). Evidence base update for autism spectrum disorder. *Journal of Clinical Child & Adolescent Psychology, 44*(6), 897-922.

Spencer, V. G., Evmenova, A. S., Boon, R. T., & Hayes-Harris, L. (2014). Review of research-based interventions for students with autism spectrum disorders in content area instruction: Implications and considerations for classroom practice. *Education and Training in Autism and Developmental Disabilities, 49*(3), 331-353.

Spooner, F., Knight, V. F., Browder, D. M., & Smith, B. R. (2012). Evidence-based practice for teaching academics to students with severe developmental disabilities. *Remedial and Special Education, 33*(6), 374-387.

Spooner, F., McKissick, B. R., & Knight, V. F. (2017). Establishing the state of affairs for evidence-based practices in students with severe disabilities. *Research and Practice for Persons with Severe Disabilities*, 42(1), 8-18.

Stanovich, P. J., & Jordan, A. (1998). Canadian teachers' and principals' beliefs about inclusive education as predictors of effective teaching in heterogeneous classrooms. *The Elementary School Journal, 98*(3), 221-238.

Steinberg, M. P., & Donaldson, M. L. (2016). The new educational accountability: Understanding the landscape of teacher evaluation in the post-NCLB era. *Education Finance and Policy, 11*(3), 340-359.

Steinberg, M. P., & Garrett, R. (2016). Classroom composition and measured teacher performance: What do teacher observation scores really measure?. *Educational Evaluation and Policy Analysis, 38*(2), 293-317.

Steinbrecher, T. D., Selig, J. P., Cosbey, J., & Thorstensen, B. I. (2014). Evaluating special educator effectiveness: Addressing issues inherent to value-added modeling. *Exceptional Children, 80*(3), 323-336.

Stempien, L. R., & Loeb, R. C. (2002). Differences in job satisfaction between general education and special education teachers: Implications for retention. *Remedial and Special Education, 23*(5), 258-267.

Stevens, J. P. (2002). *Applied multivariate statistics for the social sciences*. Mahwah, NJ: Lawrence Erlbaum.

Sun, M., Saultz, A., & Ye, Y. (2017). Federal policy and the teacher labor market: Exploring the effects of NCLB school accountability on teacher turnover. *School Effectiveness and School Improvement, 28*(1), 102-122.

Sutcher, L., Darling-Hammond, L., & Carver-Thomas, D. (2016). *A coming crisis in teaching? Teacher supply, demand, and shortages in the U.S.*. Palo Alto, CA: Learning Policy Institute. Retrieved from https://learningpolicyinstitute.org/sites/default/files/product-files/A_Coming_Crisis_in_Teaching_REPORT.pdf

Sutera, S., Pandey, J., Esser, E., Rosenthal, M., Wilson, L., Barton, M., . . . Fein, D. (2007). Predictors of optimal outcome in toddlers diagnosed with autism spectrum disorders. *Journal of Autism and Developmental Disorders, 37*(1), 98–107.

Tabachnick, B. G., & Fidell, L. S. (2013). *Using multivariate statistics* (6th ed.). Boston: Pearson.

Tandy, J. A., Whitford, D. K., & Hirth, M. A. (2016). A review of special education teacher (SET) evaluation practices. *Journal of Ethical Educational Leadership, 3*(5), 1-16.

Taylor, J. L., & Seltzer, M. M. (2011). Employment and post-secondary educational activities for young adults with autism spectrum disorders during the transition to adulthood. *Journal of Autism and Developmental Disorders, 41*(5), 566-574.

Tsai, S. F., Cheney, D., & Walker, B. (2013). Preliminary psychometrics of the participatory evaluation and expert review for classrooms serving students with emotional/behavioral disabilities (PEER-EBD). *Behavioral Disorders, 38*(3), 137-153.

Turnbull, C., & Knapp, J. (2017). *A complete ABA curriculum for individuals on the autism spectrum with a developmental age of 7 years up to young adulthood: A step-by-step treatment (A journey of development using ABA).* Philadelphia, PA: Jessica Kingsley Publishers.

U.S. Department of Education. (2009). *Race to the Top program: Executive summary*. Washington, DC: Author. Retrieved from http://www2.ed.gov/programs/racetothetop/executive- summary.pdf

U.S. Department of Education, National Center for Education Statistics. (2016). *Fast facts: Students with disabilities*. Retrieved from https://nces.ed.gov/fastfacts/display.asp?id=64

U.S. Department of Education, Office of Innovation and Improvement. (2017). *Teacher and School Leader Incentive Program*. Retrieved from https://innovation.ed.gov/what-we-do/teacher-quality/teacher-and-school-leader-incentive-program/

U.S. Department of Education, Office of Planning, Evaluation and Policy Development. (2016). *Study of emerging teacher evaluation systems.* Washington, DC: Policy and Program Studies Service. Retrieved from https://www2.ed.gov/rschstat/eval/teaching/emerging-teacher-evaluation/report.pdf

U.S. Department of Education, Office of Postsecondary Education. (2015). *Alternative teacher preparation programs*. Retrieved from https://title2.ed.gov/Public/44110_Title_II_Issue_Brief_Altn_TPP.pdf

U.S. Department of Education, Office of Postsecondary Education. (2015). *Higher education act title II reporting system*. Retrieved from: https://title2.ed.gov/Public/44077_Title_II_Issue_Brief_Enrollment.pdf

U.S. Department of Education, Office of Postsecondary Education. (2016). *Teacher shortage areas nationwide listing, 1990-1991 through 2016-2017*. Retrieved from https://www2.ed.gov/about/offices/list/ope/pol/tsa.pdf

U.S. Department of Education, Office of Special Education and Rehabilitative Services, Office of Special Education Programs. (2015). *37th Annual Report to Congress on the Implementation of the Individuals with Disabilities Education Act.* Washington, DC: Author. Retrieved from https://www2.ed.gov/about/reports/annual/osep/index.html

U.S. Department of Education, Office of Special Education and Rehabilitative Services, Office of Special Education Programs. (2016). *38th Annual Report to Congress on the Implementation of the Individuals with Disabilities Education Act*. Washington, DC: Author. Retrieved from https://www2.ed.gov/about/reports/annual/osep/index.html

U.S. Department of Labor, Bureau of Labor Statistics. (2017, June 21). *Persons with a disability: Labor force characteristics – 2016.* Retrieved from https://www.bls.gov/news.release/pdf/disabl.pdf

Vittek, J. E. (2015). Promoting special educator teacher retention: A critical review of the literature. *SAGE Open, 5*(2). doi:2158244015589994.

von der Embse, N. P., Kilgus, S. P., Solomon, H. J., Bowler, M., & Curtiss, C. (2015). Initial development and factor structure of the Educator Test Stress Inventory. *Journal of Psychoeducational Assessment, 33*(3), 223-237.

Walker, H. M., & Gresham, F. M. (Eds.). (2013). *Handbook of evidence-based practices for emotional and behavioral disorders: Applications in schools.* New York: Guilford Publications.

Wasburn, M. H., Wasburn-Moses, L., & Davis, D. R. (2012). Mentoring special educators: The roles of national board certified teachers. *Remedial and Special Education, 33*(1), 59-66.

Webster, A., Cumming, J., & Rowland, S. (2017). Effective practice and decision-making for parents of children with autism spectrum disorder. In *Empowering Parents of Children with Autism Spectrum Disorder* (pp. 3-7). Springer: Singapore.

Westling, D. L. (2010). Teachers and challenging behavior: Knowledge, views, and practices. *Remedial and Special Education, 31*(1), 48-63.

White, S. W., Keonig, K., & Scahill, L. (2007). Social skills development in children with autism spectrum disorders: A review of the intervention research. *Journal of Autism and Developmental Disorders, 37*(10), 1858-1868.

White, S., Scahill, L., Klin, A., Koenig, K., & Volkmar, F. (2007). Educational placements and service use patterns of individuals with autism spectrum disorders. *Journal of Autism and Developmental Disorders, 37*(8), 1403–1412.

Whitehurst, G., Chingos, M. M., & Lindquist, K. M. (2014). *Evaluating teachers with classroom observations*. Brown Center on Education Policy: Brookings Institute.

Windschitl, M., Thompson, J., Braaten, M., & Stroupe, D. (2012). Proposing a core set of instructional practices and tools for teachers of science. *Science Education, 96*(5), 878-903.

Wold, H. (1980). Model construction and evaluation when theoretical knowledge is scarce: Theory and application of partial least squares. In J. Kmenta & J. B. Ramsey (Eds.), *Evaluation of Econometric Models* (47-74). New York: Academic Press.

Woolf, S. B. (2015). Special education professional standards: How important are they in the context of teacher performance evaluation?. *Teacher Education and Special Education, 38*(4), 276-290.

Wong, C., Odom, S. L., Hume, K. A., Cox, A. W., Fettig, A., Kucharczyk, S., ... & Schultz, T. R. (2015). Evidence-based practices for children, youth, and young adults with autism spectrum disorder: A comprehensive review. *Journal of Autism and Developmental Disorders*, *45*(7), 1951-1966.

Wu, M., Tam, H. P., & Jen, T. H. (2016). Classical test theory. In Authors (Eds.), *Educational Measurement for Applied Researchers* (pp. 73-90). Singapore: Springer.

Yell, M. L., Drasgow, E., & Lowrey, K. A. (2005). No child left behind and students with autism spectrum disorders. *Focus on Autism and Other Developmental Disabilities*, *20*(3), 130-139.

Yuan, K. H., & Bentler, P. M. (1999). F tests for mean and covariance structure analysis. *Journal of Educational and Behavioral Statistics*, *24*(3), 225-243.

Zhang, G., & Zeller, N. (2016). A longitudinal investigation of the relationship between teacher preparation and teacher retention. *Teacher Education Quarterly*, *43*(2), 73.