Title
Evaluating Special Education Instructional Practices Using Observation Rubrics: Investigating the Reliability of School Administrator Ratings

Author
Lawson, Janelle

Publication Date
2015

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Evaluating Special Education Instructional Practices Using Observation Rubrics: Investigating the Reliability of School Administrator Ratings

A dissertation submitted in partial satisfaction of the requirements for the degree Doctor of Philosophy in Special Education

by

Janelle Lawson

2015

ABSTRACT OF THE DISSERTATION

Evaluating Special Education Instructional Practices
Using Observation Rubrics: Investigating the Reliability
of School Administrator Ratings

By

Janelle Lawson
Doctor of Philosophy in Special Education
University of California, Los Angeles, 2015
Professor Jeffrey J. Wood, Co-Chair
Professor Lois A. Weinberg, Co-Chair

Recent federal grant programs and legislative initiatives have focused on improving measures of effective teaching within comprehensive teacher evaluation systems. While emerging research is contributing important information about the reliability and validity of evaluative measures as they apply to general educators, relatively little work has been done on the use of these same measures with special education teachers. What is unknown at present is who is best qualified to perform evaluations of special education teachers, especially when using one of the most common classroom performance measures: observation protocols. This study examined the rater reliability of school administrators—who do not possess expertise in the area of special education, and are typically responsible for conducting evaluations of special education teachers

in the actual school setting—as raters of special education teachers' instructional practice. The study used a mixed method-design, which involved a quantitative analysis of administrator ratings using generalizability (G) theory. A qualitative analysis of participants' perceptions of their rating experience, as well as special education teachers' perceptions of the evaluation process, was performed using a phenomenological approach. Findings suggest that school administrators show promise as reliable raters even without formal training in special education, but school administrators need to engage in repeated classroom visits, be invested in the evaluation process, and be properly trained on any measure used for evaluative purposes.

The dissertation of Janelle Lawson is approved.

Diane S. Haager

Jose-Felipe Martinez-Fernandez

Howard S. Adelman

Lois A. Weinberg, Committee Co-Chair

Jeffrey J. Wood, Committee Co-Chair

University of California, Los Angeles

2015

For Mom, Dad, Greg, Tim, and Marianne, whose unconditional and unfailing love and support have made this work possible.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

I would like to thank my committee members for reading each draft, answering each question, and reminding me that the obstacles are not insurmountable. To Jeff—my UCLA advisor and mentor, who so patiently listened to my thoughts as I formed research questions over the years, and who provided just the right amount of "poptimism" to get me around each corner and up another step. To Lois—my CSULA advisor and mentor, who was a steady and trustworthy support throughout this process, and was always one phone call away to offer much needed advice. To Howard—one of my favorite UCLA professors, whose warmth and encouragement sustained me during some of the more challenging moments. To Diane—an inspiration both personally and professionally, who had a fresh idea at each apparent stumbling block, and who reminded me that I have the rest of my life to do my life's work. To Felipe—whom I refer to as Grandmaster, for shaping this work into what it has become, and for being a sounding board, knowledge broker, and the person with the answers to everything.

I would like to thank Carrie Semmelroth at Boise State University for providing the foundational elements of this work, and for partnering with me throughout the initial stages of this process.

I would like to thank Mike Seaman for his unfaltering support in providing space in which to work. His thoughtfulness during this project made data collection possible.

I would like to thank Pat Lampman for his critical eye, sense of humor, and a near lifetime of mentorship.

I would also like to thank Greg Knollman—my "dissertation buddy"—for his selfless contribution to the qualitative component of this work, for being a ready and able coach, and for helping me set goals and achieve them. Our weekly meetings kept me accountable (and sane) throughout this past year of writing. Each word of encouragement was a contribution in and of

itself, and each conversation offered a light at the end of the tunnel. He so kindly and patiently listened to my intended direction, and helped formulate the steps to make it possible.

Finally, I would like to thank my mother and father for contributing their expertise on the subject matter, and special thanks to my mother for being the proofreader of all proofreaders.

VITA

2011	M.A., Special Education, Mild/Moderate Disabilities California State University, Los Angeles
2010	Education Specialist Teacher Credential, Mild/Moderate Disabilities California State University, Los Angeles
2006	B.A., Sociology University of California, Los Angeles

PROFESSIONAL EXPERIENCE

2012-2015	Adjunct Lecturer, Special Education (Mild/Moderate Disabilities) Charter College of Education, California State University, Los Angeles
2014-2015	Master Teaching Assistant University of California, Los Angeles
2014-2015	Lead Tester Project Scale-Up Evaluation of Reading Intervention for First Grade English Learners (California State University, Los Angeles in collaboration with University of Houston, Texas, and University of Colorado, Boulder)
2015	Manuscript Reviewer Journal of Special Education Leadership (JSEL)
2014	Proposal reviewer in the area of Instructional Strategies/Response to Intervention Council for Exceptional Children (CEC) Annual Convention
Summer 2014	Intern Office of Special Education Programs (OSEP), U.S. Department of Education
2014	Research and Inquiry Conference Program Committee Member Graduate School of Education and Information Studies, University of California, Los Angeles
2007-2012	Teacher, Special Education West Covina High School

SELECTED COURSES TAUGHT

California State University, Los Angeles

2012-2015	Cognitive, Linguistic, and Literacy Processes in Exceptional Individuals (adjunct lecturer)
2015	Assessment, Strategies, and Curricular Modifications for Individuals with Special Learning Needs in Diverse Settings (adjunct lecturer)
2013-2014	Demonstration of Competencies/Directed Teaching for the Mild/Moderate Credential (university supervisor)
2013	Understanding Students with Special Needs in Urban Schools (adjunct lecturer)

PUBLICATIONS AND PRESENTATIONS

Lawson, J. E., & Cmar, J. L. (2015, April). *Life after Assembly Bill 3632: Mental health service provision in California schools.* Paper presented at the 2015 Council for Exceptional Children Convention and Expo. San Diego, CA.

Lawson, J. E. (2014). Value-added modeling: Challenges for measuring special education teacher quality. *InterActions: UCLA Journal of Education and Information Studies*, *10*(1), gseis_interactions_20131. Retrieved from http://escholarship.org/uc/item/9r67085n

Lawson, J. E. (2014). *Mental Health Services for Students with Disabilities in California: Service Provision After the Repeal of AB 3632*. Poster presented at the University of California Center for Research in Special Education, Disabilities, and Developmental Risk Annual Conference. Santa Barbara, California.

Lawson, J. E. (2013). *A look at California Assembly Bill 114: Transition of special education and related services formerly provided by county mental health agencies*. Retrieved from http://smhp.psych.ucla.edu/pdfdocs/ass114brief.pdf

FELLOWSHIPS AND AWARDS

2015	Council for Exceptional Children (CEC) Division for Research Doctoral Student Scholar – Seventh Cohort
2014	California Council for Exceptional Children (CA-CEC) Student Scholarship
2013	Graduate Summer Research Mentorship University of California, Los Angeles

CHAPTER ONE: INTRODUCTION

The topic of teacher quality has become a permanent fixture in the ongoing discussion of educational reform in the United States. The federal government has long pushed for high-quality teachers in high-quality schools, and federal initiatives continue to demand that school districts employ and retain teachers who are effective, or deliver "instruction that helps students learn and succeed" (Weisberg, Sexton, Mulhern, & Kneeling, 2009, p. 5). Part of ensuring a high-quality, or effective, teacher workforce is establishing a system of teacher evaluation that precisely differentiates teacher performance and enables teachers to receive the feedback and support they need to improve as professionals. Federal incentive programs such as Race to the Top have encouraged states and school districts to improve their teacher evaluation systems, and large-scale research endeavors such as the Measures of Effective Teaching (MET) Project have investigated various tools used to measure teacher efficacy.

Research efforts have homed in on the two predominant teacher evaluative measures used by school districts: value-added models (VAMs) and observation instruments (Darling-Hammond, 2010; Goe & Croft, 2009; Goe & Holdheide, 2011; Jones, Buzick, & Turkan, 2013). While emerging research is contributing important information about the reliability and validity of these tools as they apply to educators in general, very little work has been done on the applicability of these same tools specifically to the evaluation of special educators, whose instructional practice requires a specialized skillset and unique knowledge base (Sledge & Pazey, 2013). Without proper resources and guidance, it is uncertain whether school administrators are able to meaningfully evaluate their special educators and provide them with adequate support (Sledge & Pazey, 2013).

What existing research does suggest is that value-added models are a woefully inadequate means by which to capture a special education teacher's performance (e.g., Buzick & Laitusis, 2010). Observation instruments, while imperfect, can assist in establishing a uniform vision of effective teaching (Holdheide, Goe, Croft, & Reschly, 2010) and allow school administrators to target and observe specific teaching skills they expect to see in the classroom (Stuhlman, Hamre, Downer, & Pianta, 2010). For special educators, an observation instrument tailored to their particular instructional practices, classroom settings, and student populations would potentially enable administrators to systematically and equitably evaluate their classroom performance (Holdheide et al., 2010). Observation instruments also allow evaluators to provide special educators with targeted feedback, which is critical to any educator's professional growth and development (Kane & Staiger, 2012; Pianta & Hamre, 2009).

Creating and refining a valid instrument is critical to the evaluation of special educators' instructional practices, but tantamount to the instrument itself is the individual responsible for using it. Some researchers have called into question the ability of school administrators, who typically perform teacher evaluations, to reliably and equitably perform evaluations of special educators (Sledge & Pazey, 2013). This concern stems from the reality that most school administrators do not possess formal training or have expertise in the area of special education, and, therefore, may be ill-prepared and unqualified to provide special educators with a meaningful evaluation of their instructional practices (Sledge & Pazey, 2013). Research is needed to explore whether a lack of knowledge regarding specific special education instructional practices systematically biases school administrators in their evaluations of special education teachers (Jones & Brownell, 2013).

The following section begins with a discussion of teacher evaluation in general and the specific policy and research efforts contributing to reform. The two most frequently used teacher evaluation measures—VAMs and observation instruments—are discussed, along with the implications for using both measures in the evaluation of special educators. A case will be made for using observation instruments, especially those specifically designed to reflect the unique nature of special education classrooms, to evaluate special education teachers' instruction. The following section will conclude with a discussion of the individuals—namely school administrators—tasked with evaluating special education teachers in the actual school setting, and whether those individuals are best able to produce fair and reliable evaluations.

It should be noted that while discussions of effective teaching often include the term teacher quality, this paper distinguishes between the constructs of teach*er* quality and teach*ing* quality (Jones & Brownell, 2013). According to Jones and Brownell (2013), measuring teacher quality requires extending inferences about a teacher's performance in a specific context at a particular time to that teacher's ability overall. Any tool that purports to measure teacher quality assumes stable attributes across contexts and focuses on fixed teacher characteristics. Measuring teaching quality, however, ties a teacher's performance directly to the instructional context in which it occurs. In this paper, the evaluation tools used to measure effective teaching are discussed as measures of teaching quality (i.e., context-specific) rather than teacher quality, which cannot be captured by any one instrument at a single point in time.

Evaluating Educators

Research indicates that teachers are important predictors of students' future achievement (Gordon, Kane, & Staiger, 2006; Lockwood et al., 2007; Rivkin, Hanushek, & Kain, 2005; Sanders & Rivers, 1996; Wright, Horn, & Sanders, 1997). Teacher evaluation systems, however,

have been criticized for failing to differentiate efficacy among teachers (Weisberg et al., 2009), and for failing to provide teachers with meaningful feedback on their performance (Harris, Ingle, & Rutledge, 2014). Formal evaluation practices, which are often based upon classroom observations that are typically short (i.e., one class period or less than 60 minutes) and infrequent (i.e., two or fewer classroom observations per academic year)[1] (Weisberg et al., 2009), give nearly all teachers the highest possible ratings (Little, 2009). In a survey of 15,000 teachers and 1,300 administrators in 12 districts across four states, Weisberg et al. (2009) found that 99 percent of teachers receive a rating of satisfactory in districts that use binary rating evaluations (i.e., *satisfactory* or *unsatisfactory*); in districts that use more rating options, less than one percent of teachers are rated as unsatisfactory (Weisberg et al., 2009).

Although formal evaluations by school administrators show alarmingly little variability (Weisberg et al., 2009), other evidence suggests that principals are able to separate teachers on the basis of quality/efficacy when asked (Armor et al., 1976; Harris & Sass, 2009; Harris et al., 2014; Jacob & Lefgren, 2008). Harris et al. (2014) interviewed 30 principals from elementary, middle, and high schools in Florida, and asked the principals to rate a list of 10 of their teachers on a scale from 1(*low*) to 9 (*high*) on a variety of personal and professional attributes. The principals were also asked to explain their ratings, provide specific examples and lengthy descriptions of each teacher's characteristics, and give each teacher an overall effectiveness rating. The principals applied the full range of the scale (i.e., 1 to 9) when rating teachers, although 69% of the teachers were rated in the top three categories (Harris et al., 2014). The results of the study suggest that principals can distribute teachers among a wider range of rating

[1] Formal evaluations are typically performed once per academic year for non-tenured teachers in public schools. Tenured teachers in public schools are typically evaluated once every other academic year.

categories than merely *satisfactory* and *unsatisfactory*, but the study also confirms previous findings that rating distributions tend to be skewed such that a large proportion of teachers fall in the high end (Little, 2009; Weisberg et al., 2009).

From a measurement perspective, evaluation tools that are limited in rating options, especially those that are binary, prevent school administrators from more precisely differentiating teacher efficacy. Beyond the measures themselves, administrators are giving a disproportionately large number of teachers remarkably high ratings, which assumes equal classroom effectiveness in a higher range than may be accurate. As Weisberg et al. (2009) suggest, if equal classroom effectiveness is assumed and teacher efficacy is not reliably and validly measured, then "[e]xcellent teachers cannot be recognized or rewarded, chronically low-performing teachers languish, and the wide majority of teachers performing at moderate levels do not get the differentiated support and development they need to improve as professionals" (p. 4). Given that teacher performance is important to student performance, teacher evaluations should do much more than deem a teacher satisfactory or unsatisfactory; strong evaluation systems should recognize truly effective teachers, identify and support teachers who need to improve, and serve as a basis for removing ineffective teachers.

Federal Initiatives

The federal government has supported the reform of evaluation measures that identify effective teachers and improve the quality of the teacher workforce overall. The No Child Left Behind Act (NCLB, 2002) focused its efforts on ensuring that teachers possessed observable qualifications such as education, licensure, and subject matter competency as measured by passing a relevant exam. According to NCLB, a highly qualified teacher is one who possesses a bachelor's degree, has full state certification or licensure, and demonstrates competency in the

subject area to be taught. Although NCLB mandates that a teacher be deemed "qualified" in order to provide instruction, observable qualifications do not necessarily make a teacher "effective." Results of studies examining teacher qualifications such as undergraduate degrees, graduate degrees, and years of experience as possible factors in differentiating effective from ineffective teachers have been mixed (see the review section in Harris & Sass, 2011). The inconsistent findings have caused researchers and practitioners to question whether teacher qualifications have any impact at all on student achievement.

To explore this issue further, Harris and Sass (2011) sought to identify specific and observable teacher qualifications that contribute to teacher productivity using student-level achievement test data for both math and reading as an outcome measure. The study analyzed the impacts of teacher experience, post-baccalaureate degrees, in-service professional development, and pre-service undergraduate education. The study also distinguished between forms of training, types of coursework at the undergraduate level, and the quality of undergraduate training controlling for the innate ability of future teachers as measured by college entrance exam scores. The results of the study indicated that experience increases teacher productivity at the elementary and middle school levels, but formal training acquired while teaching does not. The attainment of advanced degrees did not impact student achievement, in-service professional development had little or no effect on teacher productivity, and specific undergraduate coursework in education also had no effect.

Rivkin et al. (2005) found that obtaining a master's degree did not improve teacher skills, and the authors found that experience may improve teacher quality in the first three years, but not subsequently. In a validity study of the Classroom Assessment Scoring System (CLASS-S) using data from 82 Algebra classrooms, Bell et al. (2012) found that CLASS-S scores were not

related to teachers' knowledge of Algebra. Although NCLB requires teachers to demonstrate subject matter competency, the Bell et al. (2012) study suggests that knowledge of a particular subject area does not guarantee high quality teaching, which, in turn, does not ensure student achievement. In sum, observable teacher qualifications appear to be only weakly related to student achievement (Harris & Sass, 2011; Koedel & Betts, 2007).

Race to the Top. In a departure from the observable teacher qualifications required under NCLB, the Obama administration has targeted teacher quality by way of linking teacher performance directly to student achievement. The Obama administration has launched a competitive grant program entitled Race to the Top, which offers large monetary incentives for states that demonstrate success in raising student achievement, as measured by standardized test scores. Unlike NCLB, which defined high quality teachers by their education, credentials, and subject matter knowledge, the Race to the Top executive summary, authored by the U.S. Department of Education (2009), defines effective teachers as those "whose students achieve acceptable rates (*e.g.*, at least one grade level in an academic year) of student growth" (p. 12) and highly effective teachers as those "whose students achieve high rates (*e.g.*, one and one-half grade levels in an academic year) of student growth" (p. 12).

According to the U.S. Department of Education (2009), states are eligible for funding if they establish an approach to measuring student growth, design fair evaluation systems that take student growth into account as a significant factor, and provide teachers with data on the growth of their students. Schools must also use the data on student growth to inform decisions regarding compensating, retaining, and removing tenured and untenured teachers. In essence, Race to the Top places emphasis on measuring student achievement and linking student growth directly to teachers and principals. School districts are required to create systems whereby teachers receive

recognition for their direct contribution to student growth or are removed should their instruction fail to result in increases in student achievement. Race to the Top will only consider funding states that evaluate teachers "in significant part" (p. 12) by student growth; consequently, many states have already adopted or are in the process of developing teacher assessment systems that use student achievement data as the primary outcome variable of interest (Winters & Cowen, 2013). According to Winters and Cowen (2013), 19 states have developed policies that dismiss teachers for ineffective teaching, and 13 of those states use student achievement data as the primary determinant of ineffective teaching.

Measures of Effective Teaching (MET) Project

Race to the Top has encouraged states to improve their teacher evaluation systems, but teacher evaluation reform has also had strong support from research initiatives. In response to the growing consensus that teacher evaluation systems in the United States are considerably lacking (Kane & Staiger, 2012), the Bill and Melinda Gates Foundation launched a wide-scale study of teacher evaluation. With 3,000 teacher volunteers and dozens of research teams, the MET project is thus far the largest study of instructional practice and its relationship to student performance; it includes multiple instruments for classroom observations, student and teacher perceptual surveys, and multiples measures of student achievement gains.

While the MET project culminated in a series of reports including various findings and recommendations, one example of an area under investigation during the course of the project was the use of observation instruments as measures of effective teaching. Many school districts rely heavily upon observation rubrics to evaluate teachers (Holdheide et al., 2010), but large-scale projects had not examined the instruments themselves and the optimal conditions necessary for reliable and valid observation scores. In one of the MET project studies, three video-

recorded lessons were collected from 1,333 teachers across North Carolina, Texas, Colorado, Florida, New York, and Tennessee. The video lessons were scored by MET project raters using the Classroom Assessment Scoring System (CLASS), Framework for Teaching (FFT), Protocol for Language Arts Teaching Observations (PLATO), Mathematical Quality of Instruction (MQI), and UTeach Teacher Observation Protocol (UTOP) observation instruments. Some findings included the following: the five observation instruments were positively associated with student achievement gains in both English Language Arts (ELA) and math (Kane & Staiger, 2012), multiple raters needed to score multiple lessons to achieve high levels of reliability (Kane & Staiger, 2012), and combining observation scores with student feedback and achievement resulted in greater and more stable reliability (Kane & Staiger, 2012). From these findings, along with those from several other studies, the MET project team offered recommendations for improving teacher evaluation systems and practice.

Applying Teacher Evaluation Measures to General and Special Educators

As the federal government and research initiatives have brought much needed attention to teacher evaluation practices, many school districts have reformed their practice or are in the process of revising their evaluation systems. Currently, the two predominant evaluation tools in use—used alone or in combination—are observation protocols and value-added models (VAMs) (Darling-Hammond, 2010; Goe & Croft, 2009; Goe & Holdheide, 2011; Jones et al., 2013). The majority of states and districts utilize an observation protocol as the primary component of their evaluation systems (Holdheide et al., 2010), but VAMs, which rank order teachers based on the achievement gains of their students, have gained recent momentum due to encouragement from Race to the Top.

Research efforts are underway to establish the validity of both VAMs and observation instruments as they apply to general educators[2], but little has been done to address how these same practices can be fairly applied to special educators. The following section begins with a discussion of VAMs and observation protocols as evaluative measures for general education teachers in public schools—there are methodological and practical concerns associated with both VAMs and observation protocols, and these concerns serve as a necessary conceptual foundation for discussing the particular application of these measures to special educators. The following section will also discuss the implications of using both VAMs and observation protocols in their current form to evaluate special educators.

VAMs. A value-added model involves student achievement data, as measured by standardized test scores from two or more years, matched to teacher and/or school-level data (Buzick & Laitusis, 2010). VAMs are statistical models that take into account student prior achievement on standardized tests to estimate a teacher-specific effect on achievement (Holdheide, Browder, Warren, Buzick, & Jones, 2012). Some VAMs attempt to control for other variables, including student characteristics such as race and peer influence (e.g., Kane & Cantrell, 2013), and school characteristics such as the percentage of students receiving free and reduced price meals (see McCaffrey, Lockwood, Koretz, Louis, & Hamilton [2004] for a technical description of VAM and Braun [2005] for a non-technical one).

According to Holdheide et al. (2012), VAMs are an improvement upon other systems of teacher evaluation because they provide a standardized, common metric; are based on large-scale standardized assessments with more desirable psychometric properties; and do not require

[2] General education teachers, sometimes referred to as regular education teachers, are those licensed/certified to teacher specific grade levels and/or specific subject areas, not including special education.

students to meet set proficiency levels.[3] VAMs take into account students' prior achievement so as to measure growth, rather than focusing on a uniform achievement target across all populations. VAMs are also intended to make causal inferences about a teacher's direct influence on student achievement (Holdheide et al., 2012). In other words, VAMs are designed to isolate and quantify a teacher's direct impact on student learning, and a teacher's quality score can be ranked relative to the scores of other teachers in the same school or district.

Most value-added modeling, which purports to measure teacher quality or efficacy, focuses exclusively on standardized test scores as an outcome of interest (McCaffrey et al., 2004). Research suggests that dependence on students' standardized test scores as an exclusive measure of teacher quality is problematic, and caution should be taken when interpreting the results from growth models. Some major areas of concern associated with VAMs include, but are not limited to, the following: a teacher's value-added score appears to be dependent on the type of student achievement assessment used (Lockwood et al., 2007; Papay, 2011); it is difficult to disentangle the effects of multiple teachers over time and the variety of other factors that may influence how well students are able to perform on standardized tests (Baker et al., 2010; Braun, 2005; Hill, 2009; Koretz, 2002; McCaffrey, 2012); non-random sorting of students into classrooms and of teachers into schools and classrooms interferes with causal inferences about a teacher's isolated effect on student learning (e.g., Baker et al., 2010); and VAM rankings for teachers appear to be inconsistent from year to year (Kane & Staiger, 2008; McCaffrey, Sass, Lockwood, & Mihaly, 2009).

[3] Under NCLB, states are mandated to set proficiency targets on standardized tests. Committees in each state determine a "cut score" (e.g., 90 out of 100 correct answers on a given test) that students must reach to demonstrate proficiency in an academic area.

There are several other practical concerns that states and districts should consider before incorporating VAMs into teacher assessment systems. Currently, there is a lack of appropriate tests for all grade levels and subjects, which affects a district's ability to use student achievement data as an outcome measure for all teachers. Also of practical concern are the computing resources required to perform high-quality longitudinal data analysis. Many districts simply do not have the equipment, personnel, and/or expertise to perform the necessary computing for VAMs (McCaffrey et al., 2004). For states and school districts that do possess the knowledge and equipment, there is increased concern over model and policy designs (Goldhaber, Goldschmidt, & Tseng, 2013); that is, how districts design their policies, set performance criteria for teachers, and choose the specific VAM to employ will all affect teacher rankings.

What is especially disconcerting are the implications that the rampant use of VAMs will create disincentives for teachers to work with the neediest and most challenging student populations, and that teachers will be less likely to work cooperatively with other teachers as the field becomes increasingly competitive (Baker et al., 2010). It is possible that teachers will limit sharing of instructional techniques and materials with colleagues and that individual performance will take precedence over the collaborative efforts necessary to improve schools as a whole.

Using VAMs to evaluate special educators. There is a paucity of research on the use of VAMs to measure the quality of special education teachers because special education programs, classrooms, and students present unique challenges that complicate straightforward growth models. One such challenge is including the standardized test scores of students with disabilities (SWDs) in VAMs. SWDs often take alternative or modified state assessments if determined by

an Individualized Education Program (IEP) team[4]; scores from general, modified, and alternate assessments are on different scales and it may be impossible to combine them in some longitudinal models (Buzick & Laitusis, 2010). In a survey of 15 states using growth models for teacher assessment, 13 of those states did not include students who took alternate assessments in their growth model outcomes (Ahearn, 2009).

When SWDs do take state assessments, they have additional options that distinguish their assessments from those of their general education peers. If agreed upon by an IEP team, SWDs may have access to testing accommodations and modifications, which are special education supports designed to improve accessibility. Testing *accommodations* are those that do not alter the construct being measured; for example, the use of large print on tests that do not measure vision, or extended time on tests that do not measure speed (Buzick & Laitusis, 2010). According to Koretz and Hamilton (2006), "The psychometric function of accommodations is to increase the validity of inferences about students with [disabilities] by offsetting specific disability-related, construct-irrelevant impediments to performance" (p. 562). Testing *modifications* are changes that do alter the construct being measured; for example, providing the use of a calculator for test questions designed to measure multiplication. When SWDs make use of testing modifications, measurements of growth cannot be directly interpreted.

The annual review and amendment of IEP accommodations/modifications, along with students' freedom to refuse the accommodations/modifications, result in variability in the number and type of testing accommodations/modifications used from year to year (Fuchs & Fuchs, 2001). Variability is the result of the changing needs of SWDs over time or the changing

[4] An IEP team must include, at a minimum, the following members: the child's parent(s), at least one regular education teacher of the child, at least one special education teacher of the child, a representative of the public agency, other individuals who have expertise in areas of related services (when appropriate), and the child with the disability (when appropriate).

preferences of the student, but variation across years can also be due to external factors such as changes to state policy (Christensen, Lazarus, Crone, & Thurlow, 2008). The problem with the inconsistent application of testing accommodations is that standardized test scores can inflate or deflate depending on the addition or removal of supports (Jones et al., 2013). Research indicates that the use of testing accommodations results in differential score changes for SWDs (Sireci, Scarpeti, & Li, 2005); especially when testing accommodations are added, a performance boost may occur in one particular year for a student with a disability. According to Buzick and Laitusis (2010), "The implication is that the change in test scores from year to year may be related to inconsistency in the use of accommodations and modifications rather than true changes in knowledge, skills, and abilities over time" (p. 540). It is difficult, then, to isolate true special education teacher effects from the effects of testing supports in the growth of SWDs' scores.

An additional concern is that a dependence on standardized test scores alone as an outcome measure for special education teaching quality may provide an inaccurate and incomplete picture of the extent to which the special education teacher is contributing to student growth. Special education teachers serve as case carriers for SWDs, and they are charged with identifying areas that may be impeding academic performance; those may be social, emotional, behavioral, or other areas that adversely affect a student's ability to make progress in the general education classroom. A child with Autism Spectrum Disorder (ASD) may exhibit social deficits, which impact his or her ability to participate in academic activities within the classroom. A student with an emotional disorder may withdraw and have difficulty fully engaging with a teacher's instruction. A student with a behavioral disorder may need assistance with behaviors that are impacting his or her ability to appropriately manage academic tasks. These are merely a few examples of areas that a special education teacher will target and teach to in order to offer a

student with a disability the best chance of achieving his or her potential in an academic environment.

*Council for Exceptional Children (CEC) recommendations***.** In 2009, CEC convened an advisory group to consider and discuss the implications of "pay for performance"[5] systems and VAMs for special education teacher evaluations. CEC recognized that teacher evaluation systems were increasingly incorporating student performance, and the growing adoption of VAMs was cause for concern. The aforementioned criticisms of VAMs in general and the complications associated with their applicability to special educators prompted CEC to make initial recommendations regarding special education teacher evaluations going forward. In 2013, CEC drafted an official position on special education teacher evaluation, which included, but is not limited to, the following recommendations: (a) school districts should use one evaluation system that is appropriately differentiated based on a special educator's professional role, (b) evaluations should not use IEP goals as measures of student growth, (c) evaluations should be used to support teachers in their ongoing growth, (d) evaluators should have knowledge of special education teaching and should be trained in effective evaluation practices that accurately reflect special education teachers' roles and responsibilities; (e) evaluations based on student growth are insufficient to capture a special educator's contributions to student growth, and (f) VAMs should not be applied to any teacher until there is general consensus among researchers that the model provides a valid estimate of a teachers' contribution to student growth (CEC, 2013).

[5] Teachers typically receive pay increases based on a "step and column" formula where steps represent years of service and columns represent levels of education. Pay increases are given automatically as teachers increase their number of steps and advance across columns. Pay for performance, or merit-based pay, is a system of teacher evaluation that provides teachers with opportunities to earn pay increases based on performance, or how well they teach. This is often determined using student achievement data and is connected to growth models.

Observation instruments. Although much has been written about growth models and their potential use in teacher evaluation, little attention has been paid to observation instruments in comparison (Taylor & Tyler, 2011). Even without a robust research base, classroom observations remain the predominant data source used in teacher evaluations (Holdheide et al., 2010), and the primary source of information for school principals regarding the instructional practices of their teachers (Grissom, Loeb, & Master, 2014). Some classroom observations are informal (Blanton, Sindelar, & Correa, 2006; Stuhlman et al., 2010), but formal rubrics have been developed to support observers in their assessment of teaching skills and classroom performance. Two commonly used classroom observation systems that are commercially available are Charlotte Danielson's Framework for Teaching (FFT) Evaluation Instrument (Danielson, 2011), which was updated in 2013 to include alignment with the Common Core State Standards, and the Classroom Assessment Scoring System (CLASS; Pianta, La Paro, & Hamre, 2008). Both the FFT and the CLASS are supported by research on reliability and validity, and they are both intended for use across grade levels and content areas.

The FFT consists of four domains: (a) *Planning and Preparation*, (b) *The Classroom Environment*, (c) *Instruction*, and (d) *Professional Responsibility*. Observers rate teachers on a scale from 1-4: Level 1 – *Unsatisfactory*, Level 2 – *Basic*, Level 3 – *Proficient*, Level 4 – *Distinguished*. While the rubric is intended to capture professional duties beyond just classroom instruction, many states only use *The Classroom Environment* and *Instruction* rubric items as those two domains can be directly observed (Jones & Brownell, 2013). According to Noell, Brownell, Buzick, and Jones (2014), the FFT has been adopted by some large school districts (e.g., Los Angeles Unified School District), and by some states (e.g., Illinois, Rhode Island, and Delaware).

The CLASS observation instrument is designed to measure the interactions between teachers and students in classrooms, and is organized into three domains: *emotional support*, which focuses on positive classroom climates and responsiveness to student needs; *classroom organization*, which focuses on behavior management, classroom expectations, organized instruction, and the use of instructional time; and *instructional support*, which focuses on how teachers develop student knowledge, provide feedback to students, and support them in language development through classroom discussions. Whereas the FFT limits raters to a 4-point scale, the CLASS utilizes a 7-point scale, and scores are assigned based on anchor descriptions at the "low," "mid," and "high" ranges (Noell et al., 2014).

The advantage of observation instruments is that they enable observers—usually school administrators—to view classroom practice as it actually occurs, target specific observable skills and instructional practices for development, and provide teachers with feedback based on what is observed in the classroom setting. When using the FFT, providing clear feedback to teachers led to substantial gains in student's math achievement (Taylor & Tyler, 2011). Individual coaching for secondary school teachers also led to substantial improvement in student achievement gains when using the CLASS (Allen, Pianta, Gregory, Mikami, & Lun, 2011). Additionally, scores on several observation instruments (i.e., FFT, CLASS, UTOP, MQI, and PLATO) were related to student achievement gains in ELA and math (Kane & Staiger, 2012). According to Kane and Staiger (2012), "[t]he real potential of classroom observations is their usefulness for diagnosis and development of instructional practice" (p. 15). Observation instruments have the potential to serve as a strong foundation for providing teachers with meaningful feedback on their classroom performance, which should lead to student achievement gains as instructional practice improves.

An example of a well-developed system for classroom observations used within an existing teacher evaluation program is evident in the Teacher Evaluation System (TES) in Cincinnati Public Schools (Taylor & Tyler, 2011). According to Taylor and Tyler (2011), when teachers are on the TES evaluation cycle, they are typically observed four times during the academic year: three times by an assigned peer evaluator, who is an experienced teacher external to the school, and once by the principal or other school administrator. Teachers are evaluated using the FFT, and are scored on several items that cover classroom management, instruction, content knowledge, and planning. Evaluators receive intensive training and must accurately score videotaped teaching samples to check inter-rater reliability. After each classroom observation, evaluators must provide written feedback to the teacher within 10 days and meet with the teacher individually soon after the first observation. The rubric items are used to provide teachers with specific feedback in a given performance area, including stated characteristics of higher-level scoring categories when necessary. Taylor and Tyler (2011) found higher student achievement in classrooms taught during the TES evaluation year and in years following evaluation compared with the student achievement of students taught by the same teacher in years before he/she participated in TES.

Although Cincinnati Public Schools have integrated classroom observations within a systematic evaluation approach that includes external evaluators and regular feedback, classroom observations used in many other school districts are informal, lack a standardized process, and depend on instruments that are not research validated (Blanton et al., 2006; Stuhlman et al., 2010). Multiple published and unpublished observation systems are used at the school district level (Holdheide et al., 2010), and some observation instruments lack alignment to instructional practices backed by empirical evidence (Stuhlman et al., 2010). Other criticisms of observation

instruments include the following: when numerical ratings are used to obtain scores for evaluation purposes, a single observation by a single observer can produce scores more volatile than value-added (Kane & Staiger, 2012); multiple observers rating multiple lessons are needed to obtain high levels of reliability (Kane & Staiger, 2012), which can be costly and time intensive for school districts; observation instruments require rater judgment (Kane & Staiger, 2012), which may vary depending on the rater; the instruments may contain highly inferential rubric items; and instruments such as the FFT and CLASS require extensive training, which may affect the feasibility of their use in the school setting (Noell et al., 2014).

Using observation instruments to evaluate special educators. Whether the FFT, an adapted version, or another rubric, 93.8% of sampled school districts across the United States reported using observation protocols as the data source for teacher evaluations (Holdheide et al., 2010). Of 1,143 total respondents, 85.6% reported that the same observation protocol is used for all teachers including special educators, and 72% of districts did not allow for a slightly modified process for evaluating special educators. Although one rubric is typically applied to general and special educator evaluations, half of the respondents disagreed or strongly disagreed that special educators should be evaluated in the same way as general educators. Additionally, 61% of respondents believed that those who evaluate special educators should have experience in special education.

Existing observation rubrics like the FFT often fail to include components specifically for special educators (Jones & Brownell, 2013), and administrators feel the need to make modifications when evaluating special education instruction (Holdheide et al., 2010).
The FFT, for example, takes a constructivist view of student learning wherein students play an active role in constructing their knowledge (Noell et al., 2014). Jones and Brownell (2013) argue

that effective instruction for SWDs involves direct, explicit instruction that is largely teacher-directed. Furthermore, a meta-analysis of intervention studies found that explicit strategy instruction best predicted the magnitude of treatment outcomes for SWDs (Swanson, 2001). This instructional practice is not reflected in the FFT nor the CLASS and may contribute to a bias against special educators.

Existing observation instruments used for general and special educators alike also assume that teachers are engaging in whole-group instruction that is primarily student-driven. Special educators are tasked with providing highly individualized lessons that meet the unique needs of their students; this includes providing interventions to small groups of students. A robust body of intervention research within the field of special education has established that SWDs need direct, explicit instruction; ongoing and systematic feedback; and should be taught in small, interactive groups (e.g., Brownell et al., 2007; Brownell, Smith, Crocket, & Griffin, 2012). Observation instruments applied to all teachers fail to account for the specialized strategies special educators use in their classrooms and also the nontraditional instructional settings necessitated by the learning needs of the students. Although research studies have examined the reliability and validity of observation systems such as the FFT and CLASS, no existing studies have focused on the application of these instruments to teachers of SWDs.

RESET Observation Tool. Although using the same observation protocol for all teachers helps to create a uniform vision of instructional expectations and guidelines for teacher practice (Holdheide et al., 2010), it is suggested that the observation protocol be modified with a rubric "that is explicitly designed with clear expectations and performance criteria for special education teachers" (Holdheide et al., 2010, p. 16). With this in mind, Johnson and Semmelroth (2012) developed the Recognizing Effective Special Education Teachers (RESET) Observation Tool.

The RESET tool contains 28-67 items depending on the number of lesson components being observed, and the items are grouped into three subscales: Subscale 1 – *Lesson Objective*, Subscale 2 – *Evidenced-Based Practice Implementation*, and Subscale 3 – *Whole Lesson Review*. All items are scored on a scale from 0 to 3, and scores are averaged across dimensions to obtain a holistic score for each subscale.

Like the FFT and CLASS, the RESET was designed to be used across grade levels and content areas. The rubric items contained within the RESET tool reflect evidence-based instructional strategies (e.g., explicit instruction) that are important in classrooms of SWDs, but are absent from tools like the FFT and CLASS. The theoretical underpinning of the RESET is that there is a large body of research on evidence-based instructional strategies that are effective for SWDs (e.g., Swanson, 2001), and the use of those practices should lead to increases in student outcomes (e.g., Cook & Odom, 2013). Rubric items contained within the RESET are intended to differentiate effective from ineffective special educators by way of measuring their use of effective instructional practices supported by research (Semmelroth & Johnson, 2013).

The RESET tool is in pilot testing stages for reliability and validity, but early studies indicate its promise for use (Semmelroth & Johnson, 2013). Using generalizability theory to decompose sources of variance, the variance in scores attributable to persistent differences among teachers ranged from 15 percent to 21 percent across the RESET's subscales (Semmelroth & Johnson, 2013). This outcome is similar to that obtained by the MET study when using generalizability theory to analyze the reliability of the UTOP, CLASS, FFT, MQI and PLATO; teacher variance ranged from 14 percent to 37 percent across instruments. Additionally, inter-rater agreement for the RESET has ranged from .72-.95 with a median agreement of .85 (Semmelroth & Johnson, 2013). Generalizability studies examining sources of

variance for RESET have resulted in G-coefficients ranging from .79-.86 (Johnson & Semmelroth, 2013).

Raters. While an observation instrument such as the RESET allows for a more equitable representation of special educators' instructional practices, another concern related to using observation protocols to evaluate special educators—aside from the instruments themselves—is the expertise of the observer. The responsibility of special education teacher evaluations, including classroom observations, typically falls on school administrators (i.e., principals and assistant principals) (Liu & Johnson, 2006), but researchers have only begun to explore whether school administrators can be trained to score any teacher reliably using observation rubrics (Bell, Jones, Lewis, & Qi, 2013).

Many research studies include raters trained by project staff or use other teachers as raters of instructional practice (Kane & Staiger, 2012; Semmelroth & Johnson, 2013). Earlier studies from the MET project, for example, utilized observers trained by the Educational Testing Service (ETS) to score classroom instruction on five different observation instruments (Kane & Staiger, 2012). Semmelroth and Johnson (2013) used trained peer teachers as raters of special education teachers' classroom instruction on the RESET instrument. While testing observation instruments with trained raters is necessary to establish reliability and validity of the measures, very little is known about how school administrators apply these instruments within the school setting where the administrators know the teachers they are observing, are subject to additional factors that may not be controlled in experimental studies, and are observing classroom practice as it occurs rather than through video-recorded instructional segments.

The MET project has contributed recent findings from one study evaluating the reliability of school administrators in scoring classroom instruction using two domains of the FFT:

classroom environment and *instruction* (Ho & Kane, 2013). During the 2011-2012 school year, teachers in Hillsborough County, Florida were given digital video cameras and microphones to capture their instructional lessons; a total of 67 teachers across a range of grade levels consented to have their lessons scored by administrators and peers. Fifty-three school administrators (principals and assistant principals, some from the teachers' same schools and some from outside schools) and 76 peers (other teachers within Hillsborough) participated as raters of the video lessons. Each rater scored a total of 24 lessons: four lessons from each of six different teachers. Findings from the study included the following: administrators scored their own teachers higher on average than administrators from other schools, administrators from other schools scored teachers higher on average than certified peers, administrators were more likely to differentiate among teachers from their own schools and from other schools than peers were, and reliability was higher for administrator scores than for peers (Ho & Kane, 2013).

The findings from Ho and Kane (2013) suggest that administrators are more reliable raters than peers, but they assign higher scores on average to their own teachers, resulting in a "home field advantage" (p. 15). Despite higher ratings on average, the administrators in general were more likely to assign scores at the extreme ends of the scale, whereas peer raters were more likely to score using the middle categories. While these findings contribute important information about school personnel and peers as raters, the study only included two domains of one instrument (i.e., FFT) and did not include special education teachers in the sample.

Administrators as raters of special education instruction. School administrators are typically responsible for evaluating all teachers on their school sites, including special education teachers. These school administrators, however, often lack a knowledge base regarding evidenced-based instructional strategies recommended for students with specific disabilities

(Sledge & Pazey, 2013). A lack of knowledge may adversely impact a school administrator's ability to provide an accurate and meaningful evaluation of a special education teacher's performance (Sledge & Pazey, 2013). According to Jones and Brownell (2013), if administrators serving as teacher evaluators do not have experience within the field of special education, it could undermine the reliability of observation scores (i.e., administrators may be less likely to agree with one another on ratings of classroom instruction), and systematically bias ratings of special education teachers (i.e., particular administrators may systematically score special education teachers higher or lower on certain elements).

CEC's position on special education teacher evaluation is that evaluators should have knowledge of special education teaching and be appropriately trained in effective evaluation practices as they apply specifically to special educators. According to the Holdheide et al. (2010) survey, while 61% of respondents believed that evaluators of special educators should have experience in special education, only 12.4% of respondents indicated that evaluators were given training designed specifically for evaluating special educators. As previously mentioned, researchers have only begun to explore whether school personnel can be trained to score reliably using observation rubrics, and research has not yet addressed whether administrators without a special education background differentially rate special education instruction. Jones and Brownell (2013) suggest that research on observation tools that use administrators with different educational backgrounds as raters would help determine who is best qualified to conduct classroom observations of special educators for diagnostic and evaluation purposes.

Rationale

There is an urgent need for evaluation models that are valid and reliable measures and that support all teachers in their ongoing growth (Benedict, Thomas, Kimerling, & Leko, 2013).

According to Danielson (2011), evaluation measures should primarily serve two purposes: a) ensure teaching quality and b) promote professional development. While VAMs and similar evaluation measures have focused on the former, it can be argued that the latter is of paramount importance for special educators.

Promoting growth and development is necessary for special educators for several reasons. First, there is a chronic shortage of special education teachers (Johnson & Semmelroth, 2013). Some researchers argue that to improve the quality of the teacher workforce, VAMs should be used to dismiss the lowest performing non-tenured teachers; Johnson and Semmelroth (2013) argue, however, that if the lowest-performing special educators were to be dismissed based on VAM rankings, there would be no special education teachers to replace them. It is critical, then, that evaluation systems do more than merely dismiss ineffective teachers. Special educators need to be supported in their current positions so that SWDs receive consistent instruction from credentialed professionals. Second, educators should use research-based instructional practices designed to maximize student growth, but many special educators report using ineffective instructional methods as frequently as they do those approaches with a strong research base (Burns & Ysseldyke, 2008). Improving the outcomes for SWDs requires improving the instructional practices of the special education teachers responsible for educating them (Scruggs, Masstropieri, Berkeley, & Graetz, 2009).

Rather than depending on VAMs, observation rubrics are better able to provide special educators with the specific feedback they need to inform their professional learning goals (Benedict et al., 2013) and improve their instructional techniques overall. Brownell et al. (2009) argue that observation studies of classroom practice are essential to defining effective teaching in special education instructional settings and ultimately improving practice. As an evaluation

method, classroom observations allow the observer to capture the essence of the learning environment over the course of the instructional day and across different points in time. These observations provide insight into the nuances of individual teaching styles and especially the interaction between teachers and their students (Sledge & Pazey, 2013). Pianta and Hamre (2009) determined that teaching behaviors can be valid predictors of positive student outcomes, and can be improved when teachers are provided with feedback and support. Teachers are also able to improve when they are exposed to best practices, which can occur when they are able to observe the effective classroom practices of their colleagues. In sum, observing the instructional environment as it occurs is a necessary part of helping teachers improve their skills, and providing teachers with meaningful feedback based on the observation is critical for a teacher's overall development.

Improving and refining the instruments used in evaluation processes is a necessary part of ensuring that all teachers receive a fair and meaningful evaluation, but testing the instruments with the individuals most likely to use them is equally important. School administrators are an integral part of the evaluation process, and including them in research on the tools increases the possibility that findings will generalize to the actual school setting. For special educators, research using school administrators as raters of their classroom performance is necessary for determining if a lack of knowledge regarding special education instructional practices systematically biases the results. Even the RESET tool, while designed to reflect the instructional practices of special educators, has only been tested with experienced special education teachers as raters. The raters, then, were knowledgeable about and trained in the field of special education, which may or may not affect how they rate the instructional practices of their peers. An examination of administrator ratings is necessary to determine who is best

qualified to evaluate special education teaching and determine if specific rubric items can be practically applied by administrators in the school setting.

Purpose of the Current Study

The purpose of the current study was twofold. First, this study examined how school administrators without any formal education and/or experience in special education performed with rubric items from an observation instrument. Testing rubric items with school administrators contributes to the field of teacher evaluation research by providing information on whether administrators who lack knowledge concerning special education instruction can reliably evaluate special educators when using observation rubrics, whether they feel qualified to do so, and whether a specific set of rubric items has the potential for application in a school setting by personnel who would be most likely to use it.

Second, this study investigated rubric items from the RESET, an observation protocol that is tailored to special education instruction and is designed to be of maximum benefit to those providing the instruction as well as those responsible for evaluating special education instructional practices. Administrators need tools that can be applied to special education settings and that are valid measures of special education teaching quality. Special educators need those same tools to offer them critical and constructive feedback on their performance so that they have the most opportunity for growth. Existing observation systems such as the FFT and CLASS are not designed to assess the essential elements of instruction necessary for the unique needs of students with disabilities and special education classrooms (Noell et al., 2014); the RESET Observation Tool was developed and pilot-tested specifically for use with special education teachers, and research is necessary to further refine the tool.

Special education instructional practice framework. While the RESET Observation Tool includes several rubric items contained within three subscales, this study included seven specific rubric items that best aligned with a proposed definition of effective special education teaching proffered by Jones and Brownell (2013). Jones and Brownell (2013) argue that the following six features mark effective special education teaching; it is: (a) *explicit* - clear, systematic instruction that builds a rationale for learning, models how to use the strategy or skill, and provides opportunities for practice; (b) *intensive* – the teacher's instruction is purposeful, the pace is rapid, there is little wasted time during instruction, and there are smooth transitions between activities; (c) *cohesive* – the objective of the lesson and individual activities are clearly linked, and the current lesson relates to past or future learning; (d) *engaging* - students seem motived and interested in participating, and the teacher uses multiple techniques for student engagement; (e) *responsive* – there are multiple opportunities to evaluate student learning, and instruction is adjusted based on student need or accommodations are made such that students with disabilities have access to the curriculum; and (f) *focused on essential concepts, strategies, and skills* - essential concepts, strategies, and skills are repeatedly revisited to assist students in developing proficiency (Brownell et al., 2007; Brownell et al., 2012; Swanson, 2001).

It should be noted that the RESET tool is an observation protocol designed to evaluate special educator performance related to instruction only. A special educator's job responsibilities may include, but are not limited to, the following: assessing students with disabilities; writing IEPs; incorporating appropriate goals into the IEPs; designing curriculum-based measures to track growth on individualized student goals; collaborate with related services personnel, administrators, parents, and outside agencies; create supplemental materials for use within general education settings; plan and prepare highly individualized lessons; and monitor

overall progress for a caseload of students. Any comprehensive teacher evaluation system should take a special educator's varied job responsibilities into consideration, and this study does not intend to argue that the RESET tool be used as the sole evaluative measure of special educator teaching quality. Rather, the tool is designed to evaluate instructional strategies that special educators should be using and are distinguishable from those used by general educators, but are subsumed within the larger framework of effective teaching strategies expected of all teachers. The RESET is a tool that can supplement existing rubrics so that when districts incorporate observation protocols into their evaluation systems, special education instructional practices are equitably represented.

Study Aims, Research Questions, and Hypotheses

This study used a mixed-method design that involved quantitative analysis of administrator ratings of special education instruction, as well as qualitative analysis of administrator and special education teacher interviews to gain insight into the formal teacher evaluation process as it applies to special educators. The study was guided by the following four aims:

1. **Aim 1**: Examine administrator rater reliability using generalizability theory. Traditional inter-rater agreement measures like Cohen's kappa are insufficient because they only attend to one source of variance (i.e., the rater). Previous studies of teacher evaluation measures have indicated that variance in obtained scores typically comes from the lessons (i.e., how much we can generalize from a teacher's performance on one lesson/occasion to the next), from raters (i.e., how much of the score is dependent on which rater evaluates the lesson), from teachers, and from error (Erlich & Shavelson, 1976; Hill, Charalambous, & Kraft, 2012; Ho & Kane, 2013). Generalizability theory systematically

examines several sources of variance (i.e., teachers, raters, lessons, interactions, and measurement error) that can affect the consistency of rater scores.

a) ***Question 1***: Can school administrators without previous experience in the field of special education use observation rubrics to reliably score video-recorded instructional segments of special education teachers?[6]

2. **Aim 2**: Investigate optimal data collection conditions (i.e., number of raters and number of observed lessons) for a desired score reliability. This is important given that school districts have limited time and finite resources; determining conditions for practical application helps guide school districts in their implementation of the tool.

a) ***Question 2***: What are the number of raters and number of observed lessons necessary for a desired score reliability?

. *Hypothesis 2*: A desired score reliability can be obtained when four raters observe four lessons.

3. **Aim 3**: Examine whether school administrators without previous experience in the field of special education systematically score special education teachers higher or lower on observation rubric items when compared with trained raters who have experience in special education instruction.

a) ***Question 3***: Do school administrators without previous experience in the field of special education score the instructional segments of special education teachers higher or lower on observation rubric items when compared with trained raters who have previous experience as special education teachers?

[6] No hypothesis is provided for Question 1. According to Shavelson and Webb (1991), G studies involve variance component estimation and interpretation, and are not used to formally test hypotheses.

. *Hypothesis 3*: School administrators without previous experience in the field of special education will score the instructional segments of special education teachers higher, on average, than trained raters who have experience in special education teaching.

4. **Aim 4 (qualitative)**: Explore how school administrators without previous experience in the field of special education feel about the evaluation process as it applies to special educators, how special educators also feel about the evaluation process, and how both the administrators and special educators perceive the RESET rubric items in terms of their practical application in a school setting.
    a) ***Question 4a***: How do school administrators without previous experience in special education perceive the process of evaluating special educators, including the administrators' previous experiences, training, and beliefs regarding instructional practices?
    b) ***Question 4b***: How do special education teachers perceive the process of being evaluated by their school administrators, including the teachers' previous experiences, issues of equitability, and suggestions for improvement?
    c) ***Question 4c***: How do both the administrators and special education teachers feel about the RESET rubric items as valid measures of special education instructional practices and their potential for use in the actual school setting?

This study included 19 teacher participants, who each contributed three videotaped instructional segments and were evaluated by three raters. Previous studies of teacher evaluation measures and generalizability explanations have established that smaller sample sizes are acceptable for research purposes (Erlich & Shavelson, 1978; Hill et al., 2012; Shavelson &

Webb, 1991). Using a similar study design, Erlich and Shavelson (1978) included five teachers, three occasions per teacher, and three raters; Hill et al. (2012) included eight teachers, three lessons per teacher, and nine raters; and Semmelroth and Johnson (2013) used a two-facet partially-nested design in which five raters evaluated nine teachers who were videotaped on three occasions.

CHAPTER TWO: METHOD

Teacher Participants

A total of 19 special education teachers from urban and suburban schools in California and Idaho participated in the study. The RESET rubric items are designed to be used across various special education settings, grade levels, and subject areas provided that the observed classroom includes evidence-based instructional strategies; as such, teacher participants in this study were not limited to specific grade levels, subjects, or settings. The teachers did, however, need to be including lessons that were based on approved curriculum and that utilized evidence-based practices for supporting SWDs in academic settings. The special education teacher participants from Idaho were selected from an existing database of 21 teachers who were recruited by Boise State University for participation in previous studies using the RESET observation tool. Each teacher contributed video data files over the course of two academic years (2011-2012 and 2012-2013). The final 12 special education teacher participants from Idaho were selected for this study based on the quality of the video file (e.g., teacher instruction could be heard and the teacher was visible), the teacher as primary deliverer of instruction (i.e., instruction was not led by a paraprofessional), and the availability of at least three video lessons that met the aforementioned criteria. The video data files for the Idaho participants were captured using the Teachscape Reflect system and stored in the Teachscape secure online database. These participants provided consent for their video data to be used for future research studies; demographic information about the 12 Idaho participants is included in Table 1.

Seven special education teachers from California participated in the study; demographic information about the seven California participants is also included in Table 1. Teachers were eligible to participate in the study if they possessed a valid California education specialist teacher

credential, were delivering instruction to SWDs in a classroom setting at the time of data collection, and worked predominately with students with mild to moderate disabilities. Special education teachers could participate if they taught in co-taught classrooms wherein both the special education teacher and general education teacher provided instruction as long as the special education teacher delivered at least 50% of the daily classroom instruction. Special education teachers were also eligible to participate if at least one period of their instructional day involved teaching students with mild to moderate disabilities in a pull-out/resource classroom in a public school. The special education teacher participant pool could also include those who taught in self-contained classrooms in public, nonpublic, or residential treatment facilities. Teacher participants had to have completed at least one academic year of teaching SWDs in an instructional setting. Special education teachers were not eligible to participate if their job duties involved consulting, case managing, or other administrative tasks for the entire school day and they did not provide any direct instruction to SWDs. Special education teachers were also ineligible to participate if they taught classrooms of students with severe disabilities such that instruction was focused on daily living and transition skills and the students were not on a diploma track.

After obtaining study approval from a university-based institutional review board, this study's principal investigator recruited California teacher participants by making direct contact, via phone or email, with teachers who met the inclusion criteria. Potential participants were identified through teacher networks or were individuals with whom the principal investigator had a previous professional relationship. After potential participants were identified, a formal letter was sent to each teacher's school district requesting permission to video record several of the

teacher's lessons. Permission to be on each teacher's school site was also obtained from a site administrator.

Rater Participants

School administrators were recruited for participation as raters of the teachers' instructional videos. Administrators could be included in the study if they were retired or no longer serving in an administrative capacity in any school district so as to protect the confidentiality of the teacher participants and ensure that the teachers would not incur any potentially negative consequences from being rated by an individual who supervises them currently or could do so in the future. The administrators also had to be former school principals or assistant principals as those roles represent the professionals who typically perform teacher evaluations. The participant raters were required to have at least five years of experience within the field of school administration and at least one year of experience evaluating teacher performance using observation rubrics. The participant raters must not have worked in the field of special education in any capacity. While some administrators may have a working knowledge of special education due to exposure within the school environment, participants could not be formally educated in the field of special education (i.e., hold an education specialist credential and/or have obtained a masters degree in special education) and could not have been employed as a paraprofessional, school psychologist, speech/language pathologist, behavior analyst, or special education teacher.

The principal investigator contacted potential raters directly via phone or email. If potential raters met the inclusion criteria, they were consented for participation in the study. A total of three former school administrators consented to and participated in the instructional video rating session. The three raters included the following: one male (Rater 1) with 28 years of

experience as an elementary and secondary school principal in Los Angeles County, and a total of 28 years of experience evaluating teachers; one female (Rater 2) with 14 years of experience as a school principal at five different elementary schools in Los Angeles County, and a total of 19 years of experience evaluating teachers; and one male (Rater 3) with five years of experience as an elementary school principal in Los Angeles County, two years of experience as a district administrator, and a total of seven years of experience evaluating teachers. All three participants were retired from their positions as school administrators and were no longer involved in evaluating teachers in any formal capacity.

Materials and Procedure

As previously mentioned, video data files were collected from the 12 Idaho teacher participants during the 2011-2012 and 2012-2013 academic years. The video data files were stored on the Teachscape secure online database, were password-protected, and could not be downloaded. Video data files of special education teacher instruction in California were collected across four school districts in Los Angeles County. Over the course of twelve weeks (April 2014-June 2014), the principal investigator collected three video recordings of direct instruction provided by each of the seven special education teacher participants. Although the length of each video varied, the video observation represented what each observed teacher self-identified as one "lesson."

Three types of video recording technology were used to capture the instructional segments. The first was the Teachscape Reflect system, which was the same technology used to capture the Idaho video data. The Teachscape video capture system consists of two cameras: (a) a 360-degree camera that allows the observer to pan and zoom on various components of the classroom environment, and (b) a fixed position camera, also referred to as a "board cam," which

is focused on a classroom board. Two California teachers, who had a greater number of students and were primarily instructing in whole-group settings, used the Teachscape system to capture their video data. Since the Teachscape system is not portable and must remain stationery, it was less ideal for teachers who worked with small groups of students at various locations within the classroom. The remaining five teachers in California used either a camcorder or a Flip Video Camera, each with a tripod, to capture their video data. These five teachers were working in smaller spaces or needed to move around the room, which necessitated a smaller camera system than that provided by Teachscape.

The principal investigator demonstrated how to use the camera and assisted in setting up the necessary equipment in each teacher's classroom. Researchers involved in collecting the video data in both Idaho and California were not present during the video recordings in any of the participating teachers' classrooms so as not to disrupt the classroom environment or the teachers' instruction. Teachers were instructed to turn the camera equipment on prior to the start of a self-identified lesson, and could turn the equipment off after the completion of the lesson. Parent consent and student assent were obtained for all students who would be visible on the video recordings.

All video data files were assigned a code according to the following scheme: one number representing the school district (e.g., 1), one letter representing the teacher from that specific district (e.g., A), and another number representing the lesson from that teacher (e.g., 3). The code 3B2, for example, indicated that the video came from the third district in the dataset, the second teacher from that district, and the second lesson from that particular teacher. A total of 57 videos were coded and stored digitally either on the Teachscape system, which requires a purchased license to access, or the Common Collaboration and Learning Environment (CCLE),

which is the secure online database for the University of California, Los Angeles (UCLA). Any videos recorded using the camcorder or Flip Video Camera were stored on CCLE, were password-protected, and were only accessible for viewing during the five-day rating session.

Once all video data files were collected and successfully stored, raters were hosted in the computer lab of a comprehensive public high school in Los Angeles County for a five-day rating session (June 2014). On the first day of the rating session, raters received a full day of training on seven of the RESET tool's evaluation rubrics. Training began with a discussion of the instrument and the performance levels for each of the seven rubric items, followed by video examples of special education instruction. Raters independently scored practice videos, discussed their scorings with each other, and also received feedback from the trainer. At the end of the training, raters were required to complete a certification process similar to that used in the MET Project (Kane & Staiger, 2012). Raters were required to independently score pre-scored videos, and needed to obtain at least a 50% exact match to expert scores. Any discrepant scores had to be no more than one point from the expert score. All three raters successfully completed the certification process and were able to move on to independent scoring on the second day of the rating session.

During independent scoring, raters were seated away from one another, were given headphones to wear, and were seated in front of two computer monitors: one to view the instructional video, and another to complete scoring via Qualtrics. The 57 videos were split into two sets: one with 27 videos (Set 1; three videos each from nine teachers) and the other with 30 videos (Set 2; three videos each from 10 teachers). Raters 1 and 2 completed scoring for Set 1, and Raters 1 and 3 completed scoring for Set 2. Each rater's list of videos was randomized so that no two raters viewed the video data files in the same order. The videos ranged in length

from 13 minutes and eight seconds to 39 minutes and 27 seconds; the average length of video was approximately 26 minutes.

Rubric items. Seven RESET rubric items were selected to represent the previously discussed definition of effective special education instruction. The rubric items were evaluated and revised for clarity, with the intention of making additional improvements to each of the items after receiving feedback from the raters in this study. The seven rubric items included the following: (a) *articulation of lesson objective*, (b) *teacher communication*, (c) *sequencing*, (d) *scaffolding*, (e) *skill development*, (f) *student engagement*, and (g) *student practice and review*. Each rubric item included the rating scale and indicators for each level. The instructional segments were rated on a scale from 0 (absence of indicators during the observation) to 3 (displayed all indicators during the observation). A description of each rubric item is included in Table 2.

The raters were also given an additional four rubric items to use during the rating session. These four rubric items represented a whole-lesson evaluation of the instructional segment, and were an exact match to four rubric items used in previous ratings of the Idaho teachers only. Raters were not provided with any training on these items for two purposes: (a) the scores would then reflect how these school administrators would rate special education teachers on broad categories when the administrators were not specifically trained to rate special education instruction, and (b) the scores could be compared to those used in previous studies of the Idaho dataset, which included trained special educators as raters of the special education teachers' instructional practices. The additional four rubric items included the following: (a) *effective use of time*, (b) *the teacher appears to have a solid understanding of the content*, (c) *the teacher*

implements effective instructional practices, and (d) *the teacher effectively responds to student needs*.

Interviews. Each of the raters was interviewed individually after completing the rating scales for all assigned teacher videos. Semi-structured interviews were conducted with each rater using an interval protocol consisting of open-ended questions covering the following topics: (a) the raters' general experience observing and evaluating special education teachers in their previous districts, (b) perceptions of the RESET rubric items as appropriate and valid measures of special education instructional practices, (c) beliefs regarding the applicability of the rubric items in the actual school setting, and (d) raters' opinions as to the quality and amount of training they received both in their previous districts and during the rating session. All interviews were conducted in a private room, were approximately 45 minutes in length, and were recorded and transcribed.

After submitting their video data files, the California teacher participants were again contacted regarding a potential interview. Teacher participants were informed that participation in the interview portion of the study would be entirely voluntary, and no personally identifying information would be attached to the interviews. Five teacher participants agreed to be interviewed and were re-consented. The five teachers included two males and three females, they represented four separate school districts, and they ranged in teaching experience from six to 13 years. Semi-structured interviews were conducted with each of the five teachers in a private, mutually agreed upon location. An interview protocol was developed and consisted of open-ended questions covering the following topics: (a) their general experience being observed and evaluated by a school administrator, (b) their beliefs as to whether each RESET rubric item used in the study was an appropriate and valid reflection of effective special education

instructional practices, (c) suggestions for improving the evaluation of instructional practice as it applies to special educators, and (d) suggestions for rubric items to be used in an observation protocol for evaluating special educators. All interviews were conducted by phone or in a private room, were approximately 25 minutes in length, and were recorded and transcribed.

Remuneration. Each California teacher participant who contributed at least three video-recorded instructional segments received a $50 Target gift card. The gift cards were given to participating teachers in person after the final recorded video observation and at the same time that the principal investigator retrieved the camera equipment from the classroom. Participant raters each received $500 in cash after completion of the five-day rating session, including the individual interview.

CHAPTER THREE: ANALYSIS

Aim 1

The first aim of the study was to examine administrator rater reliability using a generalizability study (G study) design. Similar studies, including those using the MET data, have used G studies to estimate sources of error and improve evaluation systems by varying the conditions (i.e., number and type of raters and number and length of lessons) that could be applied in actual school settings (Ho & Kane, 2013; Kane & Cantrell, 2013). The purpose of a G study "is to evaluate the characteristics of a given measurement procedure and to estimate measurement precision" (Cardinet, Johnson, & Pini, 2010, p.11). In this study, the measurement procedure was the rating of special education instructional practices using observation rubrics, and the measurement precision was determined by quantifying the relative significance of error contributors (Cardinet et al., 2010).

The first step in conducting a G study is to identify the facets (i.e., factors) and their inter-relationships that are at play in the measurement process (Cardinet et al., 2010). Facets are either "crossed" (i.e., every level of one facet is combined with every level of the others), or "nested" (i.e., a facet is associated with one and only one level of another) (Cardinet et al., 2010). This study included four facets: three instrumentation facets (raters, lessons, and rubric items); and one differentiation facet (teachers), which was the object of measurement. Raters (*r)* and rubric items (*i*) were both crossed facets because all raters in the study provided ratings on all seven rubric items for the observations/lessons in their datasets. Lessons (*l*) was a nested facet because not all teachers were delivering the same three lessons. Although all teachers contributed three lessons, a fully crossed design would require the standardization of lessons such that every teacher was teaching the same thing (Meyer, Liu, & Washburn, 2013). Since teachers each

contributed three lessons of their choosing, the lessons (*l*) facet could only be associated with one teacher; consequently, lessons (*l*) was nested within teachers (*t*), *l:t*, in the design. The final design was a three-facet, partially-nested design wherein lessons (*l*) was nested within teachers (*t*), *l:t*, and crossed with raters (*r*) and rubric items (*i*), (*{l:t} x r x i*) (Shavelson & Webb, 1991). The G study provides a decomposition of the variance over *t*, *l*, *r*, and *i* of single teacher-lesson-rater-item scores[7]:

$$\sigma^2(X_{tlri}) = \sigma^2(t) + \sigma^2(l,tl) + \sigma^2(r) + \sigma^2(i) + \sigma^2(tr) + \sigma^2(ti) + \sigma^2(rl,trl) + \sigma^2(il,til) + \sigma^2(ri) + \sigma^2(tri) + \sigma^2(ril,tril,e),$$

which is referred to as "total variance"—analogous to total sums of squares in analysis of variance (ANOVA)—as it is the sum of the G study variance components (Brennan, 2000). A description of each variance component is included in Table 3.

The second step in the estimation design is to determine if the facets will be treated as "fixed" or "random." A facet is fixed if all of its levels are featured in the dataset and no sampling of levels has occurred (Cardinet et al., 2010). Conversely, facets are "said to be random when the levels taken into consideration are randomly selected from the respective population or universe of interest" (Cardinet et al., 2010, p. 18). In this study, the facets of teachers (*t*), lessons (*l*), and raters (*r*) were random because each was randomly selected from the respective population of interest. The rubric items (*i*) facet, however, was determined to be fixed because the teachers were scored on specific rubric items designed to reflect the effective instructional practices of special educators; these rubric items were not randomly sampled from a larger universe of rubric items. Although other observation instruments and rubric items exist,

[7] With nested designs, some sources of variability cannot be estimated separately due to confounding (Shavelson & Webb, 1991). For example, the variance component for lessons $\sigma^2(l)$ is confounded with the teacher-by-lesson interaction $\sigma^2(tl)$. The variance component for confounded effects is denoted with a comma, such as in $\sigma^2(l,tl)$.

the rubric items used in this study were explicitly identified to represent a construct of interest, namely effective special education instruction, and do not exist within a larger pool of rubric items that represent the same construct.

The consequence of a facet being fixed is that any potential contribution to error variance from this source is eliminated (Cardinet et al., 2010). In order to ensure that the rubric items (*i*) facet was not a relatively large contributor to error in the measurement process, a preliminary G study was performed with the rubric items (*i*) facet treated as random. Results of that analysis revealed that rubric items (*i*) had an estimated 2.4 percent of the total variance, which suggested that the rubric items were not a significant contributor to error variance in the measurement process. In the subsequent G study analyses, rubric items (*i*) was determined to be a fixed facet in the estimation design and its contribution to error variance was eliminated.

The three-facet, partially-nested design was applied to two separate G studies, which were each a subdivision of the larger dataset. According to Chiu and Wolfe (2002), raters seldom evaluate all objects of measurement due to the time-consuming nature of the task. In order to save operational time and make the task manageable for raters, the data can be organized into batches and then randomly assigned to raters (Chiu & Wolfe, 2002). Ho and Kane (2013), for example, conducted a G study using four lessons from each of 67 teachers (i.e., 268 total lessons), but the number of lessons scored per observer was limited to 24. In this study, the 57 videos were divided into two batches, or smaller datasets, which were then randomly assigned to a pair of raters. The first G study included the data from Set 1, which was a total of 30 lessons from 10 teachers; Rater 1 and Rater 2 observed and independently rated all 30 lessons using the seven pre-selected rubric items. The second G study included the data from Set 2, which was a total of 27 lessons from nine teachers; Rater 1 and Rater 3 observed and independently rated all

27 items using the same seven rubric items. G studies were performed on each dataset, and the variance components were then weighted by their sample sizes across all subsets according to the following formula: for any particular effect *f* one obtains,

$$\hat{\sigma}_f^2 = \bar{\sigma}_f^2 = \frac{\sum_{t=1}^{T}\sum_{s=1}^{S_t} n_{p,t,s}\hat{\sigma}_{f,s}^2}{\sum_{t=1}^{T}\sum_{s=1}^{S_t} n_{p,t,s}},$$

s is the *s*th data subset, *t* is the *t*th structural design (e.g., nested or crossed) and $n_{p,t,s}$ is the number of examinees in the *s*th data subset of the *t*th structural design (Chiu & Wolfe, 2002, p. 327). All analyses were performed using EduG, a software program specifically designed for G studies.

Aim 2

The second aim of the study was to investigate optimal data collection conditions (i.e., number of raters and number of observed lessons) for a desired score reliability. After conducting a G study, a decision study (D study) can be utilized to further analyze the characteristics of the measurement procedure. A D study involves changing the measurement procedure (e.g., increasing the number of raters), to explore and evaluate different improvements that could be made (Cardinet et al., 2010). This study was concerned with the number of lessons and the number of raters, as these two characteristics of the procedure are potentially the most time consuming and costly aspects of teacher evaluation processes. The D study performs "what if?" analyses (Cardinet et al., 2010) to determine what is both optimal as well as feasible in the practical application of a measurement; the D study was used in this study to determine the minimum number of raters and observed lessons necessary to sufficiently increase reliability and reduce error.

In a D study, the variance components that are contributing to measurement error are identified, and the number of conditions for each facet are optimized to achieve a desired score reliability. Measurement error and generalizability (reliability) coefficients are interpreted for the purpose of making two types of decisions: *relative* and *absolute* (Shavelson & Webb, 1991). If for example, a decision maker is interested in how well a teacher performs relative to his or her peers, a *relative* decision would be appropriate. If, however, a decision maker is concerned with how well teachers perform against a pre-determined criteria (e.g., achieving a level of mastery in their instructional practice), then an *absolute* decision would be considered.

Relative, *absolute*, or both decisions can be interpreted depending on the interests of the decision maker. In this study, the reliability coefficients for both *relative* and *absolute* decisions were compared. For *relative* decisions in a three-facet, partially nested design, only the variance components representing interactions with teachers contribute to error; in this study those variance components were $\sigma^2_{l,tl}, \sigma^2_{tr}, \sigma^2_{rl,trl}$. For *absolute* decisions, all variance components except for the object of measurement contribute to measurement error, which in this study included $\sigma^2_{l,tl}, \sigma^2_{r}, \sigma^2_{tr}, \sigma^2_{rl,trl}$. Variance components including the rubric items (*i*) facet were not included because rubric items (*i*) was a fixed facet and, consequently, did not contribute to measurement error.

Optimization procedures, which involved varying the number of lessons and number of raters, were performed in EduG. Since the G study included three lessons and two raters, the options for optimization involved increasing and decreasing both the lessons (*l*) and raters (*r*) facets to achieve an acceptable reliability coefficient, which is equal to at least .80 (Cardinet et al., 2010). A total of 16 options were selected, which included all possible combinations of one to four raters rating one to four lessons. D studies were performed on each of the G studies (i.e.,

Raters 1 and 2, Raters 1 and 3) separately due to the discrepant results in rater variance components.

Aim 3

The third aim of the study was to examine whether school administrators without previous experience in the field of special education systematically score special education teachers higher or lower on specific rubric items when compared with trained raters who have experience in special education instruction. Data for this analysis included ratings using the four previously mentioned rubric items (i.e., *effective use of time*, *the teacher appears to have a solid understanding of the content*, *the teacher implements effective instructional practices*, and *the teacher effectively responds to student needs*). In a previous study, trained raters with special education teaching experience used the four rubric items to rate the video-recorded instructional segments of special education teachers in Idaho. In the current study, administrators without experience in special education used the same four rubric items to rate the same instructional segments.

From the larger database of Idaho teachers, six special education teachers were selected for this analysis because they received scores on all four rubric items for at least three lessons, and ratings were performed by the same pair of raters for all lessons. Each of the six teachers, then, had a score from each of two special education teacher raters and two administrator raters per rubric item per lesson. Each teacher's scores were averaged across all rubric items and raters to achieve an average score per lesson. A one-way repeated measures ANOVA with three levels (lessons) was performed with each of the teacher's average scores for Lesson 1, Lesson 2, and Lesson 3 to determine if there were significant differences between the teacher's scores from one lesson to another. Results indicated that there were no significant differences across lessons for

each teacher; consequently, the scores for each lesson were collapsed to achieve a holistic score per rubric item per teacher.

To determine if there was an effect of type of rater, a 2 (rater type) x 4 (rubric item) repeated measures ANOVA was conducted. The dependent measure was the teacher's composite score (averaged from two raters and across three lessons) on a four-point scale (0-3). *Post-hoc* analyses were performed to determine the nature of the interaction between type of rater and rubric item. Four paired samples *t*-tests, using a Bonferroni correction to maintain an alpha level of 0.05, were conducted to compare the special education teacher rater scores to administrator rater scores on each of the four rubric items.

Aim 4 (Qualitative)

The final aim of the study was to explore how school administrators without previous experience in the field of special education felt about the evaluation process as it applied to special educators, how special educators also felt about the evaluation process, and how both the administrators and special educators perceived the RESET rubric items in terms of their practical application in a school setting. Semi-structured interviews were conducted with the three rater participants and five of the teacher participants, which resulted in a total of eight audio recordings. The audiotapes of the eight interviews were transcribed verbatim, assigned an alpha code, stripped of any personal identifiers, and sent to each participant for review and approval. Once approval was obtained from each participant, the transcriptions were prepared for analysis.

Analysis of the transcripts initially involved a grounded theory approach, which generates an explanation of a process, action, or interaction (Creswell, 2007). The grounded theory approach was appropriate because participants in the study had all experienced the same process, either as the individual conducting the teacher evaluation or as the teacher being evaluated;

developing a theory as it relates to these experiences helps to explain the practice and provide a framework for future research (Creswell, 2007). In keeping with grounded theory, the transcript analysis included three stages. First, open coding was used to code the data for major categories of information. A priori codes based on an existing theory were not utilized; rather, appropriate codes were determined inductively after a first reading of all transcripts. Each transcript was transferred to an Excel spreadsheet, where participant responses were appropriately segmented into smaller chunks/lines of data. The transcripts were read line-by-line, and extensive notes were taken on each segment of data. A thorough first reading of the transcripts and a review of the notes resulted in the identification of major categories of information. For example, some notes included the following statements: *no training for special ed evaluations*, *no training for teachers in general*, *what training should look like*, and *more training was needed*. These preliminary notes/codes were aggregated into the major category of *training*, which encompassed any statements participants made regarding their training both during the current study and in their previous school districts.

Once categories of information had been determined, axial coding was then applied. Axial codes were developed based on the categories determined through open coding, and were then included in a coding sheet to be used for the second reading of transcripts. The coding sheet included each axial code (e.g., *training*), a description of each code, and illustrative quotes from the transcripts to further clarify the code's intended meaning. A selection of codes and descriptions is included in Table 4. The coding sheet was reviewed and discussed, and minor refinements were made according to group consensus. For the rater/administrator interviews, a total of seven final codes were included in the coding sheet: (a) *experience*, (b) *training*, (c) *feedback* (to teachers), (d) *instructional practices*, (e) *evaluation/observation*, (f) *personal bias*,

and (g) *evaluation instruments*. For the teacher interviews, a total of nine final codes were included in the coding sheet: (a) *administrators*, (b) *evaluation*, (c) *fairness*, (d) *process*, (e) *modifications*, (f) *feedback* (from administrators), (g) *instruction*, (h) *professional development*, and (i) *suggestions for improvement*.

The principal investigator and a research assistant, who had extensive experience with qualitative research in the area of special education, independently coded the transcripts using the coding sheets. Transcripts were read line-by-line and each segment of text was assigned one primary code. If a coder believed an additional code could be applied, that code was added to an additional column of the transcript and noted for later discussion. Although coders were not limited to one code per segment of text, only the primary code was used to establish inter-rater agreement. After independent coding of each transcript, Cohen's (1960) kappa was calculated to determine rater agreement in the assignment of codes. Kappa is an appropriate measure of reliability; it indicates the proportion of agreement beyond that which could be expected by chance alone. According to Everitt (1996), kappa values above .60 are satisfactory and values above .80 are regarded as nearly perfect agreements.

After independent coding of the first transcript, kappa was calculated for levels of agreement, and the two coders compared their primary codes for each segment of text. Where disagreement occurred, the coders discussed their reason for applying a particular code, the description of the relevant code was clarified, and discussion continued until consensus was reached over the most appropriate code to be applied and/or necessary refinement of the coding sheet. This process was performed first for the administrator rater transcripts, and then repeated again for the teacher transcripts. Kappa values indicated that coder agreement improved after

each transcript as discussion resulted in increased consensus. Kappa scores for each transcript can be seen in Tables 5 and 6.

The percent agreement achieved across raters met established criteria for qualitative research (Boyatzis, 1998); however, the coders were less concerned with agreement and more with the discussion surrounding disagreements. While it was important that evaluators understood the coding categories and applied them systematically, percent agreement may be misleading in that two evaluators may assign the same code, but that code may not be the best reflection of the interviewee's intended meaning. Ultimately, a rich and in-depth analysis of qualitative data involves a discussion of seemingly divergent interpretations because those interpretations may actually reflect concordance on some level within a wider framework (Armstrong, Gosling, Weinman, & Marteau, 1997). Consequently, the evaluators spent considerable time discussing each of their codes and the participants' intended meaning so as to develop an in-depth analysis of themes within each transcript and across participants.

While the final stage in analysis should involve selective coding, which is the process of developing hypotheses regarding the interrelationships of the categories (Creswell, 2007), it became apparent that the emerging themes reflected a description of a common experience among the participants, and did not lend themselves to the development of a theory. Rather than continue with a grounded theory approach, which should culminate in a theoretical framework, the analysis shifted toward a phenomenological approach, which focuses on a description of the experiences of participants (Creswell, 2007). During the coding process, illustrative quotes were extracted from the narratives to support categories identified during axial coding. In phenomenology, the illustrative quotes are referred to as significant statements, the researcher

formulates meaning from the statements, and the formulated meanings are clustered into themes (Creswell, 2007).

After completing coding of all transcripts, the principal investigator and research assistant discussed significant phrases and sentences that represented each of the coding categories, and then formulated meaning from those statements. For example, one administrator stated the following: "We were given the evaluation instrument and then we were told to go down through it, and then you really learn from going in and doing it." This statement was interpreted to mean that the administrator received very little training on an evaluative measure in his school district, but was able to learn how to evaluate his teachers through experience. Table 7 includes selected examples of significant statements and related formulated meanings from the teacher transcripts.

Once relationships between the formulated meanings were determined, those connections were clustered into themes, which allowed for the emergence of four total themes common to the administrator participants, and four total themes common to the experiences of the teacher participants. Table 8 includes an example of one theme cluster with its associated formulated meanings. The theme clusters, as discussed in the results, represent an integrated and in-depth description of the experiences—related to the evaluation of special education teachers—of the administrator and teacher participants interviewed in this study.

Validation. Multiple validation methods were used to establish credibility and transferability of the data. The principal investigator had pre-established professional relationships with the participants in the study or with those who worked closely with the participants. These relationships resulted in an established trust and rapport, which made the participants comfortable with the interview process and more willing to provide detailed and thoughtful responses to the interview questions. Member checking techniques included

summarizing interview notes for the interviewee to ensure an accurate reflection of the interviewee's position, and providing a draft of the relevant transcript to each participating member for review and comment (Mertens, 2010). Data triangulation was accomplished through the use of one method (interviews) from multiple sources (different individuals and sites).

CHAPTER FOUR: RESULTS

Aim 1: Examine administrator rater reliability using generalizability theory.

As previously mentioned, two separate G studies were performed and the estimated variance components were weighted and averaged to obtain one decomposition of variance across all potential sources of measurement error. The results of each separate G study as well as the combined results can be seen in Table 9. The combined results will be discussed first, but the results of the initial two G studies will also be discussed due to their discrepant outcomes.

The estimated variance component for the object of measurement (teachers) was .1893, or 18.9% of the total variance. This was the second largest variance component among the averaged weighted estimates, and it indicates that teachers varied somewhat systematically in their observed performance during the instructional lessons. The largest variance component was that for the lesson-by-rater interaction (.2628, 26% of the total variance). Since lessons are nested within teachers, it is impossible to separate the lesson-by-rater interaction from the three-way interaction between teachers, lessons, and raters. We do not know whether some raters scored some lessons higher than others, or whether the relative standing of lessons varied by teacher and by rater. The third largest variance component was that for the highest order interaction effect and/or the residual error, (.1878, 18.8% of the total variance). This component includes the rater-by-item-by-lesson interaction, which is confounded with the teacher-by-rater-by-item-by-lesson interaction, and is also confounded with unmeasured variation. The relatively large residual component (18.8%) indicates that a sizeable amount of variation is due to these confounding sources.

The component for raters was relatively smaller than the previously mentioned sources, but still large enough to indicate substantial variability among rater scores (.1048, 10.5% of the

total variance). This indicates that raters differed in how they scored the observations, averaging over teachers, lessons, and items. The percent of variance for lessons (confounded with the teacher-by-lesson interaction), the teacher-by-rater interaction, the teacher-by-item interaction, and the rater-by-item interaction were 4.6, 4.9, 4.8, and 4 respectively. While non-negligible, these sources are relatively much smaller than the other sources of variance in the design.

When combined, the results of the G studies suggest relatively low rater variability when compared with other sources of variance. The percent, however, was substantial enough (10.5%) to indicate weak rater reliability. When analyzing the G studies separately, however, the results are much different. In the G study performed with scores from Raters 1 and 2 (nine teachers and 27 total lessons), the highest variance component was that for teachers (31.3%), which indicates systematic variability among teacher performance. The variance component for raters was extremely low both in the relative and absolute sense (0.2% of total variance). This result suggests that these two raters were well calibrated in that they applied the same part of the scale and did so consistently at each observation. There was, however, a large variance component from the lesson-by-rater-interaction (21.1%), which is confounded with the teacher-by-lesson interaction. This result suggests that although raters consistently applied the same standards overall, there was systematic disagreement in the relative standing of lessons, confounded with teacher scores. Finally, the highest-order interaction effect, along with residual error, contributed a relatively large portion of variance (17.1%), suggesting error from several interactions and unaccounted for sources of variance.

The results from Raters 1 and 3 (10 teachers and 30 total lessons) revealed a remarkably different outcome from that of Raters 1 and 2. Similar to Raters 1 and 2, the variance components for the lesson-by-rater interaction (31.6% of total variance) and the highest-order

interaction effect (20.5% of total variance) were both very large. While the component for raters was negligible for Raters 1 and 2, the variance component for raters in the Raters 1 and 3 G study was the second highest source of variance (20.9% of total variance). This result suggests that these two raters were not well calibrated and there was a high degree of variability among their scores. The variance component for teachers, which should ideally be the highest source of variance according to this measurement design, was relatively small (only 6.4% of total variance). Raters 1 and 3 would not, then, be considered reliable when rating the instructional practices of special education teachers. The distribution of all three rater scores across each of the seven rubric items can be seen in Figures 1, 2, and 3.

Aim 2: Investigate optimal data collection conditions (i.e., number of raters and number of observed lessons) for a desired score reliability.

Table 10 and Figures 4 and 5 show the relative G coefficients and standard error of measurement (SEM) for Raters 1 and 2. For *relative* decisions (e.g., evaluating a teacher's lessons for the purpose of determining performance relative to his or her peers), acceptable levels of reliability were achieved with three raters observing three (.82) or four lessons (.84), and with four raters observing two (.81), three (.85), or four lessons (.87). While .80 is considered to be the minimum level necessary to indicate reliability in the measurement process (Cardinet et al., 2010), several rater/lesson combinations achieved a *relative* G coefficient approaching .80, which may be sufficient for practical purposes. For example, with only two raters, a *relative* G coefficient of .72 was obtained with two lessons and .77 with three lessons. The results indicate that with only one rater, *relative* G coefficients for any of the lesson combinations are well below .80; however, two raters observing two lessons is an arguably reliable combination for school

districts with limited resources to staff additional raters and/or provide time for additional classroom observations.

For *absolute* decisions (e.g., dismissing low performing teachers or rewarding high performing teachers regardless of their relative standing), acceptable G coefficients were obtained with three raters observing three lessons (.82) and four lessons (.84), and four raters observing two lessons (.81), three lessons (.85), and four lessons (.87). Similar to the *relative* coefficients, several other rater/lesson combinations approached .80; however, *absolute* decisions may require a stricter cutoff given that the performance evaluation may be tied to reward or dismissal. For both *relative* and *absolute* decisions, the highest G coefficients and lowest SEMs resulted from four raters evaluating four lessons (.87, .24).

What should be noted from the D study results for Raters 1 and 2 is that—for both *relative* and *absolute* decisions—there is a remarkable increase in utilizing two raters as opposed to one. While there are small increases in the reliability coefficients with each additional rater, the most substantial gain results from moving from one rater to two. The same is true of observing two of a teacher's lessons as opposed to one; each additional observed lesson results in a higher reliability coefficient, but the greatest gain lies in increasing the number of observations from one to two. Similarly, error is reduced with the addition of each additional rater and lesson, but a greater reduction occurs from utilizing two raters as opposed to one, and by observing two lessons as opposed to one. One rater observing one lesson, for Raters 1 and 2, resulted in a SEM of .57, which indicates a probable error of .57 on a 0-3 scale. The SEM decreases to .43 with two raters, .37 with three raters, and .34 with four raters. If, for example, two raters are used, the SEM is .43 with one observed lesson, but decreases to .34 with two lessons; the SEM decreases

to .30 when two raters observe three lessons and to .28 with four lessons, both much smaller decreases than from one observed lesson to two.

The D study optimization results for Raters 1 and 3 can be seen in Table 11 and Figures 6, 7, 8, and 9. As expected, the G coefficients are far lower due to the greater variability in these raters' scores. For *absolute* decisions, acceptable G coefficients were not obtained for any of the lesson/rater combinations. The highest *absolute* G coefficient and lowest SEM was achieved with four raters evaluating four lessons (.43, .27), but this value is well below what would be considered acceptable. This result suggests that, for less reliable rater pairs, *absolute* decisions such as retention and tenure cannot be made with confidence due to discrepant rater scores. With lower reliability, there is greater risk that the score assigned by one of these raters is not a true reflection of the teacher's classroom performance, but may instead be the result of inadequate training, a weak instrument, rater bias, or other sources of error.

The *relative* G coefficients were greater than the *absolute* coefficients for Raters 1 and 3, though still lower than what was achieved with Raters 1 and 2. The highest *relative* G coefficients resulted from two raters observing four lessons (.53), three raters observing three lessons (.54), four raters observing three lessons (.59), three raters observing four lessons (.61), and four raters observing four lessons (.66). For *relative* decisions, these combinations come closer to acceptable levels of reliability. What is evident from these results is that the higher the variability in rater scores, the greater number of raters and lessons necessary to achieve reliable results. For both *relative* and *absolute* decisions, fewer resources can be expended if raters are well trained and calibrated on the instruments used.

Aim 3: Examine whether school administrators without previous experience in the field of special education systematically score special education teachers higher or lower on

observation rubric items when compared with trained raters who have experience in special education instruction.

To determine if there was an effect of rater, a 2 (rater type) x 4 (rubric item) repeated measures ANOVA was conducted; results are presented in Figure 10. The analysis revealed a main effect of rater, $F(1,5) = 12.58$, $p < .05$, $n_p^2 = .72$, such that administrators without special education experience rated teachers significantly higher on rubric items than the trained special education raters. The analysis also revealed a main effect of rubric item, $F(3,15) = 13.96$, $p < .001$, $n_p^2 = .74$, and a significant interaction between rater and rubric item, $F(3,15) = 6.30$, $p < .05$, $n_p^2 = .56$.

Post-hoc analyses were performed to determine the nature of the interaction between type of rater and rubric item. A test of simple main effects, using a Bonferroni correction to maintain an alpha level of .05, was conducted to compare the special education teacher rater scores to administrator rater scores on each of the four rubric items. The analysis revealed that the administrator raters scored teachers significantly higher ($M = 1.94$, $SD = 0.17$) than the special education teacher raters ($M = 1.03$, $SD = 0.29$) on Rubric Item 2 (*the teacher appears to have a solid understanding of the content*) ($t_{(5)} = 6.82$, $p = .001$, $d = 3.87$). There were no significant differences between administrator rater scores and special education teacher rater scores on Rubric Item 1 (*effective use of time*) ($t_{(5)} = 1.63$, $p = .163$, $d = 0.98$), Rubric Item 3 (*the teacher implements effective instructional practices*) ($t_{(5)} = 3.27$, $p = .022$, $d = 0.94$), and Rubric Item 4 (*the teacher effectively responds to student needs*) ($t_{(5)} = 2.87$, $p = .035$, $d = 1.72$).

Although a statistically significant difference was not detected on Rubric Items 1, 3, and 4 in the post-hoc pairwise comparisons, an examination of *p* values alone is misleading. Due to the small sample size in this analysis, the effect size of each difference is a stronger indicator of

the practical significance of the mean differences between the groups of raters on each rubric item. Cohen's *d* for Rubric Items 1, 2, 3, and 4 was 0.98, 3.87, 0.94, and 1.72 respectively, which are all large effect sizes (Cohen, 1988). These results suggest that there are meaningful differences between the scores of administrator raters and trained special education teacher raters on all four rubric items, and a larger sample size could result in statistically significant mean differences between the groups.

Aim 4: Explore how school administrators without previous experience in the field of special education feel about the evaluation process as it applies to special educators, how special educators also feel about the evaluation process, and how both the administrators and special educators perceived the RESET rubric items in terms of their practical application in a school setting.

Evaluating special educators from an administrator perspective. After systematic analysis of the three rater interviews, four themes emerged and are discussed below.

Theme 1: training matters, but experience matters more. When asked about training specific to evaluating special education instructional practices, the three administrators in this study remarked that they received no training prior to or during their time evaluating special educators in their respective school districts. They also remarked, however, that they received very little training on evaluating teachers in general. Rater 3 stated the following:

> In terms of specific training for a special education credentialed person, I had none. Zero. Zero training. And that's basically the same with the general education teachers as well. We didn't get an awful lot of training.

With very little informative instruction to guide them in the evaluation process, these administrators consistently stated that they learned how to evaluate instructional practice by

actually performing the teacher observations. The experience of observing and evaluating is what allowed them to learn the process and expectations, and develop their own skills over time. Each of the three raters depended on a combination of their own past experiences as classroom teachers, outside resources on evaluation that they sought on their own time, and the repeated observations of classroom instruction of many teachers across many points in time and over several years. The cumulative process of observing a variety of teachers across time and context provided a foundational knowledge of what these administrators believed to be the varied abilities and efficacy of teachers. Training they received, then, was not something that the administrators in this study believed to be a valuable, let alone present, part of the development of their teacher evaluation skillset.

While these administrators perceived their experience as evaluators to be the most valuable part of their ability to effectively evaluate teacher instruction, that is not to say they believed training to be unnecessary to the process. Although they themselves did not receive adequate training, they all stated that better training would have made them better evaluators. All three administrators in this study stated that they greatly benefited from the training process they experienced in this study, which consisted of reviewing and discussing each rubric item, viewing several videos of special educators' lessons, and engaging in rich dialogue about classroom instruction seen in the videos. Referring to the training procedures in this study, Rater 1 remarked:

> This would have been great if I would have had something like this 30 years ago.

Rater 2 stated:

> I think you learn from discussing with somebody else rather than sitting and doing it by yourself in a vacuum.

While the three administrators espoused a "learn by doing" mentality, they also recognized that good training on an instrument, in concert with collaboration and dialogue while learning the procedures, can be highly beneficial to strengthening administrators' skills in the process of evaluating teachers' instructional practices.

Theme 2: administrators without special education experience can evaluate special educators. Despite a lack of training and a formal background in special education, the administrators did not feel ill equipped to perform teacher evaluations generally and special educator evaluations specifically. All three administrators stated that they believed they had enough training and knowledge to evaluate special educators and provide them with feedback on their instruction. The administrators remarked that they either had some experience with students with disabilities in previous classrooms or school sites, and they had an understanding of what they believed to be good teaching in general. The combined knowledge and experience was enough for them to feel that they could meaningfully evaluate special educators, though with some limitations.

When asked whether she believed she could give special educators meaningful feedback on their instruction, Rater 2 stated the following:

> I think with some of them, yeah, with others they were working with kids that had more disabilities than I had had a lot of training in, so that was a little more difficult. Particularly, ironically, enough, it was the discipline that the teachers always had problems with. How do you keep kids focused and how do you keep them from running around…We worked a lot on strategies on how to get kids to be refocused.

Raters 1 and 3 stated similarly that they were able to provide special educators with meaningful feedback because they expected to see the same instructional strategies as general educators, albeit in varying instructional contexts and with smaller groups of students. They also provided feedback to teachers in the same general areas, like classroom management, as stated by Rater 1. Rater 1, however, acknowledged that an unfamiliarity with certain types of disabilities resulted in what she believed to be an inadequate knowledge base on her part. She stated that there were likely many more strategies available, and an increased knowledge about specific disabilities and relevant instructional strategies would have better enabled her to provide specific feedback to her special education teachers.

When asked whether they applied differential rating strategies when observing and evaluating special education teachers' instruction (for teachers at their school sites, not the teachers in this study), two administrators stated that they did not, and one administrator stated that he did. The two administrators who applied the same rating strategies for all teachers remarked that they held the same teaching standards for both general and special educators. Rater 1 stated:

> As far as their teaching, yes, [I evaluated special educators in the same way as general educators]. There were common teaching elements that I would look for in both special ed and general ed classrooms. For example, check for understanding; you can apply that to a large classroom and you can apply that to an individual kid.

When asked whether she felt she was more lenient or more stringent when evaluating special education instruction, Rater 2 stated that her expectations shifted only as a product of the teachers' level of training:

> I really think I expected if they had a credential that they would be able to be successful with the kids. Two of the ones that I remember were working on their special ed credentials so they were still learning. Those I probably cut a little slack because they were still trying to figure out what was going on.

Raters 1 and 2 did not feel they adjusted their expectations of a special educator's instructional practices purely because the teacher was a special educator; these administrators applied the same teaching standards and had the same expectations of all of their teachers.

Rater 3, however, acknowledged his own personal bias when evaluating special educators in the classroom:

> I always went into that very pro-special ed teachers because all that I had seen that they had done. And I was probably more forgiving in terms of everyone following a lesson plan that they had given me, recognizing that in the moment things would require their attention. I was probably pretty lenient…Give them credit for showing up for work. And that was always my perspective. Probably being more lenient and more understanding of the classroom environment and the unique challenges that they have.

Rater 3 stated additionally that he believed administrators have lower standards for academic rigor in classes of students with disabilities, and that classroom management is more of a priority. Although Rater 3 felt equipped for and capable of evaluating special education teachers in the classroom, he also felt that he was more forgiving in his evaluations.

Theme 3: good instruction is good instruction. When asked specifically about strategies and practices that make for effective special education instruction and that these administrators would expect to see in a classroom of students with disabilities, the three administrators

identified strategies that they believed made for good instruction in general regardless of the type of teacher or instructional setting. The administrators, then, had general expectations for good instruction that they felt applied to all teachers, including special educators. Rater 1, for example, stated the following about his expectations for and experiences with special education classroom instruction:

> The lesson should be articulated, the strategies and materials should be appropriate for the lesson, and most of the time the failure of a lesson was due to poor classroom management—kids are up running around, the teacher is writing and talking to the chalkboard.

In his experience, what made a lesson effective (i.e., articulating the lesson objective and using appropriate materials) were necessary strategies in both general and special educator classrooms. Similarly, what prevented a lesson from being effective (i.e., poor classroom management) could equally impact general and special educator classrooms.

All three administrators remarked that when observing the classroom practice of special or general educators, they expected to see a complete lesson that began with the articulation of the lesson objective, was connected to prior and future learning, had a clear sequence and purpose for the learning activities, and concluded with an appropriate activity. The administrators also expected the teacher to communicate clearly, incorporate a variety of activities, provide opportunities for practice, and encourage student engagement. All three administrators also vehemently stated that they expected teachers not to be tied to an instructional manual. While some schools and districts utilize curriculum that includes scripted lessons, the administrators in this study expected teachers to deviate from scripted lessons when necessary. Although some special educators are required to deliver standardized lessons or

interventions, the administrators still wanted to see a variety of instructional practices, active student engagement, and an awareness of students' needs that may require a departure from the script.

The one instructional practice that the administrators noted as separate or more important in a special education classroom was differentiation/individualization. Rater 3 commented that "differentiation should be the mainstay of a special education classroom." The administrators expected special educators to be responsive to the individual needs of the students, and to organize the instructional lessons around those individual needs:

> You may have 12 to 15 kids in there and they are all at very different places [academically]. Most of the time, you have an instructional aide, and you've got to figure out a way to take these kids and to move them forward based on where they are to the best of your ability.

While all three raters had the same instructional expectations of their teachers, including those in classrooms of students with disabilities, they acknowledged that instructional practice very often looks different in a special educator's classroom. The administrators expected to see smaller groups organized around the students' instructional levels, varying activities depending on the needs of the students (e.g., enrichment activities for advanced learners), and alternate assignments or accommodations to ensure that students could be successful. For example, if students could not yet read and interpret a provided text, the administrators expected the teachers to provide a less complicated text about the same subject. Thus, while the administrators expected special educators to have good classroom management, provide complete lessons, articulate goals and objectives, and engage students, they also expected the actual lessons and activities to vary depending on the individual needs of each student with a disability.

The administrators were also asked specifically about the teaching strategies they saw in the videos of special educators' lessons during the rating session in this study. Raters 1 and 2 identified one teacher in particular whose lessons appeared to be exemplary. Each of the raters independently gave this teacher's lessons the highest scores of all the lessons they viewed. The raters stated that the teacher provided clear instructions to the students, was articulate in her delivery, incorporated activities that were appropriate for the students' levels, had a brisk pace, and did not waste instructional time. The raters also stated that the students appeared to be enjoying the activities and were engaged with the teacher throughout her lessons. The raters, then, could identify what they believed to be a strong special education teacher, and provide descriptions to justify the strength of the teacher's performance.

The three administrators also identified similar problems across lessons for teachers whom they identified as weaker. Some of these problems included not providing students with opportunities to apply specific skills, wasting class time, continuing with a lesson without redirecting problem behaviors or disengaged students, cursory explanations of the lesson objective and/or instructional activities, and failing to close a lesson and connect it to prior or future learning. As previously mentioned, both the identified strengths and weaknesses of the special educators' instructional practices in this study did not differ from those that would potentially be associated with more or less efficacious general education teachers.

Theme 4: striking the balance between broad application and instrument precision. When conducting formal classroom observations in their previous districts, the three administrators commented that they were provided with observation protocols intended to be broad in their application. The same observation protocol was to be used across grade levels, content areas, and instructional settings. Because the instruments and forms were designed to be

universally applied, the administrators felt that they needed to make modifications when applying them to their observations of special educators' instructional practices. Rater 1 stated the following:

> In the special ed classroom, I would use parts of the evaluation that were appropriate for that teacher. Since there wasn't as much group teaching, I would give the teacher feedback based on the individual student productivity.

Rater 1, then, made adjustments to the evaluation tool based on the instructional setting (i.e., whole class or small group). Rater 2 remarked that she modified the instrument in the same way she would modify it for different grade levels. Rater 3 believed that the instruments should have included different categories to reflect the differences in special education classrooms, and that, in his experience, observation protocols failed to do this.

The rubric items in this study were designed to be more narrowly applied in the sense that they are intended for use only with special educators, but they are still broad enough to be used across different grade levels and content areas. The raters in this study believed that the rubric items were still not narrow and specific enough in their descriptions and scoring criteria. The raters wanted a clear definition of the item (e.g., *scaffolding*), and concrete examples of what this should look like at each score level. Although the raters were provided with lengthy descriptions of the items during the training, they wanted those specific descriptions to be written across the top of the rubric items themselves so that they were visible as a reminder during each rating. In other words, the administrators wanted to know clearly and precisely what they should be looking for in order to provide a score for each lesson.

Rater 3 mentioned, however, that the advantage of broader observation tools is that they leave some room for interpretation. Given broader categories, raters can use their own

knowledge and expertise to note nuances in a teacher's instructional practices. Rater 3 suggested that any rubric items used include sections for commentary so that administrators could add relevant information that may not be captured within the indicators for a specific rubric item. Thus, while all three rates wanted increased specificity so that they could score appropriately, they also wanted some freedom to expand beyond the rubric items if they were to be used in evaluations in the actual school setting.

The three raters noted additionally that the rubric items in this study were difficult to apply when teachers were working with individual students or in small-group settings. One rater mentioned, for example, that a rubric item included the statement, "All students were provided with opportunities for practice." His question was whether this statement applied only to the small group of students with which the teacher was working, or if this included all students in that teacher's classroom even if there were additional students present who were working independently or with an instructional assistant. Some of the teachers in the videos worked directly with individual students on very specific skills. The instruction, then, included a great deal of repeated practice, but was not part of a complete lesson. The administrators had a difficult time applying the rubric items to this type of instruction, which begs a much larger question regarding how best to evaluate special educators who provide one-on-one instruction focusing specifically on skill development.

Special education teacher perspectives on the evaluation process. Five special education teachers were interviewed regarding their general perceptions of the formal teacher evaluation process including specific measures used in their respective school districts. The teachers were also asked their opinions as to whether or not the individual rubric items included in this study accurately represented what they believed to be essential elements of effective

special education instruction. Prior to discussing the themes that emerged from their interviews, a description of the evaluation process these teachers have experienced is warranted.

Four of the five teacher participants had nearly identical descriptions of their districts' teacher evaluation process. They were each formally observed twice per year by the school principal or assistant principal. Prior to these observations, each teacher met with the administrator who would be conducting the evaluation, and together they reviewed goals for the year. The teacher could select these goals, and then they were discussed and agreed upon during the course of the meeting. The formal observations were scheduled in advance so that the teacher had prior notice of the administrator's classroom visits, and the teacher was instructed to submit lesson plans for the scheduled observations. Subsequent to each formal observation, the teacher met with the evaluator and the observation notes were reviewed. Although it was understood by the teacher that the administrator could visit classrooms for informal observations, this rarely occurred for participants in this study. According to these four participants, the two formal classroom observations were the majority, if not all, of their evaluations. The same measures and processes were applied to general and special educators; these participants did not report any modifications made for special education teachers.

For the fifth teacher participant, the classroom observation component of the evaluation process was described in the same manner as the other participants, but the formal observations were given less weight in the overall teacher evaluation. This teacher is employed by a non-public school, which is a private entity and, therefore, not subject to the same tenure and retention policies of public school systems. According to the fifth participant, he is considered to be an at-will employee and cannot obtain tenure, and while his evaluations can result in dismissal, they can also result in pay increases. His end-of-year evaluation includes components

that reflect his performance as an employee (e.g., being on time to work, working effectively with his colleagues), in addition to his classroom instruction. For the purposes of this study, the responses included in the thematic analysis reflect only the classroom observation component of his teacher evaluation.

After systematic analysis of the special education teacher interviews, four themes emerged and are discussed below.

Theme 1: Administrators should do more than go through the motions. The participants were asked questions regarding the background (i.e., education and experience) of the administrators who evaluated them, and the type and quality of feedback they received during the course of the evaluation. All five of the teacher participants reported being evaluated by the school principal or assistant principal, or the school director (in the case of the teacher employed at a non-public school). Three of the participants reported that the administrators responsible for their evaluations did not have any formal experience or possess degrees/credentials in the field of special education. Each of these three participants stated, however, that they did not feel this to be problematic as long as the administrators spent enough time inside special education classrooms to be aware of differences that may occur. One participant stated the following:

> I think I had individuals [as evaluators] who had at least an understanding of what took place in special education classrooms and how that environment was sometimes different from a general education environment.

These participants believed that exposure to special education classroom environments was sufficient, and a formal background in special education was not necessary for the administrator to produce a fair and unbiased evaluation.

One of the teacher participants did report that an administrator with a background in special education, school psychology specifically, conducted her evaluation. This individual, while arguably more knowledgeable about special education generally than other administrators lacking in the same training and experience, had never been a classroom teacher. The participant felt, then, that this administrator was not a good evaluator because she did not provide meaningful feedback on classroom instruction. Perhaps even more detrimental to the teacher's evaluation, the administrator provided feedback that was incorrect in terms of instructional strategies:

> When they're not familiar with the standards…like one of them she put satisfactory, but she said that these students should not be spelling phonetically when that's a kindergarten standard. She said, 'When I went to school we learned whole word.' To me, why are you going to put that as a comment if it doesn't match the curriculum? So, I've had experiences like that.

For this teacher, the mere fact that the administrator had an education in and knowledge of special education was inconsequential if the administrator was not also an experienced teacher.

The teacher participant working in a non-public school had previously been evaluated by a school administrator without a background in special education, but was also currently being evaluated by the school's director, who had teaching experience and a doctoral degree in special education. Having been exposed to both, this teacher participant felt that having a background in special education made a significant difference in the type and quality of feedback he received. He far preferred his current

evaluator, and felt that this individual was extremely helpful in providing feedback on his instruction as well as general direction for long-term professional goals.

In terms of being observed and receiving feedback, four out of five of the teachers stated that their administrators provided them with meaningful feedback on their instruction. The only teacher who did not receive meaningful feedback was the one evaluated by an administrator without teaching experience, as previously mentioned. For the five teachers in this study, whether or not the administrator had experience in special education mattered less in the grand scheme of the evaluation process than the overall quality of the administrator, defined by these teachers as having general experience and investment in the process, which are discussed below.

One teacher participant mentioned that although he had an administrator without a special education background, she was an experienced administrator who was still able to guide his teaching. He stated the following:

> I know that when I started, the feedback I received from an administrator who had been an administrator for a very long time was very beneficial. She really helped to guide me in terms of becoming a better teacher.

The participants felt that an experienced administrator with a background in teaching could provide necessary and helpful feedback, even if not specific to special education instructional strategies. The participants all stated, however, that what they wanted ultimately was an administrator who was invested in their teacher evaluations as a process for helping them grow as teachers and professionals. They did not want to feel like the administrator was merely "going through the motions" during the course of the

evaluation. They wanted to improve their practice and they needed ongoing support to do so.

The notion that the evaluation process should be one of ongoing mentorship and support emerged from the juxtaposition of two teachers' responses that were diametrically opposed in terms of the quality of feedback they received. One participant, who felt he greatly benefited from the evaluation process, stated that his administrator was fully invested in making the evaluations meaningful:

> It's helpful feedback. We're not just going through the motions to get it done. He's invested in helping us become better teachers. [He] has an open-door policy. If his door's open and you have anything you want to discuss…essentially, it's like ongoing mentorship.

Another teacher participant expressed great frustration with the evaluation process:

> It was just a very check here, you got satisfied here, ok sign at the bottom. And I would have to ask if there was anything I could improve on or anything they wanted to observe the next time they came in, and it was more of like, 'No, we're done for the year.'

She also stated that she wanted more of a plan for long-term growth:

> Overall, they've been positive in the way that, like the feedback is positive, but I don't feel that it supports my growth at all because there's not really a plan to grow after that. There are no goals. It's just kind of like read here, this is what I saw, sign here, we're done, and that's the end of it.

Essentially, the teacher participants in this study stated that the quality of the feedback they received depended in part on the administrator's background (i.e., general

experience, experience in teaching, and experience in special education), but more so on the investment of the administrator. They wanted the administrator to act as coach or mentor and to do more than go through the motions of the evaluation.

Theme 2: The jury is still out on fairness. Each of the participants was asked his/her opinions as to the fairness of the evaluation process as it applies to special education teachers. Three of the participants believed the process to be fair while two did not, but their reasons for taking a particular side varied remarkably. One participant felt the process was fair because special education students who are ultimately diploma bound are held to the same academic standards as general education students. It follows, then, that special education teachers should be held to the same standards of instructional planning and delivery as general education teachers. Another participant expressed a similar sentiment in that "good instruction is good instruction" regardless of the type of teacher and/or classroom environment. The third participant who espoused a belief in the fairness of the process expressed confidence in the accuracy of his evaluations in terms of reflecting the strengths and weakness of his teaching. He did not take issue with what was expected of him as a classroom teacher, and he believed the feedback from his administrator to be both insightful and helpful.

Although one participant believed in the fairness of the process because of academic expectations of students with disabilities, another participant believed the process was unfair for exactly the same reason. She stated that the expectations of the Common Core State Standards put an undue burden on both the students with disabilities and their teachers to demonstrate performance at a level which may be inappropriate given the current levels of the students:

> Right now, the district doesn't have a separate evaluation process for special ed teachers and most of the things they look at now with the framework are all

> Common Core based. So if they go in wanting to see higher-level thinking questions when my students only communicate through one word or pictures, I don't think it's reasonable at all. And the administrators can't identify the different accommodations, so like right now with some students when I give them two choices to respond, the administrators don't feel that it is gearing them toward the higher-level thinking. They feel I should ask them to explain their reasoning when the little girl really can't explain.

This teacher believed that her administrators were not always aware of the accommodating and scaffolding that take place in order to support students with disabilities as they access the curriculum. Holding special education teachers to the same instructional standards as general education teachers may mean that administrators are not crediting special education teachers for using various instructional supports.

Another participant also opined that the evaluation process was unfair, but not because of the classroom observation component. He believed strongly that special education teachers do and are responsible for so much more than classroom instruction:

> I feel like most of the evaluation standards focus on classroom planning, lesson planning, and delivery, and they're not focused on the case management responsibilities of a special education teacher. The standards don't reflect how effective we are at bringing out improvement in terms of student progress and progress toward goals and moving toward inclusion. None of the teaching standards focus on that, they focus on planning instruction and instructional delivery. I don't know how they do it everywhere else, but in terms of the

standards listed on our evaluation materials, they don't capture what special education teachers do.

This participant believed that good classroom instruction was important, but he was also responsible for organizing and holding IEP meetings, communicating regularly with parents and other stakeholders, ensuring that students on his caseload were receiving appropriate services, and performed additional responsibilities not held by general education teachers. He wanted his evaluation to reflect all parts of his job so that he was acknowledged for his growth and improvement in areas outside of classroom instruction.

There was no general consensus, then, among the participants as to whether or not the evaluation process they experienced was fair for special education teachers, and no common rationale for either side. The responses indicate that special education teacher evaluation is a multi-faceted, complex issue, and the fairness of the process is still up for debate.

Theme 3: Good instruction is good instruction…but it looks different. The participants were asked whether their instructional strategies differ from those used by general education teachers, which relates to the overall issue of whether or not observation protocols equitably represent the expected classroom practices of special educators. Each of the five participants stated in similar terms that good instruction is good instruction, but it will take on different forms in special education classrooms. The participants had a difficult time defining precisely what the differences would be; they simply noted that there are differences.

The overall sentiment was so aptly expressed by one participant:

I feel like good teaching is good teaching, and whether you're doing it in kindergarten or in high school—not that it looks the same—but a lot of good teaching incorporates the same strategies and techniques.

This participant likened the differences between special education classrooms and general education classrooms to the differences between grade levels. One would not expect a kindergarten teacher's instructional practices to look identical to a 10th-grade teacher's instructional practices; adjustments are obviously made to reflect students' developmental levels and the demands of the curriculum. In the same way, special education teachers alter their instructional strategies so as to appropriately accommodate the needs of their students.

The instructional strategies and techniques that should be incorporated in any classroom, as stated by the participants, are those included in the rubric items used for this study. The participants agreed that each of the seven rubric items (i.e., *articulation of lesson objective*, *teacher communication*, *sequencing*, *scaffolding*, *skill development*, *student engagement*, and *practice and review*) were important and should be part of a special education teacher's classroom instruction. The participants believed the rubric items to be part of good instruction generally, but some items were given more or less emphasis by the participants. For example, the participants believed that *sequencing*, *scaffolding*, and *practice and review* to be essential and should be emphasized when teaching students with disabilities. The participants felt that *teacher communication* (i.e., clearly communicating instructional purpose, procedures, and directions) was less of a priority. Overall, these participants believed that their instruction should be direct, be focused on specific skills, and provide plenty of opportunities for student practice.

The participants noted that their instruction sometimes looks "different", but these differences were hard to define. Some general comments included that these teachers often teach smaller groups and do much less whole-group instruction; they re-teach and emphasize particular concepts, and focus on student mastery before moving on to another standard or part of the curriculum; and they are responsive to unique academic and behavioral challenges that may not

be present in general education classrooms. One participant, who recently took on an administrative role and was responsible for evaluating another special education teacher, made the following statement:

> The teacher does it really well, but he isn't standing up delivering a lesson. If I tried to use a rubric like I would with an English teacher, it wouldn't carry over. But he has everything planned out, he has really good classroom management, the kids are making progress toward IEP goals, they are making progress toward the standards that they have—it just doesn't look the same.

The idea of the classroom "not looking the same" was repeated several times across the participants' responses. The participants did not believe that they should be held to a different set of instructional standards than those of general education teachers, but they also wanted observation protocols/rubrics to reflect existing differences in their instructional form. It should be noted, however, that the participants were unable to provide a description of how these differences would be operationalized for the purpose of improving an evaluative measure.

Theme 4: Two classroom observations are insufficient. Irrespective of the observation measure used, a repeated assertion made by the participants was that two classroom observations were not enough for the purposes of the evaluation itself or for promoting the teachers' growth and development. The participants wanted more classroom visits, and they wanted those visits to be informal and unannounced. Some participants believed that pre-planned visits allowed the teachers too much time to prepare, and the instruction that followed was a somewhat artificial representation of what truly occurs on a day-to-day basis. One participant even mentioned the benefit of seeing a teacher under less than ideal circumstances:

> We all have bad days, but I feel like the repeated visits would allow the person to see if the teacher grew from that last experience and if there has been improvement from the first time they came in.

This participant expressed a sentiment shared by the other participants, which was that repeated, informal visits allow administrators to see true classroom performance—for better or worse—and track growth. Repeated visits also enable the evaluator to better target areas for improvement, which would support teachers in their ongoing growth. As stated previously, the participants wanted the evaluation process, which includes classroom observations, to be an investment in the teacher's overall development.

The participants were asked about specific experiences that have contributed to their growth and development as teachers. They shared that formal observations have had at least some impact, albeit to varying degrees depending on the teacher. Other responses included collaborating with other teachers, visiting other teachers' classrooms to observe their instructional practice, and mentoring from colleagues and/or university personnel. A common response was that being observed and observing others was an integral and valuable part of becoming a more effective special education teacher.

CHAPTER FIVE: DISCUSSION

This study's first broad aim was to investigate school administrators as evaluators of special educators' instructional practices. Using observation rubrics, school administrators without a formal background in special education were tasked with rating video-recorded instructional segments of special education teachers. As to the question of whether school administrators—representing those typically responsible for conducting formal evaluations of special educators—can be reliable raters of special educators' classroom performance, results from this study suggest that the answer is yes, potentially.

One pair of raters in this study demonstrated fairly strong reliability when scores were averaged across teachers and lessons, although there was less agreement in the relative standing of lessons. The second pair of raters had much lower agreement in their scores, and this was largely the result of one rater applying the scoring criteria more stringently than the other. An examination of the score distributions across raters indicates that Rater 3 did not assign the highest score (i.e., 3) to any teacher's lesson on any of the rubric items. This type of variability in rater scores represents the inherent subjectivity of the evaluation process when individuals, who each bring a different set of knowledge and experiences, are asked to observe and rate teachers in the classroom. Some school administrators may be more lenient in their evaluations, while others may adhere to higher standards and expectations of their teachers. Rater 3, in this case, was a stricter evaluator than his counterparts.

These results suggest that school administrators demonstrate promise in their ability to reliably rate special education teachers' instructional practices, but the results also indicate that agreement in ratings ultimately depends on the strength of training on the instrument and the ability to calibrate raters both before rating sessions begin and over the course of the

observations. This requires time and considerable discussion so that administrators can reach a place of understanding in terms of what they should expect to see in a special education teacher's classroom. As evidenced from the G study, more training would be necessary to decrease the variability in scores across lessons for Raters 1 and 2, and more training would also be necessary to increase agreement in general between Raters 1 and 3. Results from the rater interviews support the need for increased training both during this study and generalizing to the actual school setting. The administrators in this study thoroughly enjoyed the training process of watching sample videos and engaging in rich discussion, but they wanted more training and increased specificity in the instrument so that they would know exactly what to expect and how to score. They also expressed a deficiency in the amount and quality of training they received when performing teacher evaluations at their actual school sites, and they stated that school administrators would benefit from the type of training provided in this study.

One of the more surprising findings was that, despite a lack of training and formal experience in special education, the administrators did not feel ill equipped or unqualified to perform evaluations of special education teachers. On the contrary, the administrators expressed confidence in their ability to conduct the evaluations and a willingness to do so when it was their responsibility on their respective school sites. Furthermore, the special education teachers who were interviewed in this study did not seem particularly concerned with the fact that the administrators who evaluated them did not have backgrounds in special education. The teachers were more concerned with the administrators' general teaching experience and investment in the evaluation process.

The results from the rater and teacher interviews suggest that there is value to an administrator's general experience regardless of educational background. An administrator who

has had considerable experience within the classroom, and several years of observing teachers across school sites and campuses could arguably be a better evaluator of special education teaching quality than a less seasoned administrator with formal training in special education. One special education teacher in this study expressed that differences between special and general education classrooms are analogous to differences between grade levels and content areas. A school administrator at the high school level, for example, may have previous teaching experience in English, but must still evaluate teachers of other content areas. No administrator will be experienced in every grade level and every content area, but this does not prevent school administrators from performing evaluations of all teachers. Similarly, a lack of experience in special education may not necessarily prohibit a school administrator from providing a fair, high quality, and meaningful evaluation of a special education teacher.

On average, the administrator raters in this study were more lenient in their ratings of the video-recorded instructional segments than trained special education teacher raters. This finding is consistent with previous research that school administrators tend to skew their ratings toward the higher end (Little, 2009; Weisberg et al., 2009). This study did not test whether the administrators were more lenient with special education teachers than general education teachers, and future research is necessary to determine if there is potential bias in administrator evaluations of special education teachers' instructional practices. While modified standards may be necessary in developing equitable evaluation systems, it is critical that special education teachers not be held to lower standards than general education teachers. The special education teachers themselves expressed a desire to be held to high standards of academic rigor within their classrooms of SWDs, and they believed they should be expected to perform at the same level instructionally as any other teacher.

This study examined rater reliability through a G study analysis, and the subsequent D studies provided additional information as to the number of lessons and number of raters required to achieve acceptable levels of reliability. As stated in previous research (Weisberg et al., 2009) and corroborated by the teacher participants in this study, formal evaluations typically consist of one school administrator observing a teacher twice during the academic year. The D study analyses reveal that one rater is insufficient regardless of the number of lessons he or she observes. It is unlikely that reliable scores can be obtained with an evaluation system involving one administrator conducting only two classroom observations. The D studies also indicate that four raters observing four lessons, as discussed in previous studies (e.g., Ho & Kane, 2013) may in fact be excessive. Given the limited resources school districts often have to staff personnel who would be qualified to perform teacher evaluations, and the time required for each individual to conduct several classroom observations, four raters and four lessons could very well be beyond practical and feasible in the actual school setting. What this study does confirm is that stronger and more reliable evaluation systems involve multiple raters observing multiple lessons (Ho & Kane, 2013).

Findings from the quantitative analysis are supported by teacher statements that two classroom visits are simply not enough to truly capture the strengths and weaknesses of their classroom instruction. While the teachers did not comment on the number of raters they believed to be necessary, they did feel that their school administrators did not spend enough time observing their classrooms. The teachers wanted more observations, and they wanted them to be informal rather than formal. Results from this study suggest that school administrators have a better chance of producing reliable and valid evaluations of special education teachers if they observe their classrooms on a more frequent basis.

It should be noted that reliability coefficients produced from the D studies were compared against the value of .80, which is considered to be an acceptable level of reliability (Cardinet et al., 2010). This value, however, is an arbitrary cutoff, and may be flexible depending on the types of decisions that schools districts would like to make. Ho and Kane (2013), for example, examined various combinations of raters and lessons that produced a reliability score of .65 and above. As previously mentioned, *relative* decisions, which involve evaluating teacher performance relative to peers, could be made with slightly lower reliability coefficients than .80. High-stakes decisions such as retention and tenure warrant confidence in scores assigned to any teacher's classroom observation; therefore, school districts must carefully consider cutoff points for reliability.

Second to examining school administrators as raters of observed lessons, this study also investigated the use of specific rubric items intended to represent effective instructional practices for classrooms of SWDs. A preliminary G study revealed that the rubric items were not a significant source of error in the measurement process, and the subsequent G studies with each pair of raters resulted in relatively low variability in any component including the rubric items. This suggests that the rubric items performed quite well in the measurement process, and did not significantly contribute to error.

The interviews with the administrators and teachers provided more in-depth information regarding each rubric item. Neither the administrators nor the special education teachers contested the inclusion of any item, although some items were given more priority than others. Both the administrators and teachers believed that while the items did reflect the instructional practices that should be expected of special education teachers, they did not feel that the items were a clear departure from the instructional expectations of general education teachers. All

participants in this study clearly stated that when school districts use an observation instrument, the same one could be applied to both general and special education teachers. If one observation protocol is used for all teachers, the special education teachers in this study expressed that some modifications may need to be made to reflect the instructional settings (e.g., small group) and differentiated instructional supports (e.g., scaffolding) included in special education classrooms.

While school districts should be using observation instruments that are reliable and valid measures of effective teaching, the analysis of both rater and teacher interviews in this study revealed that the instrument matters less than the individual performing the evaluation. The school administrators expressed that observation instruments should include explicitly stated expectations and clear scoring criteria, but they also wanted the freedom to add their own commentary and extend their evaluation beyond the limitations of the rubric items. The special education teachers were also less concerned with the specifications of any given instrument; they wanted their evaluators to spend time in their classrooms, observe their teaching on a more regular basis, and provide them with meaningful feedback based on the observations. The teachers wanted their administrators to be involved in the evaluation process on more than a surface level, and they wanted their evaluations to reflect more than just checked boxes or a series of scores. While observation instruments help define effective teaching and provide a common metric, more emphasis needs to be placed on training and supporting the individuals who use the instruments to perform evaluations and, ultimately, contribute to special educators' professional growth and development.

Limitations

Results from this study should be interpreted with caution given the small sample sizes. The raters in this study were all retired from their positions as school administrators, and their

average age is likely much higher than would be typically found across school sites. It is uncertain whether younger and less experienced administrators would have different perspectives on their experiences, especially considering they may have been exposed to different types of training given the changing structures of school districts' evaluation processes. Additionally, the administrators only had experience evaluating teachers within schools in Los Angeles County. The special education teachers, while representing a range of years of experience and type of instructional setting, were drawn from only two states, one of which consisted of a relatively homogenous (i.e., Caucasian and female) population of special education teachers. Additionally, the sample of special education teachers was small and included only those teaching students with mild to moderate disabilities. Broadening both the rater and teacher samples would allow for greater external validity of the study's findings.

Another limitation of this study was the confounding factors in the comparison between average scores of administrator raters and special education peer teacher raters. The school administrators were all located in California and were not trained on the four rubric items used for the comparative analysis. The special education teacher raters were all located in Idaho and had received training on the four rubric items. Differences in locale and level of training may have impacted the results.

An additional limitation of the study was the length of training provided to the school administrator raters. The raters received a full day of training, but many commercially available observation instruments require between 17 and 25 hours of training (Kane & Staiger, 2012). Additional time allows the raters to become familiar with each rubric item and practice scoring. The raters in this study noted that they were more comfortable with the rating materials and process after independently scoring approximately 10 video-recorded instructional segments. In

the future, training should involve a longer amount of time practicing individual scoring of the lessons, and ongoing calibration would be helpful to prevent rater drift throughout the rating process.

Caution should also be taken when interpreting the results of this study for the purpose of informing evaluations of teaching quality in a live classroom. The construct under investigation in this study was effective special education teaching, but teacher performance was captured via video recording technology, and the raters scored the video-recorded instructional segments. Drawing conclusions regarding the validity of the construct, as well as the validity of the observation protocol as a measure of effective teaching, requires extending inferences about teacher performance as seen on video to that teacher's performance as observed in person in the actual classroom setting. Furthermore, the teachers in this study self-selected the lessons to record, and could stop and start the video equipment at will. The raters could also start and stop the video, rewind and fast forward segments to watch again, and were not subject to the actual classroom wherein they may affect teacher and student behavior by their sheer presence. While video-recorded instructional segments may be helpful for training school administrators, future studies would benefit from testing observation instruments during live classroom instruction.

Future Directions

This study contributes to the growing body of work on measures of teaching quality. This study was the first to utilize school administrators without a formal background in special education as raters of special educators' instructional practices. Given the paucity of research on specific instruments used to measure special educator teaching quality and on the individuals who perform these evaluations, a great deal more research is needed. The G studies performed in this study need to be replicated with larger samples of both administrator raters and special

education teachers, and the scores of special education teachers' lessons need to be compared with those of general education teachers using the same raters and instruments.

While this study investigated an instrument designed specifically for evaluating special education teaching quality, the results from the study indicate that a specially designed instrument may not be necessary. Future studies could utilize popular commercially available instruments such as the FFT and CLASS, which are research-validated, but have not been tested for use specifically with special educators. Using available instruments will allow for a better investigation of potential bias in ratings, and will also enable researchers to determine precise modifications that may need to be made for classrooms of SWDs. Rather than constructing a new instrument, it is possible that supplemental materials might be developed, or modifications to particular rubric items might be made, to equitably represent the instructional settings and practices of special educators.

Ultimately, providing a fair, reliable, and meaningful evaluation of a special educator's teaching quality requires consensus from the field regarding effective instructional practices, which can translate to rubric items for observation purposes, and sufficient training for evaluators. Participants from this study had a difficult time defining exactly how special education instruction was different from general education, even though they were adamant that it looked different. Research on evidence-based instructional strategies that are effective for SWDs should drive the conversation regarding components of observation instruments used to measure special education teaching quality. Once the fine-grained distinctions are made between special education and general education classrooms, these distinctions can be better reflected in supplemental materials, especially with school districts that use one observation protocol for all teachers.

With appropriate expectations of instruction in place, school administrators can then be trained on how to apply their observation instruments to special education teachers' classrooms. It is clear from this study that more training is necessary, and it may be that this training needs to begin in personnel preparatory programs for prospective school administrators. Future research should explore the current training that is in place both in programs and school districts. School administrators need to be sufficiently trained on teacher evaluation in general, and regular calibration should occur at the district level so that all school administrators within a given school site are providing consistent evaluations of their teachers. School administrators should also be trained specifically to evaluate special education teachers (CEC, 2013).

The special education teachers who participated in this study universally expressed the sentiment that their teacher evaluations be meaningful and support them in their growth. This is consistent with literature on teacher evaluation in that the purpose of evaluating teachers should be more than a rank ordering of their performance relative to their peers (Benedict et al., 2013; Danielson, 2011); the evaluation process should be one that identifies and targets areas of growth, removes teachers who are persistently ineffective, and rewards teachers for their contribution to student learning. In order to improve special education teaching quality, any measures used to evaluate special educators should be used by individuals who are qualified, capable, and willing to support special educators in their current positions.

Table 1.

Demographic Data for Teacher Participants

Teacher Demographic	*Idaho*	*California*
Gender (m/f)	0/12	3/4
Ethnicity		
Caucasian or White	12	3
Asian or Asian American	0	1
Hispanic or Latino	0	3
African American or Black	0	0
Experience (years)		
Range	1 to 15	2 to 13
Mean	8.3	7
Instructional Context		
Level (elem./sec.)	9/3	3/4
Setting	resource (10), special day class (2)	pull-out (5), co-teaching (1), self-contained ED (1)
Content Area	Math, ELA	Math, ELA

Table 2.

Rubric item descriptions

Rubric Item	Description
Articulation of lesson objective	The teacher states the objective of the lesson; explains how the planned activities connect; and makes direct, meaningful connections to previous and/or future learning activities.
Teacher communication	The teacher clearly communicates the instructional purpose of the lesson, including where it is logically situated within broader learning. Procedures and directions are clearly explained, and the teacher does not make any errors when delivering content. The teacher's spoken and written language is expressive, and the teacher finds opportunities to extend students' vocabularies.
Sequencing	The instructional sequence is seamlessly and briskly paced. The teacher utilizes direct instruction features like modeling, highlighting, feedback, review, and opportunities for student practice in an organized and deliberate way. The teacher smoothly guides students from initial practice to generalized skill training (if applicable).
Scaffolding	The teacher has pre-determined the difficulties that may be encountered in a new task and provides appropriate support. Strategies to help students overcome the anticipated difficulties are provided. Activities are provided within a structured learning environment and provided intentionally to help move students to a new level of learning. Scaffolding is presented to provide a gradual transfer of control to the student for the learning activity.
Skill development	The teacher has planned for instruction that clearly accounts for developing, maintaining, and generalizing skills that students can apply in the classroom and across environments. The skill development is effectively integrated

	within the larger learning objective.
Student engagement	The teacher provides multiple opportunities for student participation, including differentiated activities intended to promote guided and independent student practice for all students. The teacher has created a learning environment that encourages active participation from all students, as well as maintains active levels of self-determination and self-advocacy.
Practice and review	The teacher provides consistent corrective feedback, frequent checks for understanding, and periodic reviews of instruction that integrate knowledge within a structured learning environment. All students are provided with opportunities for practice and receive individual attention when necessary. All student practice activities and exercises are designed so that new information/skills are clear and manageable for students.

Table 3.

Describing variance components in a {l:t} x r x i design

Source	Description
T	Teacher or "true score" variance (object of measurement).
L:T	Variance due to lessons. Confounded with the teacher-by-lesson interaction.
R	Variance due to raters.
I	Variance due to rubric items.
T x R	Some raters score some teachers higher than others.
T x I	Some teachers score higher on some rubric items than others.
R x (L:T)	Some raters score certain lessons higher than others. Confounded with the teacher-by-lesson-by-rater interaction.
I x (L:T)	Some rubric items receive higher scores on some lessons than others. Confounded with the teacher-by-lesson-by-item interaction.
R x I	Some raters score some rubric items higher than others.
T x R x I	Some raters score some teachers higher on certain rubric items.
R x I x (L:T), *e*	Highest order interaction effect, confounded with residual error variance.

Table 4.

Selected codes and descriptions

Code	Description
Training	Any comments related to previous teacher evaluation training participants had received/experienced. This could be district-provided training or training participants sought on their own. This could also include training they would have liked to receive or their idea of an ideal training structure. This code also includes any comments related to the training experience during this study.
Personal Bias	Statements suggesting that the participant has a tendency or preference toward a particular perspective of special educators, and that tendency may interfere with the participant's ability to be impartial when evaluating special educators (in the past, not during this study).
Feedback	Description of the process of providing feedback to special educators on their instructional practice during evaluations. This may include participants' opinions as to their own ability to offer meaningful/constructive feedback. This may also include the types of feedback they provided, areas they felt they could speak to in terms of instructional practice, and areas about which participants felt they could not speak.

Table 5.

Kappa scores (and standard errors) for rater/administrator transcripts

Transcript	Kappa (SE)
1	.71 (.05)
2	.76 (.07)
3	.89 (.06)

Table 6.

Kappa scores (and standard errors) for teacher transcripts

Transcript	Kappa (SE)
1	.77 (.08)
2	.83 (.06)
3	.81 (.06)
4	.84 (.11)
5	.85 (.09)

Table 7.

Selected examples of significant statements and formulated meanings from teacher interviews

Significant Statement	Formulated Meaning
No, [the assistant principal was not able to give me feedback on my teaching] because they've never taught before.	The ability of a school administrator to provide meaningful feedback on instruction depends on having had previous classroom teaching experience.
When they're not familiar with the standards, like one of them she put satisfactory, but she said these students should not be spelling phonetically when that's a kindergarten standard. She said, 'When I went to school we learned whole word.' To me, why are you going to put that as a comment if it doesn't match the curriculum?	If administrators are unfamiliar with the curriculum and standards, this will impact their ability to provide an accurate evaluation.
I think I at least had individuals who had at least an understanding of what took place in special education classrooms and how that environment was sometimes different from a general education environment, especially from those classes where you have nothing but high-achieving students.	At least some exposure to special education classroom environments is necessary for administrators to understand differences they may encounter.
I would have preferred that [the administrator] came three days, for two hours, different parts of the day, to really know my teaching style…The one specific we're going to come in at 8:30 so you have	Frequent and informal classroom visits are necessary to truly capture a teacher's strengths and weaknesses in instructional practice.

all your kids prepped; I've seen teachers that teach totally different that one time the administrator is there versus what goes on throughout the year.	
Overall, they've been positive in the way that, like the feedback is positive, but I don't feel that it supports my growth at all because there's not really a plan to grow after that. There are no goals. It's just kind of like read here, this is what I saw, sign here, we're done, and that's the end of it.	The evaluation process lacks meaning and ultimate benefit to the teacher when the administrator treats it as a formality.
It's helpful feedback. We're not just going through the motions to get it done. He's invested in helping us become better teachers.	The evaluation process takes on meaning and benefit when the teacher perceives that the administrator is invested in his or her growth.

Table 8.

Example of a theme cluster with associated formulated meanings

Administrators should do more than go through the motions
Most administrators do not have a formal background in special education, but exposure to special education classrooms can be sufficient.
An administrator with a special education background, but without teaching experience, is arguably a weak evaluator.
Experienced administrators can help guide teaching practices regardless of educational background.
Teachers want administrators to be invested in the evaluation process.
The evaluation process should be one of ongoing mentorship and support.
Educational background matters less than the administrator's commitment to the evaluation process as something more than a mere formality.

Table 9.

Percent of variance by source

Source	Raters 1 and 2	Raters 1 and 3	Average weighted estimates
T	31.3	6.4	18.9
L:T	4.6	4.5	4.6
R	0.2	20.9	10.5
I	1.9	1.7	1.8
T x R	8.8	1.1	4.9
T x I	6.1	3.4	4.8
R x (L:T)	21.1	31.6	26.2
I x (L:T)	1.1	3.3	2.2
R x I	3.5	4.6	4
T x R x I	4.4	2.1	3.1
R x I x (L:T), *e*	17.1	20.5	18.8

Table 10.

Decision study results for Raters 1 and 2 with rubric items as a fixed facet

Number of raters and lessons	Relative G Coefficient (SEM)	Absolute G Coefficient (SEM)
1 rater, 1 lesson	.48 (.57)	.47 (.57)
1 rater, 2 lessons	.59 (.45)	.59 (.45)
1 rater, 3 lessons	.64 (.41)	.64 (.41)
1 rater, 4 lessons	.67 (.38)	.67 (.38)
2 raters, 1 lesson	.62 (.43)	.61 (.43)
2 raters, 2 lessons	.72 (.34)	.72 (.34)
2 raters, 3 lessons	.77 (.30)	.77 (.30)
2 raters, 4 lessons	.79 (.28)	.79 (.28)
3 raters, 1 lesson	.68 (.37)	.68 (.37)
3 raters, 2 lessons	.78 (.29)	.78 (.29)
3 raters, 3 lessons	.82 (.25)	.82 (.26)
3 raters, 4 lessons	.84 (.24)	.84 (.24)
4 raters, 1 lesson	.72 (.34)	.72 (.34)
4 raters, 2 lessons	.81 (.23)	.85 (.23)
4 raters, 3 lessons	.85 (.24)	.84 (.24)
4 raters, 4 lessons	.87 (.21)	.87 (.21)

Table 11.

Decision study results for Raters 1 and 3 with rubric items as a fixed facet

Number of raters and lessons	Relative G Coefficient (SEM)	Absolute G Coefficient (SEM)
1 rater, 1 lesson	.15 (.56)	.10 (.70)
1 rater, 2 lessons	.25 (.40)	.14 (.58)
1 rater, 3 lessons	.33 (.33)	.16 (.54)
1 rater, 4 lessons	.39 (.29)	.17 (.51)
2 raters, 1 lesson	.24 (.42)	.17 (.51)
2 raters, 2 lessons	.38 (.30)	.23 (.42)
2 raters, 3 lessons	.47 (.25)	.27 (.39)
2 raters, 4 lessons	.53 (.22)	.29 (.37)
3 raters, 1 lesson	.29 (.36)	.22 (.43)
3 raters, 2 lessons	.45 (.26)	.30 (.35)
3 raters, 3 lessons	.54 (.21)	.34 (.32)
3 raters, 4 lessons	.61 (.19)	.37 (.31)
4 raters, 1 lesson	.34 (.33)	.26 (.39)
4 raters, 2 lessons	.50 (.23)	.35 (.31)
4 raters, 3 lessons	.59 (.19)	.40 (.29)
4 raters, 4 lessons	.66 (.17)	.43 (.27)

Figure 1. *Rater 1 Score Distribution*

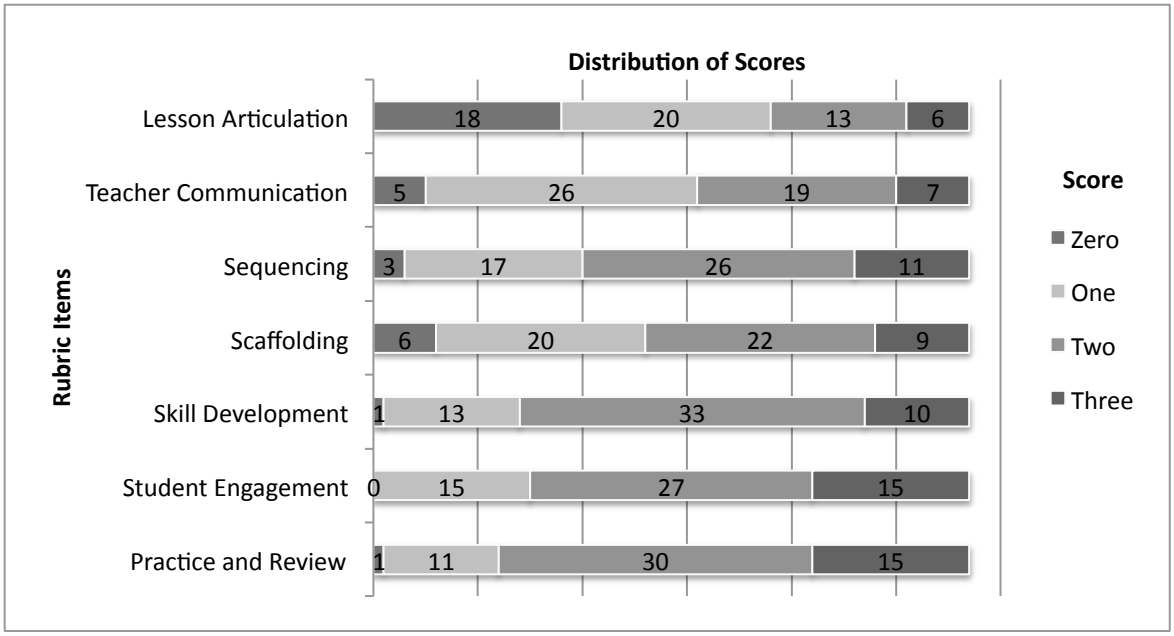


Figure 2. *Rater 2 Score Distribution*

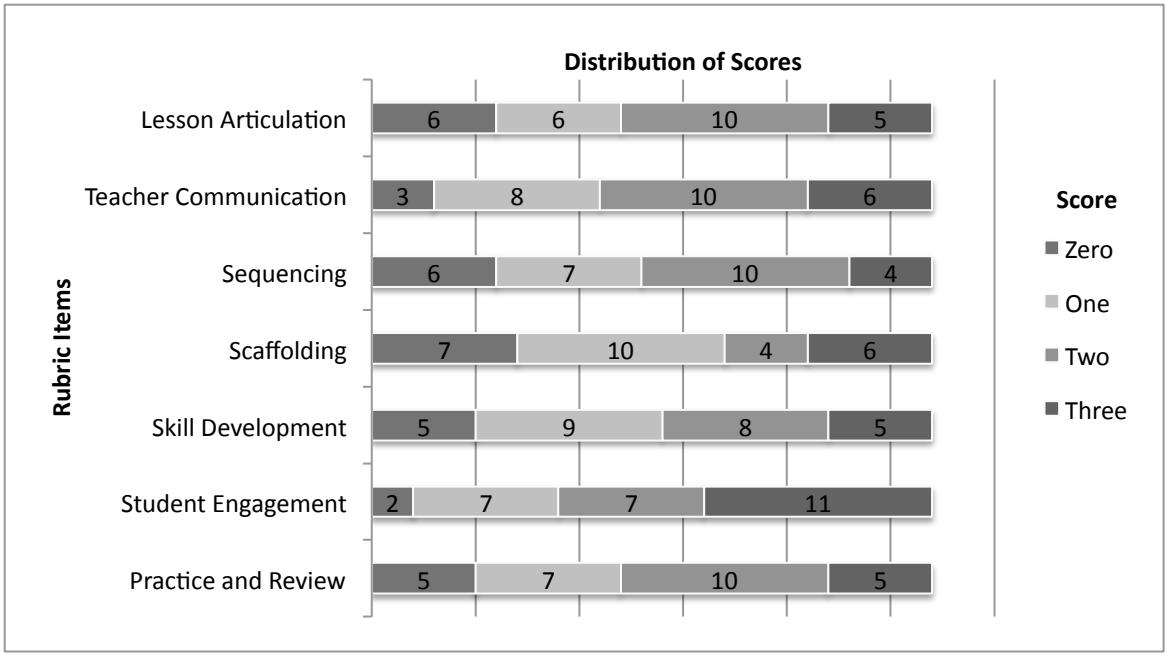


Figure 3. *Rater 3 Score Distribution*

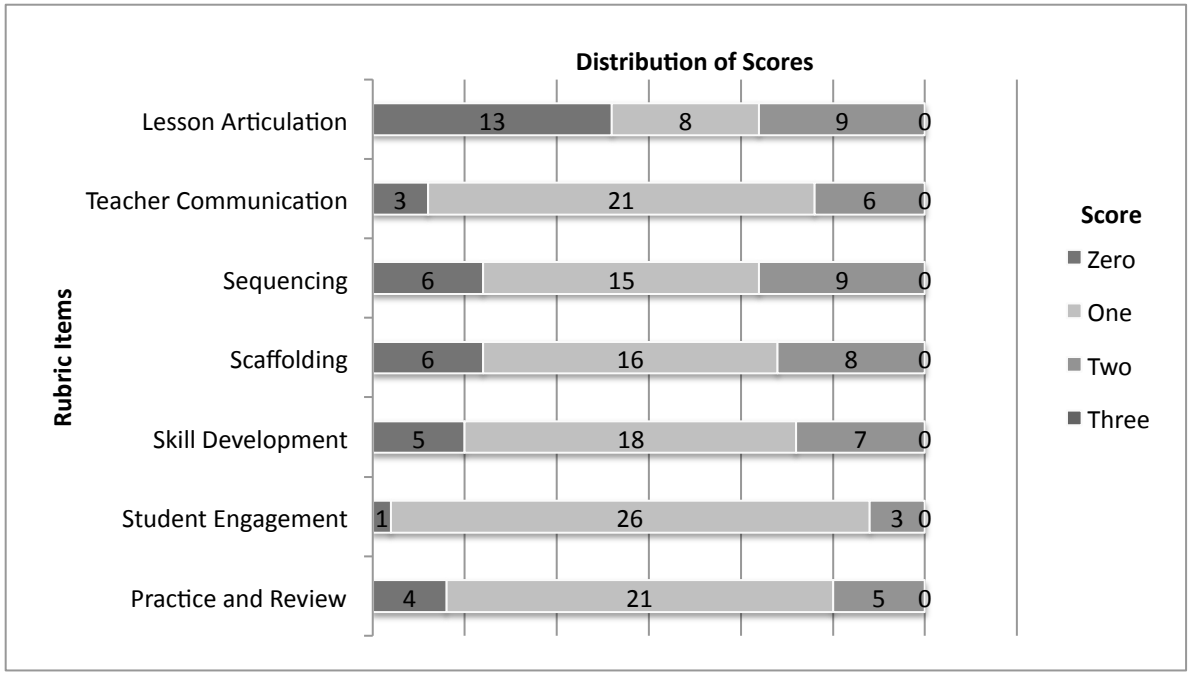
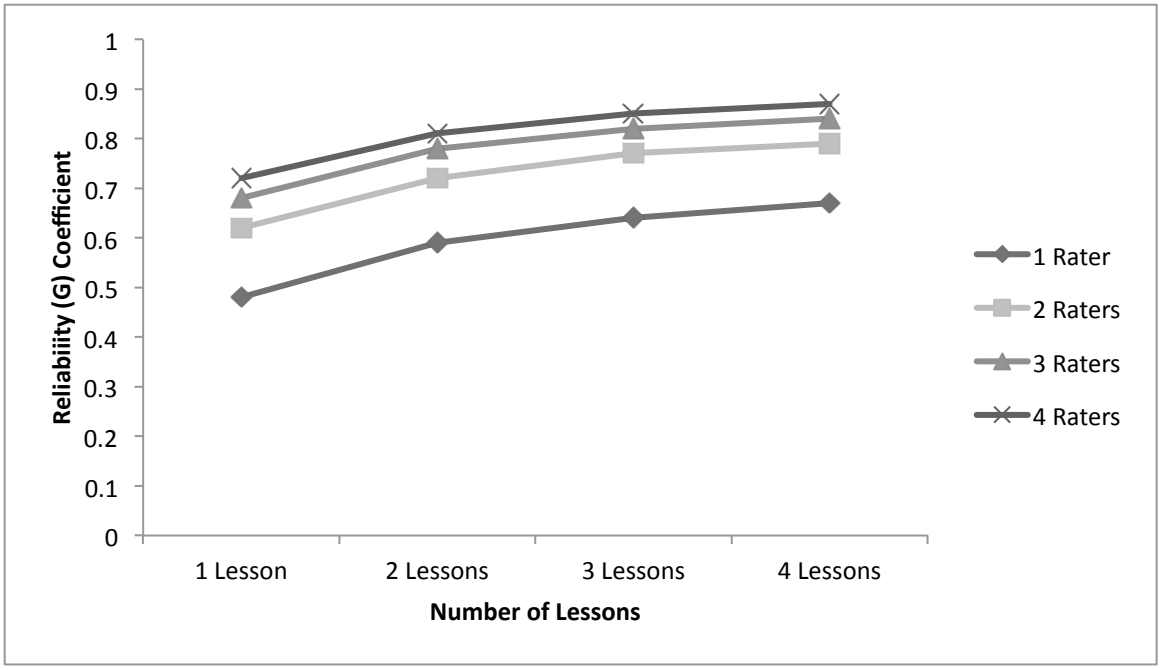
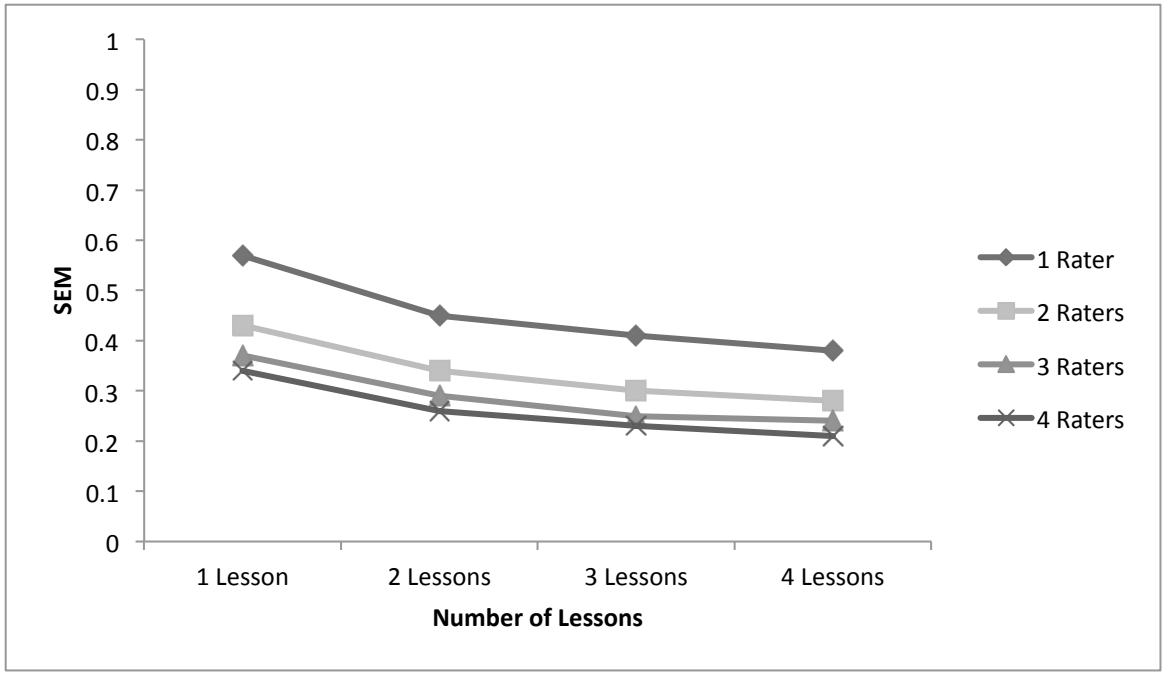


Figure 4. *D Study Reliability Coefficients for Raters 1 and 2*



Note: The figure includes *relative* reliability coefficients. *Absolute* reliability coefficients have been omitted due to the nearly identical values.

Figure 5. *D Study SEM for Raters 1 and 2*



Note: The figure includes SEM for both *relative* and *absolute* reliability coefficients due to identical values.

Figure 6. *Reliability Coefficients (relative) for Raters 1 and 3*

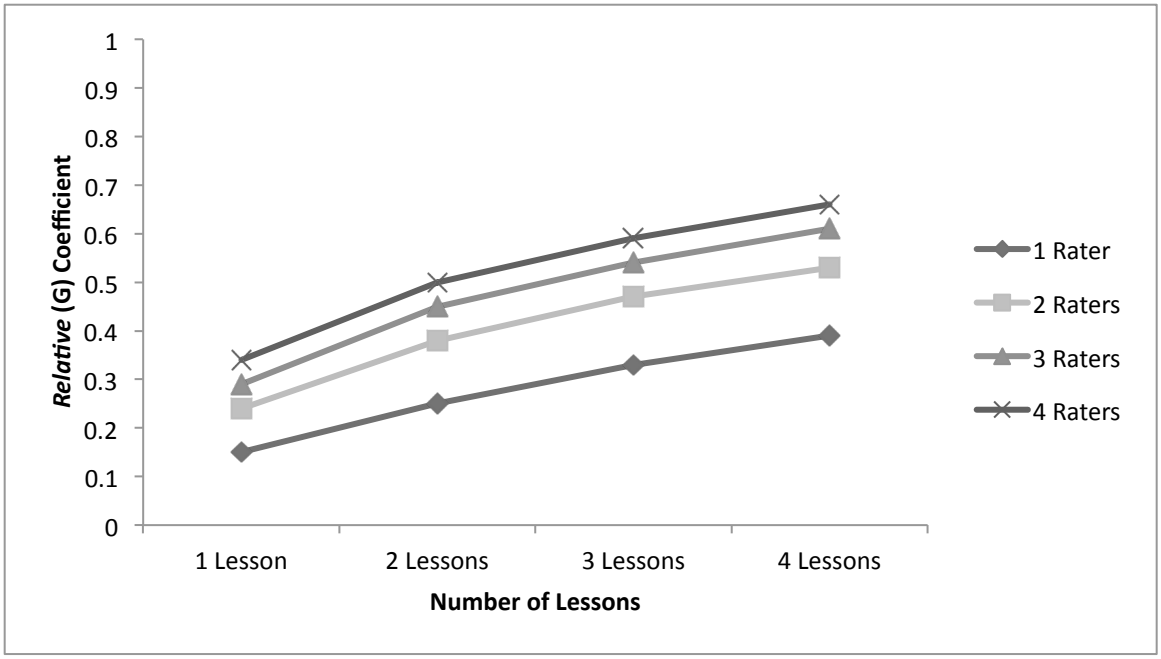


Figure 7. *SEM (relative) for Raters 1 and 3*

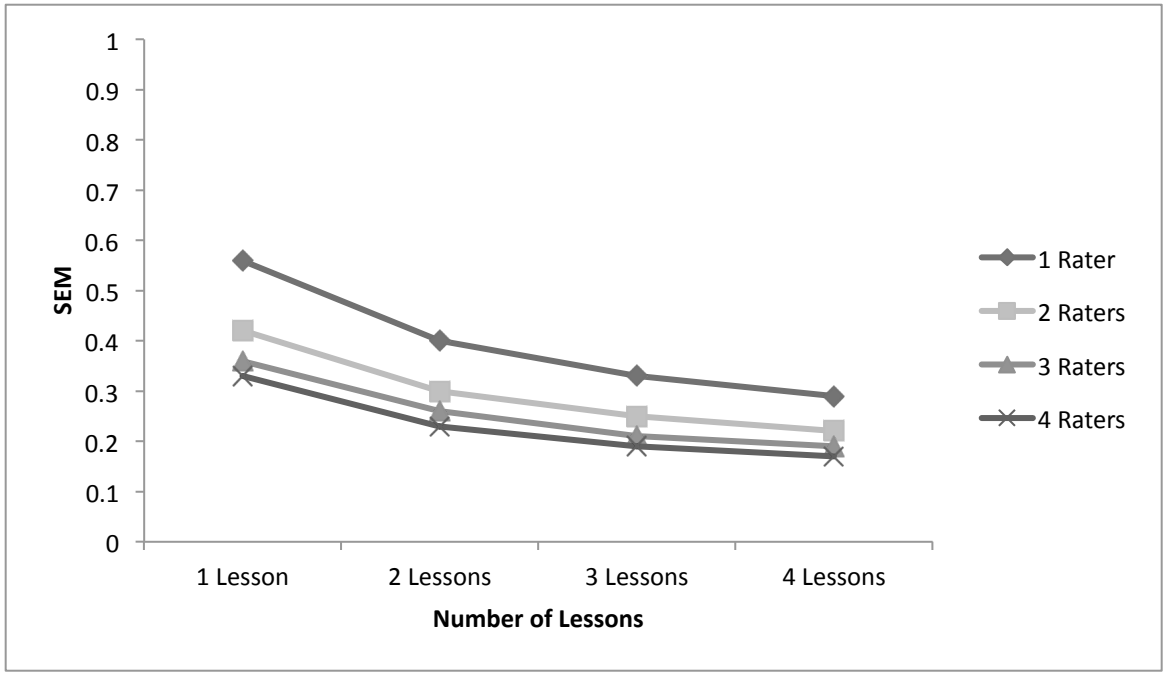


Figure 8. *Reliability Coefficients (absolute) for Raters 1 and 3*

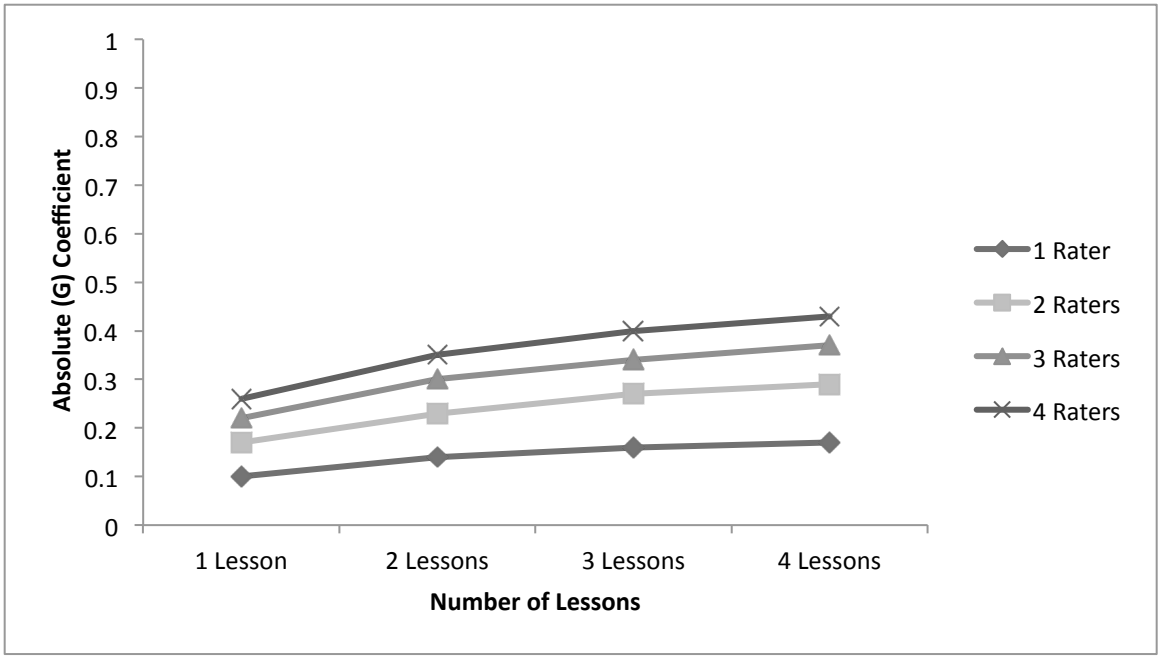


Figure 9. *SEM (absolute) for Raters 1 and 3*

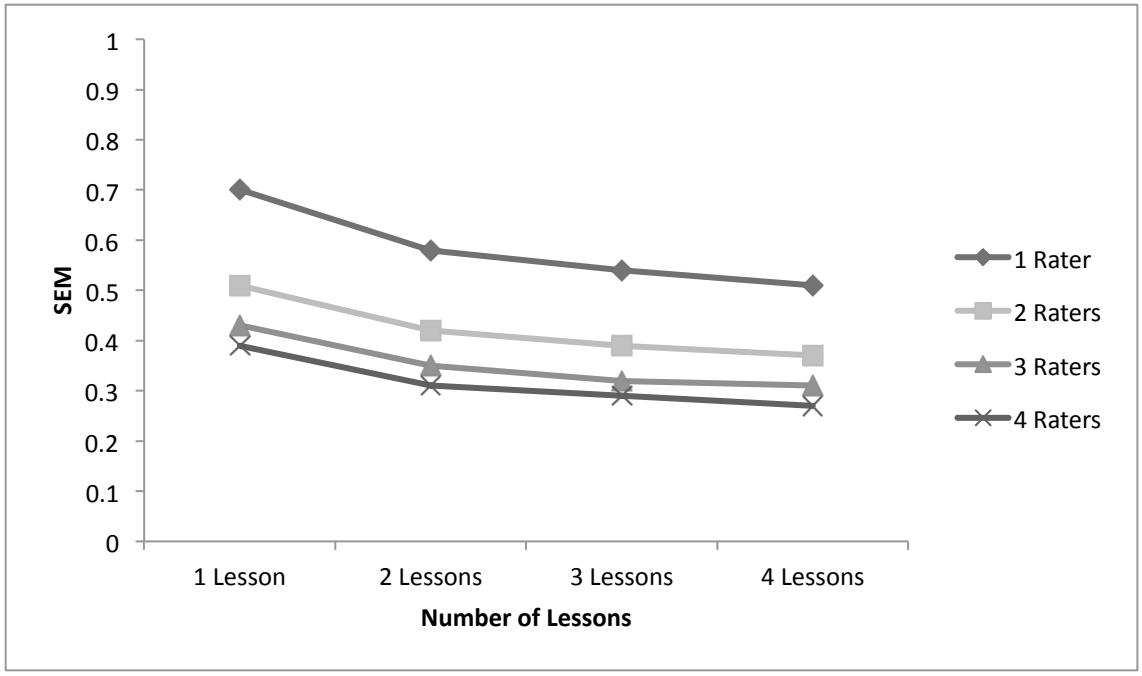
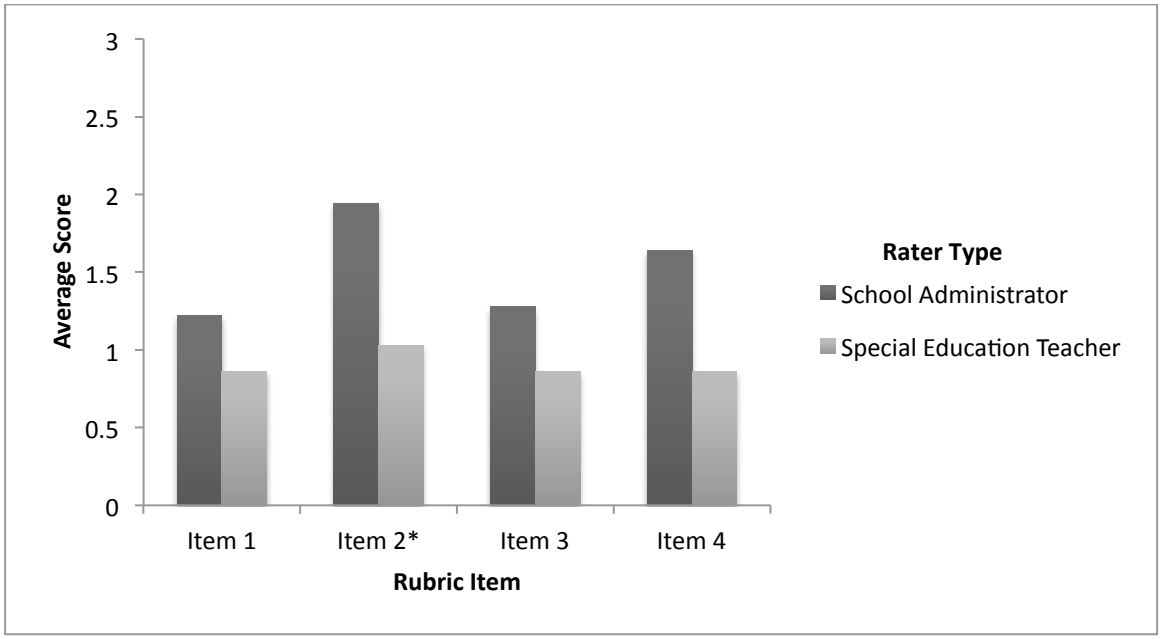


Figure 10. *A comparison of average scores by rater type on each of four rubric items.*



Note: * $p < .05$

References

Ahearn, E. (2009). *Growth models and students with disabilities: Report of state interviews*. Alexandria, VA: National Association of State Directors of Special Education. Retrieved from http://www.projectforum.org

Allen, J. P., Pianta, R. C., Gregory, A., Mikami, A. Y., & Lun, J. (2011). An interaction-based approach to enhancing secondary school instruction and student achievement. *Science 333*(6045), 1034–37. doi: 10.1126/science.1207998

Armor, D., Conry-Oseguera, P., Cox, M., King, N. J., McDonnell, L. M., Pascal, A. H., …Zellman, G. L. (1976). *Analysis of the school preferred reading program in selected Los Angeles minority schools*. Santa Monica, CA: Rand Corporation.

Armstrong, D., Gosling, A., Weinman, J., & Marteau, T. (1997). The place of inter-rater reliability in qualitative research: An empirical study. *Sociology*, *31*(3), 597-606. doi: 10.1177/0038038597031003015

Baker, E. L., Barton, P. E., Darling-Hammond, L., Haertel, E., Ladd, H. L., Linn, R. L., …Shepard, L. A. (2010). *Problems with the use of student test scores to evaluate teachers*. Washington, DC: The Economic Policy Institute.

Bell, C. A., Gitomer, D. H., McCaffrey, D. F., Hamre, B. K., Pianta, R. C., & Qi, Y. (2012). An argument approach to observation protocol validity. *Educational Assessment, 17*, 62-97. doi: 10.1080/10627197.2012.715014

Bell, C., Jones, N., Lewis, J., & Qi, Y. (2013, March). *Predicting observer training satisfaction and certification*. Paper presented at the spring meeting of Society for Research on Educational Effectiveness, Washington, DC.

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

Blanton, L. P., Sindelar, P. T., & Correa, V. I. (2006). Models and measures of beginning teaching quality. *The Journal of Special Education, 40*(2), 115-127. doi: 10.1177/00224669060400020201

Boyatzis, R. E. (1998). *Transforming qualitative information: Thematic analysis and code development.* Thousand Oaks, CA: Sage Publications, Inc.

Braun, H. I. (2005, September). *Using student progress to evaluate teachers. A primer on value-added models.* Washington, D.C.: ETS, Policy and Information Center. Retrieved June 6, 2013, from http://www.ets.org/Media/Research/pdf/PICVAM.pdf

Brennan, R. L. (2000). Performance assessments from the perspective of generalizability theory. *Applied Psychological Measurement, 24*(4), 339-353. doi: 10.1177/01466210022031796

Brownell, M., Bishop, A., Gersten, R., Klingner, J., Penfield, R., Dimino, J., …Sindelar, P. (2009). The role of domain expertise in beginning special education teacher quality. *Exceptional Children, 75*(4), 391-411.

Brownell, M., Haager, D., Bishop, A., Klingner, J. Menon, S., Penfield, R., & Dingle, M. (2007, April). *Teacher quality in special education: The role of knowledge, classroom practice, and school environment.* Paper presented at the Annual Meeting for the American Education Research Association, Chicago, IL.

Brownell, M., Smith, S., Crockett, J., & Griffin, C. (2012). *Inclusive instruction: Evidence-based practices for teaching students with disabilities*. New York, NY: Guilford Press.

Burns, M. K., & Ysseldyke, J. E. (2008). Reported prevalence of evidence-based instructional practices in special education. *The Journal of Special Education, 43*(1), 3-11. doi: 10/1177/0022466908315563

Buzick, H. M., & Laitusis, C. C. (2010). Using growth for accountability: Measurement challenges for students with disabilities and recommendations for research. *Educational Researcher, 39*(7), 537-544. doi: 10.3102/0013189X10383560

Cardinet, J., Johnson, S., & Pini, G. (2010). *Applying Generalizability Theory Using EduG*. New York, NY: Taylor & Francis Group.

Chiu, C. W., & Wolfe, E. W. (2002). A method for analyzing sparse data matrices in the generalizability theory framework. *Applied Psychological Measurement, 26*(3), 321-338. doi: 10.1177/0146621602026003006

Christensen, L. L., Lazarus, S. S., Crone, M., & Thurlow, M. L. (2008). 2007 *State policies on assessment participation and accommodations for students with disabilities* (Synthesis Report 69). Minneapolis: University of Minnesota, National Center on Educational Outcomes.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*(1), 37-46.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Earlbaum Associates.

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

Council for Exceptional Children. (2013). The Council for Exceptional Children's position on special education teacher evaluation. *Teaching Exceptional Children, 45*(3), 73-76. Retrieved from http://cec.metapress.com/content/022w828643484g12/

Creswell, J. W. (2007). *Qualitative inquiry and research design: Choosing among five approaches* (2nd ed.). Thousand Oaks, CA: Sage Publications, Inc.

Danielson, C. (2011). *The framework for teaching evaluation instrument.* Princeton, NJ: The Danielson Group.

Darling-Hammond, L. (2010). *Evaluating teacher effectiveness: How teacher performance assessments can measure and improve teaching*. Retrieved from http:www.american-progress.org/issues/education/report/2010/10/19/8502/evaluating-teacher-effectiveness/

Erlich, O., & Shavelson, R. J. (1976). *The application of generalizability theory to the study of teaching* (Beginning Teacher Evaluation Study). San Francisco, CA: Far West Laboratory.

Erlich, O., & Shavelson, R. J. (1978). The search for correlations between measures of teacher behavior and student achievement: Measurement problem, conceptualization problem, or both? *Journal of Educational Measurement, 15*, 77-89.

Everitt, B. (1996). *Making sense of statistics in psychology: A second-level course.* Oxford, UK: Oxford University Press.

Fuchs, L., & Fuchs, D. (2001). Helping teachers formulate sound test accommodation decisions for students with learning disabilities. *Learning Disabilities Research and Practice, 16*(3), 174-181. doi: 10.1111/0938-8982.00018

Goe, L., & Croft, A. (2009). *Methods of evaluating teacher effectiveness*. Retrieved from http://wwwgtlcenter.org/sites/defaults/files/docs/RestoPractice_EvaluatingTeacherEffectiveness.pdf

Goe, L., & Holdheide, L. (2011). *Measuring teachers' contributions to student learning growth for nontested grades and subjects*. Washington, DC. Retrieved from http://www.tqsource.org/publications/MeasuringTeachersContributions.pdf

Goldhaber, D. D., Goldschmidt, P. & Tseng, F. (2013). Teacher value-added at the high-school level: Different models, different answers? *Educational Evaluation and Policy Analysis, 35*(2), 220-236. doi: 10.3102/0162373712466938

Gordon, R., Kane, T., & Staiger, O. (2006). *Identifying effective teachers using performance on the job*. Washington DC: The Brookings Institution.

Grissom, J. A., Loeb, S., & Master, B. (2014). Effective instructional time use for school leaders: Longitudinal evidence from observations of principals. *Educational Researcher, 42*(8), 433-444. doi: 10.3102/0013189X13510020

Harris, D. N., & Sass, T. R. (2009). *What makes for a good teacher and who can tell?* (CALDER Working Paper No. 30). Washington DC: Urban Institute.

Harris, D. N., & Sass, T. R. (2011). Teacher training, teacher quality, and student achievement. *Journal of Public Economics, 95*(7-8), 798-812. doi: 10.1016/j.jpubeco.2010.11.009

Harris, D. N., Ingle, W. K., Rutledge, S. A. (2014). How teacher evaluation methods matter for accountability: A comparative analysis of teacher effectiveness ratings by principals and teacher value-added measures. *American Educational Research Journal, 51*(1), 73-112. Doi: 10.3102/0002831213517130

Hill, H. C. (2009). Evaluating value-added models: A validity argument approach. *Journal of Policy Analysis and Management, 28*(4), 700-709. doi: 10.1002/pam.20463

Hill, H. C., Charalambous, C. Y., & Kraft, M. A. (2012). When rater reliability is not enough: Teacher observation systems and a case for the generalizability study. *Educational Researcher, 41*, 56-64. doi: 10.3102/0013189X12437203

Ho, A. D., & Kane, T. J. (2013). *The reliability of classroom observations by school personnel*. Retrieved from http://www.metproject.org/downloads/MET_Reliability_of_Classroom_Observations_Research_Paper.pdf

Holdheide, L. R., Goe, L., Croft, A., Reschly, D. J. (2010). Challenges in evaluating special education teachers and English Language Learner specialists. [Research and Policy Brief]. National Comprehensive Center for Teacher Quality. Retrieved from http://www.isbe.net/peac/pdf/tchr_eval_res_sped.pdf

Holdheide, L. R., Browder, D., Warren, S., Buzick, H., & Jones, N. (2012). *Using student growth to evaluate educators of students with disabilities: Issues, challenges, and next steps.* Washington, D.C.: National Comprehensive Center for Teacher Quality. Retrieved June 6, 2013, from http://www.isbe.state.il.us/peac/pdf/using_student_growth_summary0112.pdf

Jacob, B. A., & Lefgren, L. (2008). Can principals identify effective teachers? Evidence on subjective performance evaluation in education. *Journal of Labor Economics, 26*(1), 101-136. doi: 10.1086/522974

Johnson, E., & Semmelroth, C. L. (2012). Examining interrater agreement analyses of a pilot special education observation tool. *The Journal of Special Education Apprenticeship, 1*(4). Retrieved from http://josea.info/index.php?page=vol1no2

Johnson, E., & Semmelroth, C. L. (2013). 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. doi: 10.1177/1534508413513315

Jones, N. D., & Brownell, M. T. (2013). Examining the use of classroom observations in the evaluation of special education teachers. *Assessment for Effective Intervention, 39*(2), 112-124. doi: 10.1177/1534508413514103

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. doi: 10.3102/0013189X12468211

Kane, T. J., & Cantrell, S. (2013). Ensuring fair and reliable measure of effective teaching: Culminating findings from the MET project's three-year study. Retrieved from http://www.metproject.org/downloads/MET_Ensuring_Fair_and_Reliable_Measures_Practitioner_Brief.pdf

Kane, T. J., & Staiger, D. O. (2008). *Estimating teacher impacts on student achievement: An experimental evaluation* [Working Paper Series 14607]. Cambridge, MA: National Bureau of Economic Research.

Kane, T. J., & Staiger, D. O. (2012). *Gathering feedback for teaching: Combining high-quality observations with student surveys and achievement gains*. Retrieved from http://www.metproject.org/downloads/MET_Gathering_Feedback_Research_Paper.pdf

Koedel, C., & Betts, J. (2007). *Re-examining the role of teacher quality in the educational production function* [Working Paper 2007-03]. Nashville, TN: Vanderbilt Peabody College, National Center on Performance Incentives.

Koretz, D. (2002). Limitations in the use of achievement tests as measures of educators' productivity. *Journal of Human Resources, 37*(4), 752-777.

Koretz, D. M., & Hamilton, L. S. (2006). Testing for accountability in K-12. In R. L. Brennan (Ed.), *Educational Measurement* (4th ed., pp. 531-578). Westport, CT: American Council on Education and Praeger.

Little, O. (2009). *Teacher evaluation systems: The window for opportunity and reform*. Washington DC: National Education Association.

Liu, E., & Johnson, S. M. (2006). New teachers' experiences of hiring: Late, rushed and information-poor. *Educational Administration Quarterly, 42*(3), 324-360. doi: 10.1177/001361X05282610

Lockwood, J. R., McCaffrey, D. F., Hamilton, L. S., Stecher, B., Le, V., & Martinez, J. F. (2007). The sensitivity of value-added teacher effect estimates to different mathematics achievement measures. *Journal of Educational Measurement, 44*(1), 47-67. doi: 10.111/j.1745-3984.2007.00026.x

McCaffrey, D. F. (2012). *Do value-added methods level the playing field for teachers?* Carnegie Knowledge Network. Retrieved June 6, 2013, from http://carnegieknowledenetwork.org/briefs/value-added/level-playing-field/

McCaffrey, D. F., Lockwood, J. R., Koretz, D., Louis, T. A., & Hamilton, L. (2004). Models for value-added modeling of teacher effects. *Journal of Educational and Behavioral Statistics, 29*(1), 67-101. doi: 10.3102/10769986029001067

McCaffrey, D. F., Sass, T. R., Lockwood, J. R., & Mihaly, K. (2009). The intertemporal variability of teacher effect estimates. *Education Finance and Policy, 4*(4), 572-606. doi: 10/1162/edfp.2009.4.4.572

Mertens, D. M. (2010). *Research and evaluation in education and psychology: Integrating diversity with quantitative, qualitative, and mixed methods* (3rd ed.). Thousand Oaks, CA: Sage Publications, Inc.

Meyer, J. P., Liu, X., & Mashburn, A. J. (2014). A practical solution to optimizing the reliability of teaching observation measures under budget constraints. *Educational and Psychological Measurement, 74*(2), 280-291. doi: 10.1177/0013164413508774

No Child Left Behind (NCLB) Act of 2001, Pub. L. No. 107-110, § 115, Stat. 1425 (2002).

Noell, G. H., Brownell, M. T., Buzick, H. M., & Jones, N. D. (2014). *Using educator effectiveness measures to improve educator preparation programs and student outcomes* (Document No. LS-1). Retrieved from University of Florida, Collaboration for Effective Educator, Development, Accountability, and Reform Center website: http://ceedar.education.ufl.edu/tools/literature-syntheses/

Papay, J. P. (2011). Different tests, different answers: The stability of teacher value-added estimates across outcome measures. *American Educational Research Journal, 48*(1), 163-193. doi: 10/3102/0002831210362589

Pianta, R. C., La Paro, K. M., & Hamre, B. K. (2008). *Classroom Assessment Scoring System: Manual K-3.* Baltimore, MD: Brookes.

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

Rivkin, S. G., Hanushek, E. A., & Kain, J. F. (2005). Teachers, schools and academic achievement. *Econometrica, 73*(2), 417-458. doi: 10/1111/j.1468-0262.2005.00584.x

Sanders, W. L., & Rivers, J. C. (1996). *Cumulative and residual effects of teachers on future student achievement* (Research Progress Report). Knoxville: University of Tennessee Value-Added Research and Assessment Center.

Scruggs, T. E., Mastropieri, M. A., Berkeley, S., & Graetz, J. E. (2009). Do special education interventions improve learning of secondary content? *A meta-analysis. Remedial and Special Education, 31*, 437-449. doi: 10.1177/0741932508327465

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

Sireci, S. G., Scarpeti, S. E., & Li, S. (2005). Test accommodations for students with disabilities: An analysis of the interaction hypothesis. *Review of Educational Research, 75*(4), 457-490. doi: 10/3102/00346543075004457

Semmelroth, C. L., & Johnson, E. (2013). Measuring rater reliability on a special education observation tool. *Assessment for Effective Intervention, 39*(3), 131-145. doi: 10.1177/1534508413511488

Sledge, A., & Pazey, B. L. (2013). Measuring teacher effectiveness through meaningful evaluation: Can reform models apply to general education and special education teachers? *Teacher Education and Special Education, 36*, 231-246. doi: 10.1177/0888406413489839

Stuhlman, M. W., Hamre, B. K., Downer, J. T., & Pianta, R. C. (2010). *A practitioner's guide to conducting classroom observations: What the research tells us about choosing and using observational systems*. Charlottesville: Center for Advanced Study of Teaching and Learning (CASTL), University of Virginia. Retrieved from http://curry.virginia.edu/resource-library/practitioners-guide-to-classroom-observations

Swanson, H. L. (2001). Searching for the best model for instructing students with learning disabilities. *Focus on Exceptional Children*, *34*(2), 1-15.

Taylor, E. S., & Tyler, J. H. (2011). *The effect of evaluation on performance: Evidence from longitudinal student achievement data of mid-career teachers*. [Working Paper 16877]. Cambridge, MA: National Bureau of Economic Research. Retrieved from http://www.nber.org/papers/w16877

U.S. Department of Education. (2009). *Race to the top*. Retrieved from http://www2ed.gov/programs/racetothetop/executive-summary.pdf

Weisberg, D., Sexton, S., Mulhern, J., & Kneeling, D. (2009). *The widget effect: Our national failure to acknowledge and act on teacher effectiveness*. New York, NY: The New Teacher Project.

Winters, M. A., & Cowen, J. M. (2013). Who would stay, who would be dismissed? An empirical consideration of value-added teacher retention policies. *Educational Researcher, 42*(6), 330-337. doi: 10.3102/0013189X13496145

Wright, S. P., Horn, S. P., & Sanders, W. L. (1997). Teacher and classroom context effects on student achievement: Implications for teacher evaluation. *Journal of Personnel Evaluation in Education, 11*(1), 57-67. doi: 10.1023/A:1007999204543