USING GENERALIZABILITY THEORY TO MEASURE SOURCES OF VARIANCE ON A SPECIAL EDUCATION TEACHER OBSERVATION TOOL

Carrie Lisa Semmelroth

A dissertation

submitted in partial fulfillment

of the requirements for the degree of

Doctor of Education in Curriculum and Instruction

Boise State University

August 2013

BOISE STATE UNIVERSITY GRADUATE COLLEGE

DEFENSE COMMITTEE AND FINAL READING APPROVALS

of the dissertation submitted by

Carrie Lisa Semmelroth

Dissertation Title: Using Generalizability Theory to Measure Sources of Variance on a Special Education Teacher Observation Tool

Date of Final Oral Examination: 12 June 2013

The following individuals read and discussed the dissertation submitted by student Carrie Lisa Semmelroth, and they evaluated her presentation and response to questions during the final oral examination. They found that the student passed the final oral examination.

Evelyn Johnson, Ed.D.	Chair, Supervisory Committee
Keith Allred, Ph.D.	Member, Supervisory Committee
Jonathan Brendefur, Ph.D.	Member, Supervisory Committee
Keith Thiede, Ph.D.	Member, Supervisory Committee

The final reading approval of the dissertation was granted by Evelyn Johnson, Ed.D., Chair of the Supervisory Committee. The dissertation was approved for the Graduate College by John R. Pelton, Ph.D., Dean of the Graduate College.

ACKNOWLEDGMENTS

Thank you foremost to all of the special education teachers who have participated in this project by allowing me access into their classrooms over the past two years, as well as to all of the teachers who have given up holiday and summer weekends to attend project data coding sessions. Without the support of these committed and dedicated special education professionals, this project, and the results presented in this dissertation, would not exist.

Similarly, I will forever be indebted to Dr. Evelyn Johnson, for without her tremendous support, I would have ended up a long ways away from the actual completion of my EdD! Thank you to Evelyn for your guidance, mentorship, tough talks, and good times! You have provided me with countless opportunities that I will undoubtedly benefit from for the rest my professional career. I hope to be able to continue working with you in the years to come.

Thank you to Dr. Jonathan Brendefur and Dr. Keith Thiede, two of the most influential professors in my doctoral program work. I am very grateful to have both of them on my dissertation committee because of their experience as researchers and professional educators. Dr. Brendefur's commitment to the classroom is evidenced by his dedicated time and visits with teachers to improve their instructional practice, and I hope to one day be able to follow his path as a true "bridge" between research and practice. And without Dr. Keith Thiede's guidance through my first interactions with quantitative

research, I'm not sure that I would have ever had the confidence to conduct this dissertation study.

Thank you to department chair and committee member Dr. Keith Allred for his ability to ask the "big" questions that helped drive deeper levels of inquiry in this study, especially those questions related to the study design and analysis.

Thank you to the countless public school teachers, support staff, and administrators who have paved the way for me to arrive where I am today. I am a true product of Idaho's public education system, having attended all of my K-12 years in the Boise School District at Franklin Elementary, North Junior High, and Boise Senior High School, as well as completing all three of my higher education degrees at Boise State University. I hope to begin giving back to the public education system that has shaped me into the person I am by dedicating my professional career to improving the lives of students through both research and practice.

Thank you to Bar Gernika for always taking such good care of me.

Lastly, thank you to my supportive network of friends and family who have put up with my student schedule for so many years! I owe you all many dinners, brunches, and lunches, as well as hikes, bike rides, dog walks, thrift store tours, summer nights, and vacations! I am very fortunate to have a life rich with such great and diverse people.

ABSTRACT

This study used generalizability theory to identify sources of variance on a pilot observation tool designed to evaluate special education teacher effectiveness, and was guided by the question: *How many occasions and raters are needed for acceptable levels of reliability when using the pilot RESET observation tool to evaluate special education teachers?* At the time of this study, the pilot Recognizing Effective Special Education Teachers (RESET) observation tool included three evidence-based instructional practices (direct, explicit instruction, whole-group instruction, and discrete trial teaching) as the basis for special education teacher evaluation. Eight teachers (raters) were invited to attend two sessions (October 2012 and April 2013) to evaluate special education classroom instruction collected from the 2011-2012 and 2012-2013 school years, via the Teachscape 360-degree video system. The raters were trained on the pilot RESET observation tool, and participated in whole-group coding sessions to establish interrater agreement (minimum of 80%) before evaluating assigned videos.

Data collected from raters were analyzed in a two-facet "partially" nested design where occasions (*o*) (observations/lessons) were nested within teachers (*t*), *o:t*, and crossed with raters (*r*), *{o:t} x r*. Using the results from the generalizability study analyses, decision studies were then completed to determine optimal facet conditions for the highest levels of reliability (the relative G coefficient and standard error of measurement scores were used to inform the decision study analyses). Results from this

study are in alignment with similar studies that found multiple observations and multiple raters are critical for ensuring acceptable levels of reliability. Recommendations for future studies include investigating the use of different raters (e.g., principals, university faculty, etc.), and using larger facet sample sizes to increase the overall measurement precision of the RESET tool. Considerations for the feasibility of practice must also be observed in future reliability and validity studies on the RESET tool.

Keywords: special education teacher evaluation, pilot observation tool, evidence-based instructional practice, generalizability theory

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: INTRODUCTION

Introduction

There are significant measurement and systemic challenges to evaluating special education teachers, and these challenges have become more prolific as the stakes have been made higher with recent changes to teacher evaluation policy (Holdheide, Browder, Warren, Buzick, & Jones, 2012; McGuinn, 2012). Special education teachers work under a variety of conditions, serve a heterogeneous population with diverse needs, do not enter the profession well-prepared, require a higher level of instructional skill to meet the needs of struggling learners, and face a field with higher levels of turnover and vacancies than other teachers (Billingsley, 2004; Connelly & Graham, 2009; Boe, Cook, & Sunderland, 2008; Gersten, Keating, Yovanoff, & Harniss, 2001; Holdheide et al., 2012). These factors make it difficult to 'fit' special education teachers into both existing and proposed models for teacher evaluation. Whether special education teachers are evaluated using mainstream tools like Charlotte Danielson's *Framework for Teaching* (2007) observation instrument, or evaluated using newer measurement systems like the Value-Added Model, there continues to exist significant gaps in teacher evaluation models that fail to: 1) provide relevant, specific feedback about special education instruction, and 2) address the significant challenges facing the profession, including the significant research-to-practice gap (Briggs & Domingue, 2011; Council for Exceptional Children, 2009; Chetty, Friedman, & Rockoff, 2011; Goe & Holdheide, 2011; Hanushek & Rivkin, 2010a; Ho & Kane, 2013; Holdheide et al., 2012; Holdheide, 2012; Kane & Cantrell, 2013; National

Council on Teacher Quality, 2012; Rockoff & Speroni, 2010; Semmelroth, et al., in press; Council for Exceptional Children, 2012).

In addition to the present issues that challenge the need for fair and reliable ways to evaluate special education teachers, new federal requirements for teacher evaluation systems have compelled states to include student outcomes as a primary component (U.S. Department of Education, 2012a; U.S. Department of Education, 2012b; Goe, Bell, & Little, 2008; McGuinn, 2012; National Council on Teacher Quality, 2011). As a result, states have been quick to adopt teacher evaluation policies and systems that may or may not be supported with empirical, or even historical, corroboration (McGuinn, 2012; Riley, 2012). Accordingly, these federal and state policy changes have shaped existing and new issues in relation to special education teacher evaluation and the profession.

Thus, in order to have an effective, fair special education teacher evaluation system that defines teacher effectiveness using student outcomes, the system must be able to not only meet the diversity found within special education teacher placements, but also address the current and historical challenges facing the profession (Danielson, 2011; Holdheide et al., 2012; Semmelroth et al., in press). The next section will explain in further detail how revisions in teacher evaluation policy have contributed to the challenges facing the special education profession.

Background

The last few decades of U.S. public education policy have addressed the issue of teacher effectiveness and teacher evaluation methods, but never as directly as within our current policy context. The era of No Child Left Behind (NCLB) helped to lay the groundwork for school accountability as an accepted part of public school culture (Baker

et al., 2010; Ravitch, 2010), facilitating the shift to the current policy focus on teacher accountability. From 1965 to the present, there have been a small number of influential federal policies that have influenced the focus of teacher evaluation policy from one that is removed from salary and compensation systems (National Council on Teacher Quality, 2011) to one that compels states to implement systems that define teacher effectiveness through some measure of student achievement (Holdheide, 2012; National Council on Teacher Quality, 2012). While most of the major legislative policies within the past 50 years have addressed teacher performance and competency to some extent, only now have policy efforts been this explicit.

Initiated by the funding attached to Race To The Top (RTTT) applications for states, followed by No Child Left Behind (NCLB) state exemptions through the Elementary and Secondary Education Act (ESEA) flexibility waiver, U.S. Secretary of Education Arne Duncan and the Obama Administration have prioritized federal education policy as one that: values the use of multiple methods to evaluate teachers, and prioritizes the use of student achievement as a primary measurement of a teacher's effectiveness (U.S. Department of Education, 2012a; U.S. Department of Education, 2012b; Murphy & Rainey, 2012). Although current bodies of evidence point to the importance of an effective teacher in a student's life (Chetty et al., 2011; Darling-Hammond, 2010; Hanushek & Rivkin, 2010a, 2010b; Martineau, 2006), the empirical body of evidence has yet to definitively answer to what extent teachers can affect student outcomes, as well as how to measure, define, and reward this effectiveness (Kane & Cantrell, 2013). This body of empirical evidence is especially scarce for special education (Buzick & Laitusis,

2010; Holdheide et al., 2012; Holdheide, 2012; Rockoff & Speroni, 2010; Rothstein, 2010; CEC, 2012).

Due to the changing federal policies and shifting focus of public school accountability, teacher evaluation systems that use multiple methods to measure teacher effectiveness have steadily risen amongst states (McGuinn, 2012; Council for Exceptional Children, 2012), and although student achievement is regarded as one of the primary predictors of a teacher's effect (Mihaly, Mccaffrey, Staiger, & Lockwood, 2013), there is a lack of research-based models or empirical evidence to support the various approaches (Goe & Holdheide, 2011; Newton, Darling-Hammond, Haertel, & Thomas, 2010), despite the increasing pressure for use of empirical evidence in public policy (Prewitt, Schwandt, & Straf, 2012).

While there are studies suggesting student achievement data can be used to predict teachers' impact on student outcomes in the future (Chetty et al., 2011; Hanushek & Rivkin, 2010a, 2010b; Kane & Staiger, 2012; Rockoff, 2004), there is no singular, research-based model or approach to measure teacher effectiveness (Goe et al., 2008; Goe & Croft, 2009). Additionally, measures used to assess teacher effectiveness are diverse and cannot be captured through the use of only one or two indicators (Partee, 2012). Similarly, there is little empirical evidence to inform how multiple-method teacher evaluation systems might weight each measure within a teacher effectiveness composite score (Kane & Cantrell, 2013; Mihaly et al., 2013), especially as each measurement in a multiple-measurement system can evaluate different aspects of teaching (Rothstein & Mathis, 2013). These measurement issues and concerns are particularly relevant within the special education context (Holdheide, 2012; Council for Exceptional Children, 2012)

as the profession is characterized in ways that can be problematic for valid and reliable measurements (e.g., small sample sizes, changing populations, individualized goals, etc.). Likewise, there is a significant gap of empirical evidence for many of the newer approaches for evaluating teacher effectiveness in relation to non-tested subject areas like special education (Buzick & Laitusis, 2010; Rockoff & Speroni, 2010; Rothstein, 2010).

Also known as "performance metrics" (Ehlert, Koedel, Parsons, & Podgursky, 2012), teacher effectiveness measures that are directly tied to student-achievement gains are growing in popularity and use because: 1) there is some empirical research showing that schools and teachers can differ in terms of their effect on test score (Hanushek & Rivkin, 2010b), and 2) both researchers and practitioners have had difficulty directly linking performance differences between schools and teachers to readily-observable characteristics (Hanushek & Rivkin, 2010b; Kane & Cantrell, 2013; Kane & Darling-Hammond, 2012). The two most common of these performance metrics that define a teacher's effectiveness through student achievement are the Value-Added Model (VAM) and student growth percentiles (SGP) (Betebenner, 2009; Hanushek & Rivkin, 2010a, 2010b; Rockoff, 2004).

VAMs are statistical models that use longitudinal data on students (usually in the form of student scores on state standardized assessments) to determine the "value added" of a particular teacher or school (McCaffrey, Lockwood, Koretz, Louis, & Hamilton, 2004). Proponents argue that while VAMs might not be methodologically ideal for all student groups or might not yet be fully tested and developed, it is still better than the current context of deficient teacher evaluation methods, approaches, and/or models (Chetty et al., 2011; Hanushek & Rivkin, 2010b; Kane & Cantrell, 2013; Kane & Staiger,

2012). However, opponents argue that VAM-based approaches to reward teacher performance are arbitrary and untested, and especially concerning for special education, none of the currently proposed systems have an empirical basis specific to the field (Baker et al., 2010; Goe & Holdheide, 2011; Holdheide, 2012; Holdheide, Goe, Croft, & Reschly, 2010; Johnson & Semmelroth, 2011; Kane & Darling-Hammond, 2012; National Council on Teacher Quality, 2011; Semmelroth et al., in press).

Similarly, while VAMs are used to answer how much "value" an effective teacher has "added" to a student's performance, SGPs seek instead to answer "How much growth did a student make?" (Betebenner, 2009, p. 42). SGPs capitalize on longitudinal data made available from over a decade of annual state assessment programs, creating what Damian Betebenner (2009) has called "an unprecedented opportunity to examine the academic growth of students" (p. 50). However, the use of longitudinal statistical models like SGPs (and VAMs) are problematic because: 1) the use of (and changes in) testing accommodations and modifications for students with disabilities are not accounted for in the homogeneity of standardized data; 2) a large percentage of students with disabilities who perform significantly below grade level may not be included in the standardized databases; 3) low-incidence disability subgroups (i.e. small populations) and changing disability classifications often translate as exclusion from state assessments and; 4) the psychometric properties of alternate and modified assessments may or may not meet state standardized assessment requirements (Buzick & Laitusis, 2010; Holdheide et al., 2012; Karvonen, Wakeman, Moody, & Flowers, 2012; van den Heuvel, Hansen, & Ilangakoon, 2012).

Opponents and critics to teacher evaluation systems that use performance metrics advocate instead for more holistic approaches and teacher quality-promoting approaches. Charlotte Danielson (2011), whose *Framework for Teaching* (2007) observation tool has been adopted as the teacher evaluation framework used in Idaho, maintains the two primary purposes of any teacher evaluation system should be to ensure teacher quality and promote teacher development. Danielson's emphasis on the importance of ensuring teacher quality and improving professional development is echoed from other leading researchers specializing in teacher evaluation. For example, Linda Darling-Hammond (2010) maintains that teaching is both an art and a science, while Ball and Forzani (2011) strongly remind us that teaching is inherently an unnatural skill that requires lots of ongoing, professional support. The inherent complexities of the need for good, quality teaching is only exacerbated within the special education context, where instruction is individualized, highly-technical, and complex (Baker et al., 2010; Browder & Cooper-Duffy, 2003; Foegen, Espin, Allinder, & Markell, 2001; Gersten, Vaughn, Deshler, & Schiller, 1997; Odom, 2009; Odom et al., 2005; Smith, Schmidt, Edelen-Smith, & Cook, 2013). Researchers like Danielson, Darling-Hammond, Ball, and others suggest that holistically ensuring teacher quality is a means to the end of improving student outcomes, as opposed to "performance metrics"-based positions that suggest student outcomes are evidence of strong teaching quality.

Furthermore, empirical studies on content-specific observation tools (e.g. Mathematical Quality of Instruction (MQI)), when used for formative teacher evaluation and tailored feedback, have been found to successfully improve teaching quality (Allen, Pianta, Gregory, Mikami, & Lun, 2011; Bloom, Hill, Black, & Lipsey, 2008; Hill et al.,

2008). Empirical work like that by Hill, Charalambous, and Kraft (2012) on content-specific observation systems that meet research-based rater criteria suggests that what might be needed is not a total rebuilding of states' teacher evaluation systems, but instead a refinement of current tools to promote the most important part of a teacher's effectiveness: classroom instruction. As Hill, Charalambous, and Kraft (2012) have noted, teaching quality is a critical element in teacher quality and that essentially "good teachers typically teach well" (p. 58).

Because special education teachers work within highly-specific but diverse instructional environments that include a variety of complex conditions, the stakes are especially high for developing a teacher evaluation system that ensures teaching quality and promotes professional development, as well as recognizes student achievement (Johnson & Semmelroth, 2012). In addition, given the current state of the special education teacher profession, an effective special education teacher evaluation system must be able to recognize and address the unique systemic challenges that special education teachers face (Boe et al., 2008; Holdheide et al., 2012; Spooner, Algozzine, Wood, & Hicks, 2010). Accordingly, in order to meet the needs of all major policy and research-based requirements, an effective special education teacher evaluation system must be characterized by features that allow for: 1) the evaluation of high-quality and evidence-based instructional techniques, 2) the measurement of teacher effectiveness using some measure of student growth or achievement, and 3) the flexibility to accommodate a variety of teaching contexts (Council for Exceptional Children, 2012). Unfortunately, current teacher evaluation methods (observation tools and performance

metrics) used for special education teachers do not support this theory of action (Holdheide, 2012; Semmelroth et al., in press; Council for Exceptional Children, 2012).

Thus, an evaluation system to measure special education teacher effectiveness should have the systematic goal of increasing attention on improving the quality and quantity of instructional services provided to students with disabilities. This study's approach to evaluating special education teachers is based on the observation of the special educator's use of evidence-based instructional practices, with future validity studies including the eventual inclusion of resulting student outcomes reported through effect sizes on evaluated evidence-based practices.

Based on the definition that an effective special education teacher is *able to identify a student's needs, implement evidence-based instructional practices and interventions, and demonstrate student growth*, a pilot observation tool has been developed to measure a special education teacher's use of evidence-based instructional practice and the resulting effect on student outcomes (Johnson & Semmelroth, 2012). The research and development on the pilot observation tool is funded by a two-year (2011-2013) grant from the Idaho State Department of Education called the Recognizing Effective Special Education Teachers (RESET) project, located in the Department of Special Education at Boise State University. The RESET project is tasked with two primary goals: 1) to define special education teacher effectiveness, and 2) to develop a tool to measure special education teacher effectiveness.

The study completed in this dissertation is part of a larger project to develop and validate a special education teacher observation measure, the pilot RESET observation tool, designed to evaluate instructional practice, provide feedback to special education

teachers about the quality of their instruction, and ultimately improve the outcomes for students with disabilities. To measure special education teacher effectiveness, the RESET observation tool evaluates a teacher's ability to implement evidence-based instructional practices that align with the classroom content and grade level, and accordingly adjusts to different placements, classrooms, grades, and exceptionalities (Johnson & Semmelroth, 2012). The tool consists of three main parts: the Lesson Objective (questions related to the lesson objective), the specific Lesson Components (questions based on specific evidence-based instructional practice components), and the Lesson Evaluation (overall evaluative questions). To construct the RESET observation tool, scoring criteria based a four-point Likert scale was developed (0-3), in alignment with Danielson's (2007) *Framework for Teaching* (the state's adopted teacher evaluation model) evaluation rubrics of: Unsatisfactory, Basic, Proficient, and Distinguished.

In summary, the past three decades of special education research has produced a foundational body of knowledge on the use and application of evidence-based instructional practices (Cook & Odom, 2013a; Cook, Tankersley, & Landrum, 2009; Graham, 2009; Horner et al., 2005; Odom et al., 2005). But, while arguably no other content area in education has produced more instructional practice research than special education, the profession itself has made little progress in practice (Smith et al., 2013). Improving special education teacher practice requires a systems-level change that includes evaluation systems that focus on measuring and improving instructional practice, and supporting teachers in professional development (Johnson & Semmelroth, 2012; Council for Exceptional Children, 2012).

Problem Statement

The current education policy emphasis on measuring teacher effectiveness using multiple measures and student outcomes has been met with disagreement coming from groups representing different interests in public education such as teacher unions, state departments, and local school districts (Baker & Santora, 2013; Baker, 2013; Baker et al., 2010; Watanabe, 2013). One of the largest areas of contention within these various interests lie within the issue of *how* teacher effectiveness can be measured using student achievement, especially when high-stakes decisions like teacher tenure, salary, and contract renewal may potentially be used based on the outcomes of these measures. These policy and measurement concerns are exacerbated when considered in the context of special education, especially given the historical problems still facing the profession (e.g. attrition, lack of qualified teachers, teacher dissatisfaction, etc.) (Boe et al., 2008; Gersten et al., 1997; Russ, Chiang, Rylance, & Bongers, 2001). Consequently, an effective special education teacher evaluation system that meets current policy requirements to define teacher effectiveness using student outcomes as a primary measurement, should 1) address the diversity found within special education classrooms, and 2) acknowledge the struggles found in the profession.

Purpose of the Study

This purpose of this study was to continue development of a pilot special education observation tool (RESET) by using generalizability theory to identify sources and levels of variance. Additionally, from the results of the generalizability studies, decision study analyses were also completed to identify optimal numbers of raters and teachers to maintain the highest levels of reliability when using the RESET tool. A total of eight special education teachers were trained to use the pilot RESET observation tool

to evaluate video observation of special education classroom instruction during two different sessions (June 2012 and April 2013). The rater data was captured using a web-based system (Qualtrics) that was then inputted into EduG v. 6.1 to run generalizability study analyses to identify sources of variances, followed by decision study analyses to determine the strongest levels of reliability in optimal observation conditions (using raters, teachers, and occasions as the facets of measurement) (Brennan, 2001; Cronbach, Gleser, Nanda, & Rajaratnam, 1972; Shavelson & Webb, 1991).

Research Questions and Hypothesis

This study sought to answer the following questions:

1) What sources of variance affect reliability across raters on the pilot RESET observation tool?

2) When organized by content subscales, which part of the pilot RESET observation tool demonstrates the strongest and weakest levels of reliability?

3) What are the optimal observation conditions to maximize reliability using the RESET observation tool?

In order to answer these questions, generalizability theory was used to identify contributing sources variance and minimize the largest sources of error with the ultimate goal of increasing the precision of the pilot RESET observation tool for future studies. Because generalizability theory answers open-ended questions related to multiple sources of contributing variance and error, the traditional null and alternative hypotheses used in a quantitative study were not used. In fact, not only are null hypotheses not needed to run

analyses, they're actually inconsequential to the design and methods of a generalizability study; instead, studies that use generalizability theory seek to ask, "How many instances of which conditions of measurement are needed for acceptably precise measurement?" (Brennan & Lee, 2013, p. 3). (A much more detailed description of the issues related to generalizability theory and measurement will appear in Chapter 2: Literature Review.)

Thus, for this study, the primary research question guiding the generaliability and decision study analyses was: *How many occasions and raters are needed for acceptable levels of reliability when using the pilot RESET observation tool to evaluate special education teachers?*

Nature of the Study

This quantitative study was designed to examine sources of variance on a pilot special education observation tool, which evaluates special education teacher effectiveness based on the teacher's use of evidence-based instructional practice. All evaluative rater data was collected from two data coding sessions (June 2012 and April 2013) that were held at Boise State University with five trained raters using the pilot RESET observation tool. For the generalizability study, a two-facet nested design was used: the object of measurement was teachers (t), and the facets were raters (r) and occasions (o) (Shavelson & Webb, 1991, pp. 52-54). Decision studies were completed to apply measurement information gathered from the generalizability theory analyses to decompose varying levels of reliability between facets (Shavelson & Webb, 1991), which informs optimal levels of raters and occasions to maximize reliability. A more detailed description of the methodology that was used for this study is provided in Chapter 3.

Overview of the Pilot RESET Observation Tool

Recent changes in federal policy have required states to apply for federal education funds by meeting a pre-determined set of criteria, with very specific requirements for teacher evaluation systems (U.S. Department of Education, 2012a; U.S. Department of Education, 2012b; McGuinn, 2012; National Council on Teacher Quality, 2012). Two of the most important new federal policy requirements regarding teacher evaluation is the shift to systems that use 1) student outcome measures as a component of teacher evaluation, and 2) multiple methods of measurement for teacher evaluation (National Council on Teacher Quality, 2012; Newton et al., 2010).

In effect, these new federal requirements have signaled the legitimization of states to rebuild their teacher evaluation systems into ones that use student outcomes as direct measurements of a teacher's ability and effectiveness. These legislative changes in teacher evaluation suggest a new focus for accountability (i.e., moving away from the whole-school accountability to teacher accountability) (Mehta, 2013). This policy movement towards a multiple-method, student-outcome based system to evaluate teacher effectiveness has compelled states like Idaho to propose new legislation like Students Come First (2011), which require local districts to revise teacher evaluation policies. It is within this context that the Recognizing Effective Special Education Teachers (RESET) grant project was established. The RESET project has two primary goals: 1) to define special education teacher effectiveness, and 2) to measure special education teacher effectiveness using student outcomes as a primary measure.

The RESET project defines effective special education teachers as those teachers who *are able to identify a student's needs, implement evidence-based instructional*

practices and interventions, and demonstrate student growth (Johnson & Semmelroth, 2012). The RESET observation tool assumes that the quality of instruction that special education teachers provide to their students is a key determinant of a student's individual growth. A significant body of research establishes a number of effective instructional practices to meet the needs of students with disabilities (e.g. Baker, Chard, Ketterlin-Geller, Apichatabutra, & Doabler, 2009; Browder & Cooper-Duffy, 2003; Browder, Spooner, Harris, & Wakeman, 2008; Chard, Ketterlin-Geller, Baker, Doabler, & Apichatabutra, 2009; Cook & Odom, 2013a; Fuchs & Fuchs, 2005; Gersten et al., 2009; National Autism Center, 2009; Odom, 2009; Odom et al., 2005; Odom, Cox, & Brock, 2013; Spooner, Knight, Browder, & Smith, 2012). Aligning the evaluation system to provide feedback on the specifics of instructional practice provides special education teachers the opportunity and information needed to improve their practice.

The pilot RESET observation tool is based on the following principles:

1. RESET is grounded in Danielson's framework with a focus on Domain 3: Instruction. However, it includes much more clearly delineated criteria for evaluating evidence-based instructional practice appropriate for students with disabilities.

2. RESET is a computerized, evaluation system that relies on the use of video capture of instruction. The video is evaluated by a trained observer who can evaluate the quality of the instruction following the RESET criteria.

3. Special education teachers evaluated by RESET will receive feedback on the specific dimensions of their teaching according to criteria derived from research identified effective practice.

4. Effective teaching is highly correlated with student outcomes based on effect sizes. Reported effect sizes serve as a reasonable estimate of anticipated student growth if a practice is implemented with fidelity (Johnson & Semmelroth, 2012).

The RESET observation tool is feasible for schools because the use of video capture will allow special education teachers and administrators the flexibility in scheduling that is often an issue for conducting evaluations (Foegen et al., 2001; Odom et al., 2003). Thus, in a profession that is characterized by high-turnover and lack of highly-qualified educators, networks of newly certified teachers, trained mentors, and consulting special education teachers can connect virtually, bridging gaps defined by distance and lack of time and resources (Boe et al., 2008; Gersten et al., 1997; Russ et al., 2001; Vannest & Hagan-Burke, 2009). From an assessment design perspective, video capture also affords the opportunity to conduct large enough datasets for statistical and psychometric analyses of RESET.

In addition to being aligned with Danielson's evaluative rubrics, the pilot RESET observation tool is grounded in research through the use of evidence-based instructional practices to evaluate special education teacher effectiveness. By creating a systematic, purposeful link between evidence-based practices developed in the research setting and the practical application found in the classroom setting, the pilot RESET observation tool aims to: 1) close the research-to-practice gap found in special education, 2) address the systemic and historical challenges found within the profession, and 3) ensure teacher quality and promote professional development. These goals are addressed through the

overall, larger purpose of the tool: to identify and measure special education teacher effectiveness.

The pilot RESET observation tool therefore focuses on the core component of teacher practice, instruction. RESET includes evaluation criteria aligned with the characteristics of evidence-based practice, so that teachers can be provided direct feedback on their ability to implement evidence-based practices to support student outcomes. When special education teachers are provided feedback on specific elements of their instructional practice, they will better understand the evidence-based practice and be able to improve their ability to implement.

Conceptual Framework of the Pilot RESET Observation Tool

The pilot RESET observation tool aims to meet the two purposes of what Charlotte Danielson (2011) maintains is critical for any effective teacher evaluation system to: 1) ensure teacher quality and 2) promote professional development. To design the pilot RESET observation tool, the five stage Evidence-Centered Design (ECD) approach to measurement outlined by Mislevy, Steinberg, and Almond (2003) was used. ECD follows five stages to developing assessments that comprehensively measure a complex construct. These stages include: a) Domain Analysis, b) Domain Modeling, c) Conceptual Assessment Framework, d) Assessment Implementation, and e) Assessment Delivery. Each of the stages is used to guide the design and conceptualization of RESET and is outlined below.

Domain Analysis

As the first stage in assessment design, Domain Analysis leads the assessment developer to understand the knowledge people use in a domain, the representational forms, characteristics of good work, and features of situations that evoke the use of valued knowledge, procedures, and strategies (Mislevy & Haertel, 2006). The Domain Analysis stage involves collecting substantive information about the domain being assessed, in this case, effective special education teaching. Pilot work on the development of the RESET observation tool has been primarily focused in the activities associated with the Domain Analysis stage.

In the Domain Analysis stage, a definition of effective special education teaching was developed. First, the research was reviewed on teacher impact to determine the critical importance of the teacher's role in affecting student outcomes. Next, a review of research within special education was completed to identify the specific instructional practices that have a research base to establish efficacy. Three primary sources informed our work in the Domain Analysis stage. These include: a) Danielson's *Framework for Teaching* (2007), Domain 3: Instruction, b) Council for Exceptional Children (CEC) Professional Standards for Special Education Teachers (2009), and c) a meta-review of the literature on effective special education instructional practice. Based on this process, the following definition was developed: an effective special education teacher is able to identify a student's needs, implement evidence-based instructional practices and interventions, and demonstrate student growth (Johnson & Semmelroth, 2012).

This gap in Danielson's framework can be filled by including the criteria that are specific to the instructional strategies that are most effective for meeting the needs of students with disabilities. The most prominent framework for defining the qualities and

characteristics of effective special education teaching is the Council for Exceptional Children's (CEC) professional standards. The CEC developed initial standards outlining the knowledge and skills that special educators should bring to both initial and advanced roles. The underlying premise is that achievement of these standards will adequately prepare special education teachers to teach students with disabilities effectively (Ashton, 2011; Council for Exceptional Children, 2009). Although the professional standards do not directly specify instructional strategies, standards -- such as, conducts task analysis to determine discrete skills necessary for instruction; designs and implements positive behavior intervention strategies; plans instruction that is appropriate to the needs of the individual student; and individualizes instruction to support student learning in various settings -- imply the importance of being well-versed in evidence-based instructional strategies. These general descriptors of effective instructional practice guided our initial research reviews on special education practice.

The research on instructional practice in special education includes over four decades worth of research on a variety of instructional strategies designed to meet the needs of various disability types. Several meta-analyses of instructional practice have been undertaken over the years in special education and provide helpful starting points for explicating the key elements of an instructional strategy (Baker et al., 2009; Bellini, Peters, Benner, & Hopf, 2007; Berkeley, Scruggs, & Mastropieri, 2009; Browder et al., 2008; Dexter, Park, & Hughes, 2011; Gersten et al., 2009; Gersten, Jordan, & Flojo, 2005; Test, Richter, Knight, & Spooner, 2010). From these meta-analyses, common definitions of different instructional practices can be developed, along with the specification of the particular elements that are essential to define the practice. In addition

to providing guidance on instructional characteristics, meta-analyses also provide data on a range of effect sizes that help gauge the expected outcomes for students with disabilities when specific instructional strategies are used.

Domain Modeling

The Domain Modeling stage in the process takes the information and relationships discovered in the Domain Analysis component and considers how to translate them into assessment design options or assessment argument (Mislevy & Haertel, 2006). For teaching, a common design option is to center the information from domain analysis into a Knowledge, Skills, and Abilities (KSA) framework to begin to suggest options for how assessment can be designed to obtain evidence of those KSAs.

To begin the Domain Modeling stage for RESET, a matrix was developed to crosswalk Danielson's Framework for Teaching (2007) with the CEC professional standards related to instruction and included specific evidence of a variety of instructional practices with a strong research base in special education. From this crosswalk, a model of effective special education teaching is defined as those who engage in the delivery of evidence-based instructional practices that support the academic growth of students with disabilities (Johnson & Semmelroth, 2012). The domain of effective teaching is best assessed through performance tasks, or observations of their instructional practice, and validated by including and analyzing the growth achieved by students who are provided with effective instruction. Other elements of special education teacher responsibilities, such as conducting IEP meetings or completing paperwork, were not included; although these are critical requirements of the job, there is currently no research base linking the successful completion of these administrative tasks to student outcomes.

Within the Domain Modeling stage, the characteristic and variable features of tasks specify aspects of the situation in which teachers produce performance tasks (Mislevy & Haertel, 2006). Characteristic features are those that all assessment tasks motivated by the design pattern should possess in some form. Variable features address aspects of the assessment that can be used to affect the focus of attention (Mislevy & Haertel, 2006). The characteristic tasks that will be common across all special education teachers include the recording of a teaching context in which a special education teacher is directly working with students in an instructional setting. Because teaching contexts and instructional settings are highly variable in special education, the variable features of RESET will include criteria for evaluating a number of instructional practices. For example, special education teachers may be working with students with autism in an extended resource room, or working with students with high-incidence disabilities in a general education classroom in a team teaching setting. These variable features are the aspects of the evaluation tool that would focus attention to a specific teaching context, allowing RESET to be flexible and responsive to the diverse contexts in which special education teachers work.

Conceptual Framework for Assessment

The Domain Analysis and Domain Modeling stages lead the measurement developer towards creating a conceptual framework for the proposed assessment. The conceptual framework guiding RESET is that through a targeted, well-defined observation that incorporates clearly explicated criteria linked to evidence-based practices in special education, teacher attention will be targeted to those instructional practices that have been demonstrated to result in improved student outcomes. RESET

will be able to discriminate when research-based instructional practices are implemented with fidelity, provide explicit feedback to the teacher on the specific components of instructional practices that need improvement, and to demonstrate a link between the implementation of research-based practice and impact on student outcomes. The operational definition derived from this conceptual framework is that effective special education teachers implement relevant (appropriate to population, context and content) evidence-based instructional practices with fidelity in order to improve student outcomes.

Assessment Implementation

The operational definition derived from the conceptual framework leads to the fourth step in evidence-centered design (Mislevy et al., 2003), Assessment Implementation. This is the stage at which assessment items are created. (This stage is an ongoing part of the RESET project, but initial work has been completed to date on assessment item development. Further work on assessment implementation will be addressed in future project activities.)

To collect evidence establishing the use of research-based practices, the assessment relies on video captures of special education teacher instruction that are evaluated according to relevant criteria based on the characteristics of effective instruction identified in the research base. As with similar studies, some of the considerations about the use of video capture that will need to be refined in future work includes the required length of each video to obtain a valid evaluation, the number of observations per teacher required to obtain a reliable evaluation, the interrater reliability across different evaluators (i.e., principal or special education director), and how to assess when more than one instructional strategy is in use (Bell et al., 2012; Hill, Charalambous,

& Kraft, 2012; Ho & Kane, 2013; Kane & Staiger, 2012; Moscoso, Tello, & López, 2006; Shavelson & Dempsey-Atwood, 1976).

Because special education teachers find themselves working in a variety of contexts, settings, and with a very heterogeneous population, the evaluation criteria -– which in RESET are equivalent to assessment items -- needs to encompass this range. To do this, the rater (evaluator) will first identify the instructional context and setting observed, and this choice will then direct the evaluator to the criteria relevant for that instructional strategy. For example, if the evaluator were rating a special education teacher providing a small group, direct-instruction reading lesson, the set of criteria used to evaluate direct instruction of reading as identified in the research would be used.

The evaluation of instructional practices will result in a score for each strategy on which the special education teacher was evaluated. This provides a 'component' score. Because RESET is grounded in Danielson's framework, the initial scoring criteria in pilot procedures for scoring (i.e. this study) are on a 0-3 scale, where a 0 is consistent with Danielson's 'Unsatisfactory', a 1 with 'Basic', a 2 with 'Proficient', and a 3 with 'Distinguished'. Scores are provided at the element, component, and domain levels in Danielson's framework. On the pilot RESET observation tool, scores are provided at the element (each individual characteristic of the instructional practice) and component (the instructional practice) levels, with an overall domain score restricted to Domain 3: Instruction. High scores indicate that the teacher has implemented the specified instructional practice in accordance with the research-based elements of that procedure, and lower scores indicate that the teacher has not implemented the specific instructional strategy with fidelity.

The assessment tasks of RESET must also provide a means for collecting evidence of student growth. Student growth measures and criteria for evaluating that growth will need to be established in future project work. As with the instructional practice criteria, student growth criteria will vary based on disability type, context and content area. For example, a relevant outcome measure for students in a reading group could be their growth, measured in effect sizes, on standardized measures of reading. The documented effect size will be compared to the research-reported range of effect size for this instructional strategy. Effect sizes for single-case research will be informed through the increasing literature found for non-overlapping techniques (Parker, Vannest, & Davis, 2011).

Validation Activities (Assessment Delivery)

The fifth stage of evidence-centered design includes assessment delivery: i.e., the stage at which the items are piloted and feedback is collected, reviewed, and integrated into the final design of the assessment tool. Although not a part of this study, in general, the validation activities planned for this project include: 1) determining the reliability of evaluations across times, across teachers, and across raters, 2) examining the results of RESET as compared to other measures of teaching effectiveness, 3) determining the extent to which ratings on instructional evaluation and student growth correlate, and 4) examining the impact of RESET feedback on instructional practice over time.

Operational Definitions

Evidence-Based Practice

Defining “evidence-based practices” can become problematic given the multiple perspectives and approaches that exist. A great illustration of this is the difference between What Work’s Clearinghouse (WWC) review of empirical studies and Robert Slavin’s and othersn work on the Best Evidence Encyclopedia (BEE). For example, both WWC and Slavin and others maintain many of the same requirements for an “evidence-based practice,” but whereas WWC requires a study to be randomized, the BEE does not adhere itself strictly to this requirement (Slavin, Lake, Davis, & Madden, 2009; Slavin & Madden, 2011; *What Works Clearinghouse: Procedures and standards handbook (version 1.2)*, 2011). Similarly, special education researchers have established a long, ongoing conversation about what it means to have an “evidence-based practice”; a conversation that crosses exceptionality, content and incidence (Browder & Cooper-Duffy, 2003; Cook et al., 2009; Graham, 2009; Horner et al., 2005; Odom, 2009; Odom et al., 2005; Odom, Collet-Klingenberg, Rogers, & Hatton, 2010; Roberts, Torgesen, Boardman, & Scammacca, 2008; Spooner et al., 2012).

Because of the complexity and discrepancies surrounding the classification of the term “evidence-based instruction,” in this study it will be defined broadly using Cook and Odom's (2013) most recent requirements for a practice to be considered evidence-based: “it must be supported by multiple, high-quality, experimental or quasi-experimental (often including single-case research) studies demonstrating that the practice has a meaningful impact on student outcomes” (p. 136).

Special Education Teacher Effectiveness

An effective special education teacher is one that "is able to identify a student's needs, implement evidence-based instructional practices and interventions, and demonstrate student growth" (Johnson & Semmelroth, 2012).

Interrater Agreement

Interrater agreement is defined as the degree to which two or more raters achieve identical results under similar assessment conditions" (Brennan & Prediger, 1981; Landis & Koch, 1977).

Generalizability Theory

Generalizability theory or "G theory" is "a statistical theory about the dependability of behavioral measurements (Shavelson & Webb, 1991, p. 1) and "the strength of G theory is that multiple sources of error in a measurement can be estimated separately in a single analysis" (Shavelson & Webb, 1991, p. 2).

Generalizability Study

A generalizability study or "G study" collects data from which "estimates can be made of the components of variance for measurements made by a certain procedure" (Cronbach et al., 1972, p. 16).

Decision Study

A decision study or "D study" collects data "for the purpose of making decisions or drawing conclusions" (Cronbach et al., 1972, p. 16). A "D study makes use of the information provided by the G study to design the best possible application of the social science measurement for a particular purpose" (Shavelson & Webb, 1991, p. 12).

Reliability

Reliability in this study is defined through the use of generalizability theory, which allows for the examination of multiple influences on score reliability within a single analysis (Brennan, 2001; Cronbach et al., 1972; Shavelson & Webb, 1991).

Assumptions, Limitations, and Scope

Assumptions for this study are: 1) as Hill, Charalambous, and Kraft (2012) noted, "good teachers teach well" and it is assumed that these characteristics can be observed through observation; 2) primary sources of variance will be coming from raters (r) and occasions (o) (as opposed to inherent flaws with the tool); and 3) the video observation data that will be used by raters is a fair and appropriate representation of special education instruction.

Possible limitations to this study include a lack of generalizability of the results to other raters and teachers because of: 1) the pilot stage of the developing RESET observation tool, 2) the convenience sampling, and 3) the small number of raters, teachers, and items. The scope of this study included the rater reliability of the evaluation of evidence-based instructional practices used by special education teachers from selected districts in Idaho.

Significance of the Study

This study sought to identify sources of variance on a pilot special education observation tool. It is expected that the results of this study will help to inform both future studies (reliability and validity) and versions of the RESET observation tool, which is the only tool known to this date that evaluates the effectiveness of a special education teacher based on his/her use of evidence-based instructional practices. While there are multiple approaches to teacher evaluation, at this time there are only two known in development that are specific to special education, the Classroom Observations of Student–Teacher Interactions (COSTI), developed to quantify the rates of specific instructional interactions that occur between teachers and their students (Doabler, Fien, Nelson-Walker, & Baker, 2012; Smolkowski & Gunn, 2012), and an "opportunities to learn"-based approach developed through MyiLOGS (Elliott & Kurz, 2012). Because of the RESET observation tool's emphasis on instructional practices, it is expected that the results of this and future studies will lead to increased, positive outcomes for student with disabilities.

Summary

There are significant challenges to designing an effective special education teacher evaluation system, and there is a growing need to improve the quality of special education teacher professionals as evidenced by the poor outcomes for students with disabilities. Current approaches to teacher evaluation have not been validated for use with special education teachers, and in their design do not adequately address the challenges of special education.

The pilot RESET observation tool offers a method that is more consistent with the use of evidence-based practices for students with disabilities, and provides a blueprint for special education teachers to improve instructional practice. Consistent with other researchers' perspectives on evidence-based instructional practices, the pilot RESET observation tool is based on the idea that increased use of effective evidence-based instructional practices will lead to increases in student outcomes (Cook & Odom, 2013; Odom et al., 2005, 2010). Effective instruction is expected to lead to gains in student performance consistent with the range of effect sizes achieved in the research on instructional practice. This very basic connection between effective instructional practice and student outcome data drives the core of the conceptual framework for RESET. The other important tenet of the RESET framework is that when special education teachers are provided feedback on specific elements of their instructional practice they will better understand the evidence-based practice and be able to improve their ability to implement each component. In this way, a special education teacher evaluation system that focuses on the effective use of evidence-based instructional practices, outcomes will include: 1) targeted, specific, corrective feedback for teacher instructional practice, 2) quantitatively defined levels of teacher effectiveness identified through appropriate use of evidence-based instructional practices, 3) the use of student growth rates (through effect sizes) to define teacher effectiveness, and 4) adaptability to do all three of these outcomes within all special education classrooms. The five stage Evidence-Centered Design (ECD) approach to measurement is the conceptual framework for the development of the pilot RESET observation tool and for future studies related to the development of the RESET teacher evaluation system (Mislevy et al., 2003).

This study analyzed sources of rater variance to further develop and refine the pilot RESET observation tool for eventual implementation and use at the practitioner level. A review of the literature is presented in Chapter 2. Key topics in Chapter 2 include an overview of teacher evaluation methods, the current state of the special education teacher profession, issues and challenges related to special education teacher evaluation, and the use of generalizability and decision studies to analyze sources of variance. Chapter 3 follows with a description of the methods and procedures for the proposed study. Chapter 4 includes the results of the study, followed by the interpretation of results, discussion, and recommendations in Chapter 5.

CHAPTER 2: LITERATURE REVIEW

Introduction

According to the Center for American Progress, "improving teacher quality has become the centerpiece of the Obama administration's education agenda and of the contemporary school-reform movement" (McGuinn, 2012, p. 3). As a result, policymakers and researchers have identified the task of developing new teacher evaluation systems as a crucial part for both improving teacher quality and increasing student achievement (McGuinn, 2012; National Council on Teacher Quality, 2011, 2012). Although teacher evaluation has emerged as a prominent educational policy issue, there has also emerged many challenges that highlight how difficult this type of reform can be (Darling-Hammond, Amrein-Beardsley, Haertel, & Rothstein, 2012).

Chapter 2 begins with an overview of teacher evaluation, followed by a discussion of the issues related to classroom observations and performance metrics currently being used to evaluate teacher effectiveness. Next, the literature review narrows the focus on special education teacher evaluation, and the unique challenges and issues associated with special education teacher evaluation. Finally, the chapter concludes with a discussion on generalizability and decision studies, and the rationale for using these types of analyses for this study.

Teacher Evaluation

Since 2009, 36 states and the District of Columbia have made policy changes both in legislation and in practice to their teacher evaluation systems (National Council on Teacher Quality, 2012). In accordance with federal guidelines (U.S. Department of Education, 2012a, U.S. Department of Education, 2012b), teacher evaluations can now include combinations of different tools (e.g., multiple method systems), as well as new, non-research-based approaches (e.g., Value Added Model), that reflect the current paradigm shift from school to teacher accountability (Mehta, 2013). However, there is yet to exist a national system of supports and incentives to ensure that all teachers are well-prepared and ready to teach all students effectively when they enter the profession, nor are there readily available methods to support the evaluation and ongoing professional development of teacher effectiveness (Darling-Hammond, 2010).

The State of Teacher Evaluation

While previous policy and systemic rationales for teacher evaluation may have been more procedural or process-oriented, current reform-based approaches to teacher evaluation are driven by accountability (Riley, 2012). Teacher evaluation is no longer a reflection of a contractual obligation, or professional development, instead it is increasingly being used as a measure to hold teachers directly responsible for student achievement (Lewis & Young, 2013; McGuinn, 2012). These changes in teacher evaluation system requirements have compelled states to redefine how teacher effectiveness is measured by: 1) creating a direct relationship between teacher effectiveness and student outcomes and 2) using multiple methods to measure teacher effectiveness (U.S. Department of Education, 2012a, U.S. Department of Education,

2012b). For example, in 2009, only four states were using student achievement as an important criterion in how teacher performance was assessed, but in 2012 that number had increased to 22 states (National Council on Teacher Quality, 2012).

Given the different ways teacher effectiveness can be defined, it is not surprising that multiple approaches for evaluating teachers exist (Goe & Croft, 2009). Currently, the two most widely used measures to evaluate teacher effectiveness are classroom observations and performance metrics (e.g., VAMs, SGPs, etc.), while other methods include principal evaluations, portfolios, teacher self-reports of practice, including surveys, teaching logs, and interviews, and student and parent ratings of teacher performance (Goe & Croft, 2009; Kane & Staiger, 2012; Winters & Cowen, 2012). Although each type of teacher evaluation measurement highlights a particular aspect of teaching quality, most reform efforts have focused on just two indicators: observations and student test scores (Jones, Buzick, & Turkan, 2013). Thus, in the next section, a discussion of these two most commonly used methods to evaluate teachers -- classroom observations and performance metrics -- will be reviewed, followed by the advantages and disadvantages of each.

Classroom Observations

Up until recently, most states have approached teacher evaluation using a combination of formative and/or summative classroom observations by principals or on the accumulation of teacher qualifications such as completion of a preparation program, number of degrees, or years of teaching experience (Ehlert et al., 2012; Goe et al., 2008; Prince et al., 2009). In Idaho and in many other states, Charlotte Danielson's (2007) *Framework for Teaching (FFT)* has been adopted as the teacher evaluation system for

use by locally controlled districts. Danielson's FFT is organized around four domains of teaching responsibility: planning and preparation, classroom environment, instruction, and professional responsibilities, which are broken into 22 components and 76 elements.

The FFT observation tool was most recently involved in the Measures of Effective Teaching (MET) study, and was compared against other teacher evaluation frameworks (Kane & Staiger, 2012). Because this is the first large-scale comparison of multiple instruments with the same group of teachers and their outcomes and the field is at an early stage in the evolution of observation instruments, the results from the study were mixed. Overall, the results from the MET study indicate there is little to no relationship between a teacher's performance on the FFT tool and student achievement (Kane & Staiger, 2012). The correlation increases when FFT is included in a multiple methods approach (i.e., VAM-based), but results are still preliminary (Kane & Cantrell, 2013; Kane & Staiger, 2012). Nevertheless, from the results of the MET study, Danielson (2011b, 2013) revised the FFT guidelines to enhance the identification of a teacher's performance levels by tightening the rubric language, adding "critical attributes," and developing illustrative examples for each component.

Performance Metrics

Results from experimental studies have shown that teachers differ in their effect (Chetty et al., 2011; Konstantopoulos & Chung, 2010; Rockoff, 2004), giving increased political and practical attention to "performance metrics" (Ehlert et al., 2012, p. 4) like the value-added model (VAM) and student growth percentile (SGP) approaches to evaluating teacher effectiveness (Rothstein, 2010). Although performance metrics can be formulated and defined in different ways, the essential purpose of the method is to use

student achievement data to predict a teacher's influence on future student performance (Betebenner, 2009; Chetty et al., 2011; Martineau, Paek, Keene, & Hirsch, 2007; Martineau, 2006; Newton et al., 2010). In general, performance metrics define a relationship between teacher effectiveness and student academic achievement through weighted statistical formulas that incorporate values primarily through a teacher's effect on a student's performance on a state assessment (McCaffrey et al., 2004).

For example, VAMs have been formulated to predict teacher effectiveness at varying levels, including at the district, whole school, and teacher/classroom. Performance metric concepts have been used as the basis for other approaches, like Damian Betebenner's (2009) development of student growth percentiles. Other performance metric approaches begin more conceptually to address the complexities of the statistically 'noisy' school environment, like Joseph Martineau's (2006) work on vertical versus horizontal alignment (see also: Martineau et al., 2007). Regardless of the statistical formula, teacher evaluations based on student achievement and growth has nevertheless stimulated discussions concerning what statistical models and properties that can be used to measure the "value-added" or student "growth" of the teacher effect (Betebenner, 2009; Chetty et al., 2011; Hanushek & Rivkin, 2010a; Heck, 2007; Konstantopoulos & Chung, 2010; Mariano, McCaffrey, & Lockwood, 2010; Mihaly et al., 2013).

Proponents of performance metric approaches to teacher evaluation argue that existing research confirms that individual teachers do have an impact on student gains and despite some fluctuation from year to year, a teacher's record of promoting achievement remains the strongest single predictor of the achievement gains of their

future students (Chetty et al., 2011). For example, Konstantopoulos and Chung (2010) found that sixth grade students who have "very effective" teachers at the 85th percentile of the teacher effects distribution (the researchers assumed that teacher effects are normally distributed) in six consecutive grades (K-5) would experience achievement increases of about one-half of a standard deviation in mathematics and reading (p. 383). In contrast, students who have "low effective" teachers (bottom half of the teacher effects distribution) from K-5 resulted in a negative effect on sixth grade achievement, and the disadvantage ranged between one-fifth and one-third of a standard deviation (Konstantopoulos & Chung, 2010, p. 383). In another example, Hanushek and Rivkin (2010a) note that eliminating "6-10 percent of the worst teachers could have strong impacts on student achievement, even if these teachers were replaced permanently with just average teachers" (p. 3).

It is also argued that performance metrics like value-added models (VAMs) can lessen the penalization for those who instruct students from less-advantaged backgrounds by accounting for changes in student scores longitudinally, using databases across individual teachers who have instructed the students (Braun, 2012); because it is just growth in achievement that is being studied, it is argued that a VAM can reduce the effect of factors intrinsic to the student and his/her background (Braun, 2012; Chetty et al., 2011; Hanushek & Rivkin, 2010b).

Opponents of performance metrics criticize the approach for multiple reasons, primarily for those based in empirical (i.e., lack of empirical evidence) and pedagogical (i.e., does not address teaching quality) rationales. The most serious of these criticisms charge the lack of empirical support for their implementation and use. Even VAM

researchers like Kane and Staiger (2012) and Hanushek and Rivkin (2010b) caution the sole use of any performance metric to define teacher effectiveness, and while the approach shows promise as a predictive tool of a teacher's performance, it should not be separated from a multiple methods approach. Methodologically, there are significant issues that remain unanswered within the varying performance metric frameworks (e.g., lack of randomization in estimating teacher effects). As Braun (2012) observes, "the fundamental concern is that, if making causal attributions is the goal, then no statistical model, however complex, and no method of analysis, however sophisticated, can fully compensate for the lack of randomization" (p. 8).

Performance metrics also fall under considerable criticism for the lack of empirical information regarding both the tested and untested participants (Goe & Holdheide, 2011; Holdheide et al., 2010). The untested groups are sometimes referred to as the "other 69%" and include: non-tested subjects (e.g., art, music, physical education), non-tested grades (e.g., pre-kindergarten to Grade 2 and high school), English language learners, and students with disabilities (Prince et al., 2009). It still unclear how to measure the teacher effects of non-tested subjects using performance metrics because: 1) there is very little empirical evidence about teacher effects outside of math and reading core content areas, 2) there is little to no empirical evidence linking the extent to which teachers of untested subjects contribute to gains in student achievement tested areas, and 3) it is more difficult in some subjects than in others to obtain reliable estimates of teachers' contributions to their students' performance, suggesting that there may be other sources of variance that are unaccounted for (e.g., principal effects, home environment, etc.) (Ballou, Sanders, & Wright, 2004; Braun, 2012; Briggs & Domingue, 2011;

Goodman & Turner, 2010; Konstantopoulos & Chung, 2010; Lipscomb, Teh, Gill, Chiang, & Owens, 2010; Lohr, 2012; Prince et al., 2009; Sawchuk, 2012).

Another area of criticism that performance metrics are subject to is through the determination of a composite score. Not only might the use of composite scores invite misleading and overly simplistic policy conclusions if they are misinterpreted or poorly constructed (especially when considering measurement error issues), but perhaps even worse, they may be misused to support predetermined policies if the process of constructing them is not transparent or not readily understood (Hanushek & Rivkin, 2010b; Mihaly et al., 2013). In circumstances with potential multiple sources of selection bias, and/or less comprehensive data than is statistically ideal, due diligence will require "a careful look under the hood" (Braun, 2012, p. 16), which may be skipped or overlooked when LEAs and SDEs are burdened with overly complicated teacher evaluation models. In addition, studies suggest that teacher effects decrease as students get older, confounding both policy and research decisions regarding how to categorize and define cutoff scores, composite scores, etc. (Konstantopoulos & Chung, 2010; Voight, Shinn, & Nation, 2012).

Lastly, a performance metric approach to measure teacher effectiveness does not nuance between varying levels of teacher quality, nor is it able to provide any formative, targeted feedback to improve instructional practice. A performance metric composite evaluation score may "disguise serious failings on some dimensions and increase the difficulty of focusing remedial action" (Mihaly et al., 2013, p. 4), leaving the performance metric teacher evaluation system unable to meet the two primary features

Danielson (2011) maintains are crucial for effective evaluation: 1) ensuring teacher quality, and 2) promoting professional development.

Special Education Teacher Evaluation

With the current emphasis in educational policy on improving teacher effectiveness, states are rapidly developing and implementing new models and methods for teacher evaluation. However, these newly developed models fail to address the unique challenges related to measuring special education teacher effectiveness, and how it relates to student growth. For example, a recent forum sponsored by the National Comprehensive Center for Teacher Quality titled "Using Student Growth to Evaluate Educators of Students With Disabilities: Issues, Challenges, and Next Steps," the expert and researcher panel concluded that "to improve teacher practices and academic outcomes for students with disabilities, it is critical that we design evaluation systems that account for diverse teacher roles, student learning goals and trajectories, and assessment means (e.g., standardized, alternative, and formative)" (Holdheide et al., 2012, p. 1). This assembled group of researchers concluded that because of the limited research and the challenges involved with measuring the academic growth of students with disabilities, they caution against using student achievement until further research and practical experience can fully support the validity of claims made by proponents (Holdheide et al., 2012).

Furthermore, there are little to no teacher evaluation approaches that are specific to the unique needs of the special education classroom, nor are there any able to recognize the historical and current challenges facing the profession. Essentially, there is a significant gap in empirical support that is specific to measurement approaches of

special education teacher effectiveness (Holdheide et al., 2010; Prince et al., 2009). In this next section, the current state of special education will be reviewed, followed by a description of the challenges associated with special education teacher evaluation, and concluding with a review of the limitations of current approaches to teacher evaluation with special education.

Current State of Special Education

Students served through special education often have the most intense instructional needs, and require specially designed instruction (Gersten et al., 1997; Vannest, Hagan-Burke, Parker, & Soares, 2011; Wehmeyer & Field, 2007). Meeting the multiple and varying needs of students with disabilities is challenging, highly-technical, and requires teachers who have strong instructional skills (Feng & Sass, 2010; Odom, 2009). Unfortunately, students with disabilities are more often served by a special education teaching force that is highly subject to attrition, turnover, and burnout; historically, special education has been characterized by high attrition rates (Billingsley, 2004; Boe et al., 2008; Holdheide et al., 2010; Sindelar, Brownell, & Billingsley, 2010), job dissatisfaction (Gersten et al., 2001; Stempien & Loeb, 2002), and personnel who are not fully certified or certified through alternate routes (Littrell, Billingsley, & Cross, 1994; McLeskey, Tyler, & Flippin, 2004). These factors lead to a profession chronically faced with teacher shortages, as evidenced by surveys in which more than 95% of all U.S. school districts reported at least one teaching vacancy in the field of special education at the beginning of the 1999-2000 school year (Connelly & Graham, 2009). Given the increase in students receiving special education of over 30% in the past decade (Connelly & Graham, 2009), this crisis continues to get worse. The combination of these challenges

have contributed to what researchers call the "substandard quality of education for students with special needs" (Gersten et al., 2001).

Special education is consistently indicated as a high-demand field, with positions filled by teachers who lack adequate preparation to meet the demands of the position (Boe et al., 2008). Even when special education teachers enter the classroom with adequate pre-service training, actual instruction time is consumed by multiple duties like case management, testing, progress monitoring, paperwork, meetings, and management of support staff (Russ et al., 2001; Santoro, 2011; Vannest & Hagan-Burke, 2009). Recent estimates suggest that as little as 20% of a special education teacher's time is dedicated to instruction (Vannest & Hagan-Burke, 2009). As a result, this lack of instructional time impacts student outcomes: as few as 30% of students with disabilities nationally are able to meet performance standards (Odom, 2009) and post-school outcomes for students with disabilities are not encouraging (Newman et al., 2011). Young adults with disabilities are less likely to have enrolled in postsecondary programs than their peers in the general population, as well as less likely to complete 4-year degrees, make less per hour, and are less likely to live independently (Newman et al., 2011).

To improve the outcomes for students with disabilities, the instructional practice of special education teachers must be improved (McLeskey, 2011; Morgan, Frisco, Farkas, & Hibel, 2008; Nougaret, Scruggs, & Mastropieri, 2005; Scruggs, Mastropieri, Berkeley, & Graetz, 2009). Promisingly, the field of special education research has a strong foundational knowledge base on the use and application of evidence-based instructional practices that can be utilized to improve the current state of the profession

(Baker et al., 2009; Cook & Odom, 2013; Gersten et al., 2009; Odom, 2009; Odom et al., 2005; Smith et al., 2013). Evidence-informed instructional practices produces better outcomes for students with disabilities, and in order to reap these benefits, implementation of these practices must be systemized (Fixsen, Blase, Metz, & Van Dyke, 2013; McLeskey, 2011).

Unique Challenges to Special Education Teacher Evaluation

Because of the historical and current difficulties facing the special education profession, as well as the highly-defined roles and responsibilities that characterize special education teaching, there are unique challenges facing special education teacher evaluation. And, because the ultimate goal of any teacher evaluation system is to improve student outcomes, students with disabilities have the most to gain (and lose) in the development of a fair and effective special education teacher evaluation system. The teaching context and individualized nature of special education pose the two primary challenges to evaluating special education teacher effectiveness.

Variety of Special Education Teaching Contexts

Special education teachers serve approximately 12% of the student population nationally (Council for Exceptional Children, 2012). Yet within this 12% student population, there is a significant amount of heterogeneity in the kinds of settings in which students with disabilities are served. Special education teachers may work in collaboration with a general education teacher in the classroom. Alternatively, they might run a resource room, in which students are pulled out from their general classroom to receive specialized instruction. For students with more significant needs, special

education teachers may provide instruction in self-contained or extended resource rooms (Browder & Cooper-Duffy, 2003; Wehmeyer, Palmer, Shogren, Williams-Diehm, & Soukup, 2010). Special education teachers may work in a consultant role, providing support to teachers to include students with special needs in the general classroom. Not only does the role of the special education teacher vary across settings, but in smaller districts with fewer resources, one special education teacher may find herself filling a number of these roles (Moore, 2012). The heterogeneity in special education settings requires a flexible approach to evaluation (Semmelroth et al., in press).

The "Technical Science" of Individualized Instruction

One of the requirements for being diagnosed as a student with a learning disability is that the student requires specially designed instruction. The instructional strategies that are appropriate to meet the needs of students with disabilities vary based on disability type, content area, and grade level. Special education instruction is not just a complex and variable profession but a technical science (Odom et al., 2005), requiring strong analytic skills as well as the ability to stay current on evidence-based instructional practices for a heterogeneous population. Students served in special education reflect a very heterogeneous and diverse population (Tyler, Yzquierdo, Lopez-Reyna, & Flippin, 2004), and defining student achievement through one universal measure, or even through a set of accepted predetermined measures, poses methodological problems (Baker et al., 2010). Even when students present with similar needs, they may function at vastly different performance levels (Karvonen et al., 2012). It is difficult to say that one type of student is just like another type of student if placed in the same classroom or determined eligible under the same exceptionality. While this is arguably true of all students, for

students with disabilities this is especially the case. Depending on their baseline performance, their opportunities to learn, and the severity of their disability, students with disabilities will experience different growth rates and consequently meet very different outcome targets. As a result, any effective special education teacher evaluation system will need to be able to account for these challenges.

Limitations of Current Teacher Evaluation Approaches

Previously in this literature review, a discussion of the two most commonly used methods to evaluate teachers, classroom observations and performance metrics, was provided to outline some of the primary advantages and disadvantages of each approach. In this next section, a review of how these two approaches are limited in the special education teacher evaluation context will be discussed, starting with classroom observations and followed by performance metrics.

Limitations of Classroom Observations for Special Education Teacher Evaluation

As previously mentioned, Charlotte Danielson's (2007) *Framework for Teaching (FFT)* is organized around four domains of teaching responsibility: planning and preparation, classroom environment, instruction, and professional responsibilities. Domain 3, the instructional domain, is based in a constructivist approach, which is not in alignment with the evidence-based practices typically used to meet the needs of students with disabilities (Odom, 2009; Roberts et al., 2008; Spooner et al., 2012). Therefore, the use of FFT to evaluate special education instruction could lead to an evaluation that is not aligned with the research base and that endorses practices that do not lead to improved outcomes for students with disabilities.

In addition, research suggests that content specific observation tools are found to have positive effects on student outcomes. For example, ongoing studies on the Mathematical Quality of Instruction (MQI) have found that "there is a powerful relationship between what a teacher knows, how she knows it, and what she can do in the context of instruction" (Hill et al., 2008, p. 496). Similarly, work on the Classroom Observations of Student–Teacher Interactions (COSTI), a special education observation tool that evaluates a teacher's interaction with students as a measurement of effectiveness, suggests that content-specific tools may be beneficial (Smolkowski & Gunn, 2012). In fact, even Charlotte Danielson has released a 2013 edition of the FFT observation tool that is more sensitive to the challenges found in the special education setting, as well as incorporating some of the instructional implications of the upcoming Common Core State Standards (Danielson, 2013; Elliott, 2012).

Lastly, through large-scale studies like the Measures of Effective Teaching (MET) project, research suggests that observer reliability is unstable unless optimized with both multiple observations and multiple raters (Kane & Cantrell, 2013; Kane & Staiger, 2012). These findings suggest that as states revise education policy to incorporate multiple-methods teacher evaluation systems, current practices of one to two formative classroom observations by the building administrator may need to be reconsidered (Kane & Cantrell, 2013; Mihaly et al., 2013). This finding is especially compelling for the field of special education as there can be significant diversity in special education teacher roles, responsibility, and specialized instructional practice, which a building administrator may or may not be sensitive.

Limitations of Performance Metrics for Special Education Teacher Evaluation

Probably the biggest criticism of the use of performance metric-based teacher evaluation systems to evaluate special education teachers is that they fail to deliver a mechanism to provide specific, targeted feedback regarding instructional practice. Given the roles and responsibilities special education teachers have, as well as the challenges facing the profession, it is important that an effective special education teacher evaluation system is able to: 1) bridge the research-to-practice gap, and 2) provide targeted, specific feedback to improve practice (Cook & Odom, 2013b; Feng & Sass, 2010; Foegen et al., 2001; Gersten & Smith-Johnson, 2001; Goe, Biggers, & Croft, 2012; Greenwood, Horton, & Utley, 2002; Kretlow & Bartholomew, 2010; McLeskey, 2011; Smith et al., 2013).

There are also other limitations of performance metrics for special education teacher evaluation. First, the number of special education students with standardized assessment data are too few to be used in quantitative analyses (of which, the standardized assessment data are already faced with measurement issues related to modifications, accommodations, etc.) (Braun, 2012; Lohr, 2012). Second, given the range of special education teacher roles and responsibilities (spanning across grades, content areas, academic areas, etc.), defining one, primary role of a teacher's "effect" in a performance metric is difficult (Holdheide et al., 2012). Some districts are experimenting with allocation of time as a way to parse the "value added" by each teacher, but these approaches are flawed because time does not directly translate into the intensity of the instruction in special education. For example, it is difficult to determine the impact that a 20-minute instructional session in reading has on a student's performance in social

studies or science. These questions regarding the allocation of time can also be found in other nontested, noncore content areas (Prince et al., 2009).

Third, students served through special education reflect a very heterogeneous and diverse population (Tyler et al., 2004), and defining student achievement through one assessment measure, that can vary based on a student's classification, poses additional challenges in analyses (Baker et al., 2010). Because the empirical and theoretical work on performance metrics has not included special education, a research-based model or approach for special educators within this type of framework does not exist (Braun, 2012; Floden, 2012; Hanushek & Rivkin, 2010a; Holdheide et al., 2010; Kane & Cantrell, 2013; Lohr, 2012).

Other issues related to the lack of empirical evidence on performance metrics for special education teacher evaluation are related to the measurement questions of special education student growth like: 1) What is a reasonable rate of growth for students with disabilities? 2) What is the impact of testing accommodations on student performance? 3) What is the impact of test difficulty on student performance? and 4) What are the longitudinal characteristics of the population of students with disabilities (Buzick & Laitusis, 2010; Karvonen et al., 2012; van den Heuvel et al., 2012)?

Therefore, in order to develop a teacher evaluation system that effectively meets the diverse needs found in special education, it must be able to account for the current challenges found in the profession and in the variety of classrooms. In the next section, a rationale for this proposed study is provided, based on a discussion of the reliability issues related to the research and development of the pilot RESET observation tool.

Rationale for Study

The Recognizing Effective Special Education Teacher (RESET) project was established to: 1) create a definition of special education teacher effectiveness, and 2) develop a tool to measure special education teacher effectiveness, using student outcomes as a primary source of measurement. The RESET project is funded by a two-year (2011-2013) grant from the Idaho State Department of Education and is located in the Department of Special Education at Boise State University. Through RESET project work, an effective special education teacher is defined as someone who is *able to identify a student's needs, implement evidence-based instructional practices and interventions, and demonstrate student growth*. This definition has been developed on the premise that instructional practice is a crucial component of promoting a student's individual growth. This premise is grounded in over three decades of empirical research that establishes a number of effective instructional practices to meet the needs of students with disabilities.

Measuring Sources of Variance

The RESET observation tool evaluates the use of evidence-based instructional practices in an observed lesson to measure special education teaching effectiveness. The tool is flexible enough to be used across multiple special education settings, but specific enough to provide targeted feedback for teachers. The pilot RESET observation tool is still in early stages of development and additional studies will be required before it is ready to be used in practice. Future studies will include establishing levels of validity to predict student outcomes based on evaluation of the observed teacher.

For this study, current research efforts on the tool were focused on identifying levels of variance across facets: raters, occasions, and teachers (generalizability study),

and determining the optimal conditions of these sources of variance to minimize error (decision study). Likened as an alternative to classical test score theory, generalizability theory allows for simultaneous examinations of multiples sources of rater variance (Brennan, 2001; Cronbach et al., 1972; Shavelson & Webb, 1991). Generalizability theory specifies the level of error that can be accounted for by various situational variables that were present when the measurements were taken (Tindal, Yovanoff, & Geller, 2010).

Previous studies on the pilot RESET observation tool found low to weak levels of agreements across raters using perfect agreement and kappa to measure observer agreement (Johnson & Semmelroth, 2012). However, interrater agreement measures like kappa can be problematic and misaligned with what is supposed to be measured, because reported levels of interrater reliability can be low even though observer agreement is high (Cronbach et al., 1972, p. 190). For example, when events (in a mutually exclusive and exhaustive set) have highly unequal baserates, values of kappa will be lower than when baserates are equal, even when observers are highly accurate (Bruckner & Yoder, 2006, p. 435). Furthermore, strong rater agreement on an observation tool can be misleading because: 1) rater agreement levels can be influenced by the number of points on a rating scale, 2) the frequency of target behaviors in classroom teaching can affect observed and expected counts, and 3) the occurrence of chance agreement can skew outcomes (Feinstein & Cicchetti, 1990; Hill, Charalambous, & Kraft, 2012). Most importantly, measures of straight rater agreement attends to only one source of variation (the rater) leaving other sources of variation (e.g., teachers, occasions, items, etc.) that affect the consistency of evaluation scores within observations (Brennan, 2001; Cronbach et al.,

1972; Hill, Charalambous, & Kraft, 2012; Shavelson & Webb, 1991). If only the observed score between raters is considered, meaningful information is lost that may have been influential in the determination of that observed score (Tindal et al., 2010). In determining rater agreement to assess observed performances (e.g., teacher observations) traditional views of reliability maintain that observed scores comprise of just two components, 'true score' and 'error,' without any way to distinguish between the variance that makes up these two components (Brennan, 2001). Thus, a single score obtained on one occasion is not fully dependable (Shavelson & Webb, 1991), making the case for the use of generalizability theory to analyze multiple sources of variance in a measurement.

Generalizability Theory

From an earlier study on the pilot RESET observation tool, rater data were analyzed to examine interrater reliability and identify the main sources of variance, using perfect agreement and kappa analyses (Johnson & Semmelroth, 2012). Results indicated weak to no agreement for many parts of the RESET observation tool, and sources of variability were not readily identified using perfect agreement and kappa analyses (Johnson & Semmelroth, 2012).

However, researchers have documented that multiple sources of variance in observational scores can be due to the number of observed lessons, differences among raters, varying characteristics of the observational instrument, and variability of the teacher's own performance over time (Erlich & Shavelson, 1978; Goe et al., 2008; Hill, Charalambous, Blazar, et al., 2012; Ho & Kane, 2013; Kane & Staiger, 2012; Newton et al., 2010; Seidel & Shavelson, 2007; Shavelson & Dempsey-Atwood, 1976; Shavelson & Dempsey, 1975). While perfect agreement and kappa analyses are used to measure rater

agreement and reliability for classroom and teacher observations, generalizability and decision studies are available to account for variability that traditional interrater agreement analyses cannot (Cronbach et al., 1972; Hill, Charalambous, Blazar, et al., 2012; Tindal et al., 2010). In fact, recent studies similar to this one (i.e. MET and MQI) have used generalizability theory to estimate sources of error, and to optimize the reliability of different 'real-world' scenarios by varying the number and type of raters and the number and length of lessons (occasions) (Hill, Charalambous, & Kraft, 2012; Ho & Kane, 2013; Kane & Cantrell, 2013).

Because this study sought to identify sources of variance and minimize measurement error, generalizability and decision studies were used instead to analyze rater data. Generalizability theory is considered to be an extension of classical test theory through an application of analysis of variance (ANOVA) procedures to measurement (Brennan, 1992, 2001; Cronbach et al., 1972; Shavelson & Webb, 1991). Generalizability theory "liberalizes classical theory by employing ANOVA methods that allow an investigator to untangle multiple sources of error" (Brennan, 2001, p. 3). While classical theory notes $X = T + E$ (where X is the observed score, T is the true score, and E is undifferentiated error), generalizability theory allows for the exploration of multiple sources of error, $X = T + (E1 + E2 + E3\ldots)$ (Brennan, 2001). Thus, at the simplest level, classical theory is too limited in an analysis of sources of variance because it assumes only one source of error, despite that in reality there are many different definitions of what this error looks like (Brennan, 2001). Generalizability theory goes beyond the application of variance components analysis to measurement issues, as it also informs which components contribute to which types of error (Brennan, 2001, p. 19). For these

reasons, generalizability theory was used in this study because it shifts measures of reliability on the RESET observation tool from one that is restricted by limited views of interrater agreement to one that accounts for comprehensive sources of variance.

In a generalizability study, or G study, an *observation* is described in terms of *conditions* (the setting, the observer, the task, etc.), and the general term for referring to conditions of a certain kind is called a *facet* (Cronbach et al., 1972). The facets, alone or in combinations, define *universes*, and it is from these universes that holistic scores and generalizations are produced. A facet in generalizability theory is considered to be synonymous with a factor in ANOVA (Cardinet, Johnson, & Pini, 2010). Each facet, or source of variance, can be decomposed and analyzed for its score effects (see Table 1), and then analyzed for optimized conditions (i.e., decision study or D study, discussed in the next section).

Facets can be defined as "fixed," "random," or "finite random" based on their sampling status. A facet is considered to be "fixed" when all levels are featured in the data set (i.e., no sampling of levels have occurred). A facet is considered to be "random" when the levels included in the analyses are randomly selected from the respective population or universe. A facet is considered to be "finite random," also known as "mixed," when random sampling can be conducted within a finite universe (Cardinet et al., 2010). A design is considered to be "complete" and "balanced" when all possible interactions have been considered (complete) and all facets included have the same number of items (balanced) (Brennan, 2001). The decision to use a complete and balanced dataset minimizes overall error variance (i.e., missing data is not calculated into measurements of error variance), but the tradeoff for this means that the data set size can

sometimes be smaller (in order to maximize the minimum requirements) (Brennan, 2001).

This study's design included two facets (raters, occasions), and one unit of measurement (teachers) to analyze sources of variance. All facets included in this study were determined to be random. Previous studies have established these sources of variance as primary facets in the study of the influences of teacher behavior on student achievement (Cronbach et al., 1972; Erlich & Shavelson, 1978; Hill, Charalambous, & Kraft, 2012). Like the studies completed by Erlich and Shavelson (1978), and Hill, Charalambous, and Kraft (2012), this study aims to identify the generalizability of measures of teacher behavior by systematically examining the effect of more than one facet (raters, occasions). A complete and balanced study design was used.

As seen in Table 1-1, a two-facet (raters, occasions), crossed design has six other sources of variability. A facet is considered to be "crossed" when every level of one of the facets is combined with every other in a data set (Cardinet et al., 2010). These sources of variability are associated with each of the measurement facets in generalizing from the sample of instructional practice (occasions) (from the video observation dataset to be used in this study) in the measurement of the universe of occasions on each teacher using the pilot RESET observation tool.

Table 2-1. Sources of Variability in the Two-Facet Observational with a t x r x o Crossed Design Measurement

Source of Variability	Type of Variation	Variance Notation
Teachers (*t*)	Universe-score variance (object of measurement)	σ^2_t
Raters (*r*)	Constant effect for all teachers due to stringency of raters	σ^2_r
Occasions (*o*)	Constant effect for all teachers due to their behavioral inconsistencies form one occasion to another	σ^2_o
t x *r*	Inconsistencies of raters' evaluation of particular teachers' behavior	σ^2_{tr}
t x *o*	Inconsistencies from one occasion to another of particular teachers' behavior	σ^2_{to}
r x *o*	Constant effect for all teachers due to differences in raters' stringency from one occasion to another	σ^2_{ro}
t x *r* x *o, e*	Residual consisting of the unique combination of t, r, o; unmeasured facets that affect the measurement; and/or random events	$\sigma^2_{tro,\,e}$

*Adapted from Shavelson & Webb, 1991, p. 9

The strength of the G study is that multiple sources of error in a measurement can be estimated separately into a single analysis (Shavelson & Webb, 1991). As seen in Table 2-1, a two-facet design allows for the range of different conditions found within a teacher evaluation. However, the two-facet design does *not* account for differences that can occur between observations for occasions and teachers; there can be multiple

occasions per teacher and the occasions can differ from teacher to teacher (Shavelson & Webb, 1991).

Thus, for this study, a different design was used where occasions were nested within teachers (as opposed to being crossed) because teachers are not expected to teach exactly the same lessons. Facets are considered to be "nested" if each level of one is associated with one and only one level of the other (Cardinet et al., 2010). In generalizability studies, nested facets are defined in the same way as in ANOVAs (Brennan, 2001; Shavelson & Webb, 1991).

In this study, a two-facet, nested design was used where occasions (*o*) (observations/lessons) were nested within teachers (*t*), *o:t*, and crossed with raters (*r*), *{o:t} x r*. Although nested facets can reduce the scope of the universe of generalization of the results, the nested "occasion" facet helps to reduce overall error variance, while also staying true to the purpose of the analysis. Table 2-2 presents the various components for this type of study.

Table 2-2. Sources of Variability in the Two-Facet Nested Design {o:t} x r

Source of Variability	Type of Variation	Variance Notation
Teachers (*t*)	Universe-score variance (object of measurement); amount of systemic variability between teachers in their instructional practice	σ^2_t
Raters (*r*)	Variance component that measures how much variability raters see over teachers and occasions	σ^2_r

Source of Variability	Type of Variation	Variance Notation
Occasions (*o:t*)	Nested variance component that measures how much variability teachers differ from one occasion to another	$\sigma^2_{o,\,to}$
tr	Variance component that measures the relative standing of teachers from one rater to another	σ^2_{tr}
o:pr, e	Residual due to confounded sources of variation	$\sigma^2_{ro,\,tro,\,e}$

*Adapted from Shavelson & Webb, 1991, p. 54

Alternative Study Design

It is important to point out that data analyses for this study could have been completed using a *three*-facet, partially nested design where individual questions from the RESET tool are kept in tact as a separate facet, Items (i). This type of design would have been {o:t} x i x r , where occasions are nested facets crossed with items and raters, and teachers remain the unit of measurement. Although it is recommended that any given data set should be "maximally exploited" so that "as many facets as possible should be identified for exploration in the analysis," this exploitation of identified facets must also be constrained within data balance (equal cell sizes), data quantity (too few observations for a facet or facet interaction will lead to unstable estimation), and software limitations (Cardinet et al., 2010, p. 40). Because of the already relative small sample sizes of this study, and given the structure of the RESET tool (that allows for raters to identify and define instructional components within each video), the data set would have been considerably constrained by issues related to data balance and data quantity. For these reasons, it was determined that the present data set would be too unstable (too small) for

a three-facet design. Instead, specific items from the tool were selected for analysis, and were combined into separate, purposeful "subscales." A more detailed explanation of each subscale is included in the following "Use of Subscales in G-Studies" section.

Additionally, expanding on the idea that the purpose of generalizability theory is to "obtain estimates of a variance components associated with a universe of admissible observations" (Brennan, 2001, p. 8), and that data sets should be "maximally exploited" for exploration in generalizability study analyses (Cardinet et al., 2010, p. 40), this study approached data analyses as diversely as possible. As discussed, this study's design included two facets (raters, occasions) and one unit of measurement (teachers) to analyze sources of variance, and used rater data collected from two separate data coding sessions (October 2012 and April 2013). However, the combined data Oct/April set was also defined multiple ways to "exploit" explorations of sources of variance. A more comprehensive description of these data sets will follow in Chapter 3 Research Method.

Generalizability Coefficient

In relation to issues of reliability, generalizability theory allows an analysis to generalize from sample to universe. Cronbach et al. (1972) explain, "the question of 'reliability' thus resolves into a question of accuracy of generalization, or generalizability" (p. 15), known as the generalizability coefficient or G coefficient. Another reliability-like coefficient is the index of dependability, or dependability coefficient (Brennan, 2001). Both of these coefficients (G coefficient and index of dependability) are defined as the ratio of universe score variance to itself, but they differ in the addition of variance: the G coefficient adds the relative error variance, while the index of dependability adds the absolute error variance (Brennan, 2001). The program

used in this study, EduG, provides only for the relative G coefficient and the absolute G coefficient; however, this is still sufficient for conducting generalizability study analyses (Brennan, 2001; Shavelson & Webb, 1991).

Broadly speaking, relative reliability corresponds to the G coefficient (Shavelson & Webb, 1991), but whereas Cronbach's alpha (α) measurement error is attributable to one source of variance, the G coefficient accounts for multiple sources of error variance that can be acknowledged and accommodated (Cardinet et al., 2010). The G coefficient is analogous to the reliability coefficient in classical theory and has a range of zero to one (Hendrickson & Yin, 2010). Estimates of the G coefficient for different numbers of raters and occasions rely on an extension of the Spearman-Brown prophecy formula, which considers only one facet of error affecting the measurement (Erlich & Shavelson, 1978, p. 78).

Conceptually speaking, the G coefficient of relative measurement indicates how well a measurement procedure has differentiated among objects of study (i.e., how well the procedure has ranked objects on a measuring scale) and where the objects concerned might be students, patients, teaching methods, training, etc. On the other hand, the G coefficient of absolute measurement indicates how well a measurement procedure has located objects of study on a scale, regardless of where the other objects might be placed (Cardinet et al., 2010, p. 6). Typically a G coefficient of absolute measurement will have lower values than the relative value because there are more potential sources of error variance (Cardinet et al., 2010). Because generalizability theory allows each observation to belong to a multitude of possible sets of observations, a test is no longer determined to be reliable or unreliable. Instead, G theory allows one to simply generalize to different

degrees from one observed score to the multiple means of the different sets of possible observations (Cardinet, Tourneur, & Allal, 1976; Cronbach et al., 1972).

Although different researchers strive to maintain specific levels of G coefficient cut scores, there is no agreed upon scale or range. For example, in the G study analyses conducted in the Measures of Effective Teaching (MET) study, Ho and Kane (2013) presented different ways (i.e., facet conditions) to ensure reliabilities of .65 or above, while Cardinet et al. (2010) and Shavelson and Webb (1991) more consistently adhere to the .80 rule to evaluate the preciseness of a measurement. Still others maintain that the reliability of the entire measurement procedure must be considered (facets, study design, unit of measurement) when interpreting the results of G and D study analyses (Brennan & Lee, 2013; Cronbach et al., 1972) as generalizability theory is much more than just the application of variance components analysis to measurement issues (Brennan, 2001).

Thus, in addition to analyses of generalizability coefficients, this study also reports the standard errors of measurement (SEMs). The absolute error variance scores (the difference between a person's observed and universe score) are reported in this study to help provide a deeper examination using generalizability theory, because as Brennan (2001) reminds, "it can be very misleading to refer to the reliability or the error variance of a measurement procedure without considerable explanation and qualification" (p. 17).

Decision Study

In this study, a G study was used to decompose levels of variance associated with the use of the pilot RESET observation tool, using three facets: teachers, raters, and occasions (lessons/observations). Following the G study analyses, the decision study procedure, or D study, was completed to identify the optimal amount of facet conditions

to achieve the lowest levels of measurement error when using the pilot RESET observation tool. Looking at it another way, the D study assists in answering the question, "How many observations and raters are needed to obtain the minimal amount of error when evaluating teachers using the pilot RESET observation tool?"

Although considered to be two separate types of analyses, G and D studies are complementary when exploring sources of error. If G studies help identify the sources of error (or variance), then D studies explore conditions to optimally minimize these sources of error. The G study and D study are often conducted in sequence: "Often, generalizability analyses may be viewed as two-stage processes. The goal of the first stage is to obtain estimated variance components for a G study design, given a universe of admissible observations. The second stage involves using these estimated variance components in the context of a D study design and universe of generalization to estimate quantities such as universe score variance, error variances, and coefficients" (Brennan, 2001, p. 53). While the G study analyzes a measurement for sources of variance, the D study uses information from the G study to optimize the analyzed facets for the least amount of error. Shavelson and Webb (1991) explain, "G studies estimate the magnitude of as many potential sources of measurement error as possible. D studies use information from a G study to design a measurement that minimizes error for a *particular* purpose. The G study is associated with the development of a measurement procedure, whereas the D study applies the procedure" (p. 83). In this study, D studies were completed to examine different conditions of occasions and raters to help identify acceptable levels of precision in the pilot RESET observation tool.

Use of Subscales in G Studies

To analyze collected data, this study used a two-facet, partially nested design ({o:t} x r) that included two facets (raters, occasions) and one unit of measurement (teachers) to analyze sources of variance. All raters evaluated all videos included in the analyses, and all scores were initially aggregated at the lesson level. Like Hill, Charalambous, and Kraft (2012) G-study measurement design using the Mathematical Quality of Instruction (MQI) observation tool, this G study measurement design was based on the view that most special education lessons classes feature purposeful differences in instructional methods as the teacher interacts with students through different phases of the lesson. That is, special education teachers will use different instructional methods not just between occasions, but within occasions themselves as they strive to meet the instructional strengths and needs of a particular group of students. For example, although one component may feature the use of explicit, direct instruction, the second component later in the lesson may intentionally feature a different type of instruction. This type of approach to evaluating a special education teacher's effectiveness (via instruction) makes it difficult to conduct direct crosses comparisons across raters within one video observation; rater disagreement can occur not just between evaluative rubric ratings, but in the determination of when one instructional component begins and ends, as well as what type of instructional practice is being used by the observed teacher. For these reasons, collected data must further be aggregated past the lesson level in order to conduct G study and D study analyses using a two-facet, partially nested design. Individually rated items must be collapsed into purposeful subscales so that collections of rater scores comprise just one facet (rater), and so that broad analyses

can be made across all collected data. Hill et al. (2012) used a similar approach when analyzing rater data from the MQI, but use the term "dimensions" instead of "subscales."

To create the subscales used in the analyses in this study, items were grouped according to evaluative purposes: Subscale 1: Lesson Objective, Subscale 2: Evidence-Based Instructional Components, and Subscale 3: Evaluative Summary. Data from the October 2012 and April 2013 data coding sessions were combined to create each subscale. Given that the RESET tool is grounded in Danielson's *Framework for Teaching* evaluative framework, when applicable, all questions appearing in the RESET tool include the same rubric scale, i.e. a qualitatively defined rating scale from 0 to 3. Subscales were created by collapsing relevant questions into a holistic score (all items in the RESET tool align with the same evaluative rubric scale). Again, just as Hill, Charalambous, and Kraft (2012) "averaged" scores across "dimensions," this study collapsed relevant items into subscales.

Subscale 1: Lesson Objective

Subscale 1 is only comprised of one question between both data sets, and all raters had to answer this question for all observations included in the study. Although three additional questions related to the lesson objective were added to the April 2013 version of the RESET observation tool, these could not be used in the G studies because they were not included in the October 2012 session (and thus there are no rater data). The question included in Subscale 1 is directly related to the lesson objective for component #1 and asks "Is component #1 objective aligned with the larger lesson objective?" and lists three possible answers: "Yes," "Partially," "No/Inconclusive." Appendix A includes the evaluative rubric for Subscale 1.

Subscale 2: Evidence-Based Instructional Components

Subscale 2 is directly related to the characteristics of specific evidence-based practices, as well as how observed teachers in the October 2012 and April 2013 video data sets implemented these practices. Recalling back to the structure of the RESET observation tool, and the variable use of rater-defined instructional components in each observation, it is especially clear for subscale 2 why direct comparisons of each rater's observation is not practical, nor does it yield a large enough data set. This is because raters individually determine: 1) when an instructional component begins and ends, and 2) what type of practice was used. Thus, in order to maintain a data set large enough to conduct G and D studies, the evaluated evidence-based instructional components for Component #1 were collapsed into one subscale score. For example, if Teacher 1 was evaluated by 5 raters that all indicated that the evidence-based instructional practice "explicit, direct instruction" (comprised of four components: organized instruction, sequenced instruction, scaffolding, student practice, and review) was used, then each rater's score for each instructional component (i.e., four components per rater) was collapsed into one holistic score. However, if Teacher 2 was evaluated by 5 raters, and 4 of those raters indicated that the evidence-based instructional practice "explicit, direct instruction" was used, but 1 rater identified "whole group instruction" (comprised of: individualized instruction, skill development, student engagement, and feedback and assessment) instead, the rated components would still be collapsed into one holistic score by rater. The rationale for standardizing Subscale 2 across different practices is based on two important reasons: 1) each evidence-based practice is comprised of four, discrete components that while they may be separate from one another by definition are not very

different in terms of purpose (e.g. “student practice and review” versus “feedback and assessment”) and 2) all components are evaluated on the same rubric (aligned with Danielson’s *Framework for Teaching*). Appendix B includes the evaluative rubric for Subscale 2, which includes the four components for each one of the instructional practices that appear in the RESET observation tool (12 total).

Subscale 3: Evaluative Summary

Subscale 3 is a broad look at the “big” questions included at the end of the evaluation, and like Subscale 1, all raters had to answer these questions for all observations in October 2012 and April 2013. Subscale 3 is comprised from the four ‘big’ questions that are related to broad, evaluative determinations of the observed teacher’s lesson. These four questions are: “Is the use of time effective for the lesson's learning objective?” “Does the teacher appear to have a solid understanding of the content/curriculum?” “Does the teacher implement effective instructional practices?” and “Does the teacher effectively respond to student needs?” As with Subscale 2, all four of these questions were collapsed into one holistic score because all evaluations were completed on the same rubric. Appendix C includes the evaluative rubrics for Subscale 3.

Thus, this study continued development of a pilot special education observation tool (RESET) by identifying sources and levels of variance using generalizability theory to analyze rater data. The rationale for use of this type of analysis is that traditional measurements of observer agreement to define interrater reliability are too limited in its scope, and they do not account for other sources of variance and measurement error. Instead, generalizability studies were used because it allowed for identification of sources of variances and error, followed by decision studies to determine the strongest levels of

reliability in optimal observation conditions (using raters, teachers, and occasions as the facets of measurement) (Brennan, 2001; Cronbach et al., 1972; Shavelson & Webb, 1991).

Summary and Conclusion

Improving teacher quality has become the focus of the contemporary school reform movement (McGuinn, 2012). This effort is dependent on the development of new teacher-evaluation systems with multiple measures of performance rooted in student achievement that can provide reliable data around levels of teacher effectiveness and quality. Classroom observation and performance metric approaches to teacher evaluation face increasing scrutiny as the stakes are raised higher for teacher effectiveness, especially when considering these changes in the context of special education (Holdheide et al., 2012; Prince et al., 2009; *The council for exceptional children's position on special education teacher evaluation*, 2012). While performance metrics can provide useful information about a teacher's performance in comparison to others, they do not provide targeted, specific feedback and there are still many unanswered measurement questions regarding both tested and nontested student groups. Similarly, classroom observations may provide opportunities for feedback, but it is not specific to special education, and recent studies suggest problems achieving and maintaining reliability. Not only do the multiple roles of the special educator cause problems for current teacher evaluation approaches, but the state of the special education profession complicates the issue as well. Issues associated with the special education profession include high levels of attrition, vacancies, and turnover; a lack of highly qualified teachers in core content

areas; and the ability of special education teacher preparation programs to adequately prepare new teachers to meet the challenges found in the classroom and the profession.

In order for a special education teacher evaluation system to ensure teacher quality, promote professional development, and improve outcomes for students with disabilities, it must reliably and consistently discriminate between effective and ineffective special education teachers; provide targeted, specific, corrective feedback for teacher instructional practice; include the use of individualized student growth rates to define teacher effectiveness; and adapt to the variety of contexts in which special education teachers work. Currently, there is no teacher evaluation system that comprehensively and holistically accounts for these specific requirements. Thus, as a first step to this call for a special education teacher evaluation system that ensures teacher quality and improves outcomes for students with disabilities, there is a need for further development of a pilot special education observation tool. The next section of this paper includes the results of the G study and D study analyses using rater data from the pilot RESET observation tool to evaluate special education classroom instruction.

CHAPTER 3: RESEARCH METHOD

Introduction

This study applied generalizability theory to identify sources of variance on a pilot observation tool designed to evaluate special education teacher effectiveness. In this study, the pilot Recognizing Effective Special Education Teachers (RESET) observation tool included three evidence-based instructional practices (direct, explicit instruction; whole-group instruction; discrete trial teaching) as the basis for special education teacher evaluation. Eight teacher coders (raters) were invited to attend two sessions (October 2012 and April 2013) to evaluate special education classroom instruction collected from the 2011-2012 and 2012-2013 school years, via the Teachscape 360-degree video system. The raters were trained on the pilot RESET observation tool, and participated in whole-group coding sessions to establish interrater agreement before evaluating assigned videos.

Data collected from raters were analyzed using generalizability theory in a two-facet "partially" nested design (Shavelson & Webb, 1991, p. 52). Generalizability study analyses were used because they are useful for understanding the relative importance of various sources of error to assist in the design of more efficient procedures (Brennan, 1992; Shavelson & Webb, 1991), and because teacher evaluation systems are complex, traditional approaches to establishing reliability, such as interrater reliability, do not adequately inform the design of these tools. Using the results from the generalizability

study analyses, decision studies were then completed to determine optimal facet conditions for the strongest levels of practical reliability.

Research Design

Due to the questions asked in this proposed study, quantitative methods were used to analyze data and discuss results. Recall that the primary question guiding this generalizability and decision study analysis was: *How many occasions and raters are needed for acceptable levels of reliability when using the pilot RESET observation tool to evaluate special education teachers?* This research study was designed to determine the sources of variance affecting reliability across raters on the pilot RESET observation tool. Additionally, this study was designed to identify different levels of reliability across content subscales on the pilot RESET observation tool using generalizability study analyses. Finally, this study analyzed sources of variance using decision studies to determine optimal conditions for reliability using the pilot RESET observation tool (i.e., the number of raters needed per lesson, and the number of lessons per teacher required to achieve the most practical levels of reliability).

Participants and Setting

Participants

Eight special education teachers were invited to participate as data coders for this study. Previous studies and generalizability theory explanations have established that smaller rater sample sizes are sufficient for research and teacher development purposes (Erlich & Shavelson, 1978; Hill, Charalambous, & Kraft, 2012; Shavelson & Webb, 1991). The teachers were selected through their participation with other university

research projects and/or identified through their district special education directors. Although the sample was one of relative convenience for this study, pre-determined criteria were observed to ensure that invited raters represented: 1) a balanced sample of the range of content, placement, and grade found in special education, and 2) that the invited raters have all completed a minimum of 5 years of certified teaching (i.e., newly certified and/or special education teachers on alternative authorizations were not invited to participate). Raters were financially compensated for their time ($500/session). All participating raters successfully completed the Collaborative Institutional Training Initiative (CITI) program in alignment with Institutional Review Board (IRB) requirements.

Table 3-1 presents a summary of rater demographics including current teaching assignment, total years teaching, and highest level of education completed. All raters were female (except Rater 7), and all raters worked in urban districts (except Raters 1 and 6). All eight raters who participated in the April 2013 session have participated in at least one previous (June 2012 or October 2012), and two of the raters (Raters 1 and 5) have participated in all three sessions.

Table 3-1. April 2013 and October 2012 Data Coding Rater Demographics, n=8 raters

Raters	Current Teaching Assignment	Years Teaching (total)	Highest Level of Education Completed
Rater 1	**Elementary EBD	30	Graduate Certificate
Rater 2	Elementary Resource	15	Bachelors
Rater 3	Elementary Resource	5	Bachelors
Rater 4	Elementary Resource/ University Adjunct	5	Masters
Rater 5	Secondary Resource	12	Masters
Rater 6	**Secondary ERR	10	Bachelors
Rater 7	*Secondary Resource	8	Masters

Rater 8	Secondary Resource	5	Masters

*male, **rural district

Although there initially were eight raters involved between the two data coding sessions, the analyses conducted in this study were reduced to five raters. The loss of two raters was due to last-minute circumstances (i.e., the April 2013 data coding session lost raters 2 and 6). Additionally, rater 7 was not able to participate in the April 2013 session but in order to keep this rater's data from the October 2012 session, a "replacement" rater was trained (Rater 8) to substitute for the April 2013 session. Therefore, the combined October 2012 and April 2013 data set experienced a loss of three raters (Raters 2, 6 and 7), leaving a total of n=5. Because of this loss, the combined October 2012 and April 2013 data sets were defined in two different ways: 1) one with the "complete" set of October 2012 and April 2013 data, n=5, and 2) one with the combined set of October 2012 and April 2013 data, with the Rater 7/8 omitted, n=4.

Setting

For both the October 2012 and April 2013 data coding sessions, raters were hosted on the Boise State University campus with all arrangements (food, parking, room, etc.) provided through the RESET grant project. The sessions were designed to protect the confidentiality of the teachers appearing in the video observation data. Training and data coding sessions were held in a reserved room on campus, which was only accessible to those participating in the project. Raters were seated away from one another, and were given headphones to wear throughout the sessions to prevent any sharing of rater visual or audio information.

Raters evaluated video observation data that was collected via the Teachscape Reflect system, the same technology used by the Measures of Effective Teaching (MET) study funded by the Bill and Melinda Gates Foundation (Kane & Staiger, 2012). The Teachscape video capture system consists of two cameras: 1) a 360-degree camera that allows the observer to pan and zoom on various components of the classroom environment and 2) a fixed position camera, also referred to as a "board cam" because it is usually focused on a classroom board (see: Appendix D for a screenshot of what a user sees when viewing a processed Teachscape video capture). Raters only had access to these videos while on campus during the session, and upon completion of the session, each rater's Teachscape accounts were deleted, preventing any outside access to the video observation data.

Video Data Collection

Video data for both the October 2012 and April 2013 data coding sessions were collected across five school districts from 25 different teachers over the course of two school years (2011-2012, 2012-2013). Data collection efforts were completely dependent upon district, school, and classroom access, and while some districts gave permission to conduct research, most teachers within each district opted not to participate. For example, in one of the larger school districts that agreed to participate in the study, out of roughly 200 special education teachers, only 3 agreed to participate in the study. In this way, establishing trusting, collaborative working relationships was a critical part of the data collection process. The exception to this was found in District 4, where teachers had a much higher rate of participation than the other much larger districts (see: Table 3-2). From the 2011-2012 school year, a minimum of three observations each were collected

from a total 12 different teachers, and 10 of these teachers were eventually included in the October 2012 data coding session. Similarly, from the 2012-2013 school year, six observations each were collected from 13 different teachers, and nine teachers were included in the April 2013 session. Teachers were excluded from the data coding sessions because either there were too many unusable video captures, or because the teachers utilized classroom instructional practices that go beyond the current capabilities of the pilot RESET tool. The amount of captures assigned to each teacher changed from the 2011-2012 (minimum of three) to 2012-2013 (at least six required) school year because rater agreement measures shifted from simple interrater agreements to this current study's use of generalizability theory. The mean time of each video was 25 minutes, with videos in the data set ranging from 72 minutes to 17 minutes.

Table 3-2. Video Data: Distribution of Teachers Across Five Districts, n=25 teachers

School Year	District 1	District 2	District 3	District 4**	District 5 **
2011-2012	5	2	--	5	--
2012-2013	3	1	2	6	1

**rural districts

Rater Training

Each rater was provided with two university-owned laptops for use: one to watch the assigned Teachscape videos and one to complete the observation tool. I was available throughout both three-day coding sessions to answer questions and provide assistance to raters. For each session, raters were provided with a half-day training presentation, followed by individually evaluating two, separate videos for the purposes of calibration and measuring interrater reliability. A 45-page user manual was provided to explain the structure and features of the pilot tool. The manual also includes operationalized

definitions and descriptions of the three evidence-based instructional practices (direct, explicit instruction; whole-group instruction; discrete trial teaching,), and the evaluation rubrics for all ratings on the pilot RESET observation tool (see: Appendices A-C for an example of each practice's component rubrics).

During the training, raters were oriented through the user manual and a blank pilot RESET observation tool. Raters were presented with the theoretical framework of the tool, followed by a walk through of the specific components of the evaluation rubrics and evidence-based instructional components. Following the presentation portion of the training session, the first video was viewed, which raters evaluated individually using the pilot tool. The rater scores from video #1 were reviewed and then compared for agreement as a whole group activity against the master ratings (predetermined by myself and the RESET Project Director). Following the first training video and whole-group discussion, a second video was viewed and the scores across the raters were again reviewed and compared for agreement against the master ratings as a whole group training activity. Although formal measures were not in place to evaluate rater agreement for the October 2012 session (besides the individual items discussed within the whole group), the April 2013 session formally measured rater agreement scores for the two training videos. (This discrepancy in rater agreement measurement procedures is due to improvements in training between the two sessions.) Table 3-3 includes the results of the interrater agreement from the April 2013 session, organized by total agreement as well as by agreement by each of the three subscales.

As can be seen from Table 3-3, the total level of agreement increased from .78 (video #1) to .82 (video #2). The rater agreement for Subscale 1 remained consistently

low at .50 for both videos, but this is partially due to a small sample size (2 questions with 5 raters each), and because the April 2013 version of the tool introduced three new questions related to lesson objectives. Raters who had previously been trained on the October 2012 and June 2012 versions of the tool expressed confusion how to answer the old and new questions related to the lesson objective. However, after clarifying evaluative criteria after the second video, all five raters confirmed understanding of the questions included in Subscale 1. Subscale 2, which evaluates specific components of evidence-based instructional practice had strong levels of agreement, at 1 (video #1) and .90 (video #2). It is hypothesized that the agreement for video #2 decreased by .10 because the technical complexity of that video was significantly higher than video #1. Lastly, Subscale 3, which are the summative "big" questions about a teacher's overall instructional effectiveness and practice, increased in agreement from .67 (video #1) to .85 (video #2).

Table 3-3. Results of Interrater Agreement Compared Against Master Ratings from April 2013 Training, n=5 raters

	Total	Subscale 1	Subscale 2	Subscale 3
Video #1	.78	.50	1	.67
Video #2	.82	.50	.90	.85

Measures

For this study, the pilot RESET observation tool was used to evaluate the special education instructional practice of teachers using trained raters during two, separate data coding sessions. The pilot RESET observation tool evaluates a teacher's ability to deliver evidence-based instructional practices that align with content and grade-level practices, and as a result, adjusts to different placements, classrooms, grades, and exceptionalities (Johnson & Semmelroth, 2012). The pilot RESET observation tool used in this study

includes three, evidence-based instructional practices: 1) direct, explicit instruction, 2) whole-group instruction, and 3) discrete trial teaching.

As mentioned in Chapter 1, this study is part of a larger project to develop and validate a special education teacher observation measure, the pilot *Recognizing Effective Special Education Teachers* (RESET) tool, designed to: 1) evaluate evidence-based instructional practice, 2) provide targeted, specific feedback to special education teachers about the quality of their instruction, and 3) improve the outcomes for students with disabilities. The pilot RESET observation tool is designed to address three important issues in the field of special education: 1) close the research-to-practice gap on special education instructional practice, 2) improve special education teacher quality, and 3) improve the outcomes for students with disabilities. The conceptual framework guiding RESET is that effective special education teachers implement relevant (appropriate to population, context, and content) evidence-based instructional practices with fidelity in order to improve student outcomes.

Based on a theory of effective special education teaching that *an effective special education teacher is able to identify a student's needs, implement evidence-based instructional practices and interventions, and demonstrate student growth*, the RESET pilot observation tool has been designed to measure a special education teacher's use of evidence-based instructional practice and the resulting effect on student outcomes (Johnson & Semmelroth, 2012).

The pilot RESET observation tool focuses on the primary responsibility of teacher practice (i.e., instruction). The RESET tool includes evaluation criteria separated into the core components of evidence-based instructional practice, so that teachers can be

provided direct feedback on their ability to implement evidence-based practices to support student outcomes. The tool consists of three parts: the Lesson Objective (introduction), specific Lesson Components (evidence-based instructional practices), and the Lesson Summary (conclusion). The pilot RESET observation tool uses a four-point Likert scale (0-3) that is in alignment with Danielson's (2007) evaluation rubrics of: Unsatisfactory, Basic, Proficient, and Distinguished, as well as Danielson's (2013) most recently revised rubric of numerical ratings (levels 1-4).

The research on instructional practice in special education includes over four decades worth of research on a variety of instructional strategies designed to meet the needs of various disability types. Several meta-analyses of instructional practice have been undertaken over the years in special education and provide helpful starting points for explicating the key elements of an instructional strategy (Bellini et al., 2007; Berkeley et al., 2009; Dexter et al., 2011; Gersten et al., 2009; Swanson, Lee, Sachse-Lee, 2000). From these meta-analyses, common definitions of different instructional practices were developed, along with specifying the particular elements that are essential to define the practice. In addition to providing guidance on instructional characteristics, meta-analyses also provide data on a range of effect sizes that will assist future project work to determine the expected outcomes for students with disabilities when specific instructional strategies are used.

Data Collection

Rater data included for this study was collected from five raters who used the pilot RESET observation tool to evaluate video observations of special education teachers during two different sessions in October 2012 and April 2013. The pilot RESET observation tool was accessible online via the university-owned Qualtrics system. The Teachscape video capture system was used to collect video observations from nine special education teachers from five school districts across southern Idaho during the 2011-2012 and 2012-2013 school years. A minimum of three lessons (mean time = 25 minutes) from each teacher was captured. Upon completion of the data collection session, the data were exported into Excel, and organized for analysis using the EduG 6.1 generalizability theory software system.

Data Analysis

For this study, a two-facet, partially nested design was used (Figure 1): the object of measurement was teachers (t) and the facets were raters (r), and occasions (o).

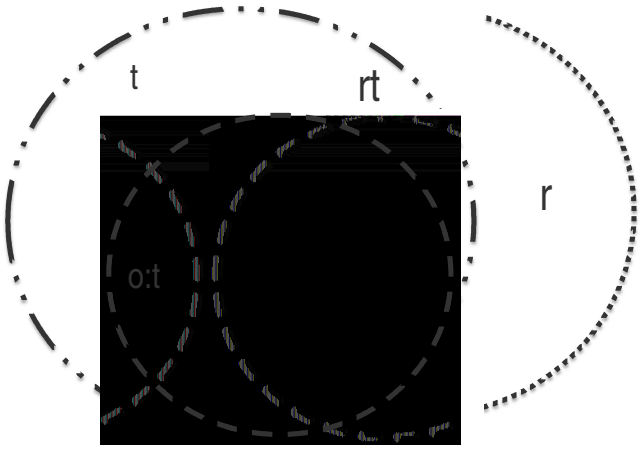

Figure 3-1. Generalizability Theory Two-Facet Nested Design Using Teachers (t) as the Object of Measurement, and Raters (r) and Occasions (observations) (o) as Facets, {o:t} x r

In this model, occasions (observations) were nested within teachers and crossed with raters{o:t} x r (Shavelson & Webb, 1991, pp. 52–54). Like Erlich and Shavelson's (1976) study of teacher behavior, in this study, different teachers (t) were observed on different occasions (o), but all raters (r) observed all teachers on all occasions. Five raters used the pilot RESET observation tool to evaluate three videos each from nine different teachers.

Decision studies were conducted to determine optimal facet conditions between raters and occasions to reduce the most amount of error (and thus increase the total amount of precision) with the pilot RESET observation tool.

Data Set Differentiation

Because the April 2013 portion of this study's data collection experienced unavoidable attrition, the data was analyzed using two differently defined data sets.

The first data set, "Data Set A," is considered to be fully complete with nine teachers (t), three occasions (o), and **five** raters (r). The second data set, "Data Set B," is considered to be missing because the Rater 7/8 combination from October 2012 (rater 7) and April 2013 (rater 8), leaving nine teachers (t), three occasions (o), and **four** raters (r). Thus, the only difference between the Data Sets A and B is Rater 7/8, but because this rater actually consisted of two separate people, and given the small sample size of each facet, it was determined that the results should included analyses without this rater.

Observation Design

In this study, three facets were identified as part of the observation design: raters, teachers, and occasions. The observation design information (i.e., facet identification, and

the numbers of levels observed) describes the structure of the data, and once defined, cannot be changed (Cardinet et al., 2010). In principle, any given data set should be maximally exploited (i.e., as many facets as possible should be identified for exploration in a G study analysis). However, the identification of facets must be considered within the constraints of data balance (equal cell sizes) and data quantity (too few observations for a facet will lead to unstable estimation).

These three facets (raters, teachers, and occasions) were used in the observation design, just as with similar studies that used the same approach including the Measures of Effective Teaching (MET) project (Ho & Kane, 2013; Kane & Cantrell, 2013), the Mathematical Quality of Instruction (MQI) G study (Hill, Charalambous, & Kraft, 2012), and illustrative analyses based on previous G studies to examine classroom and teacher characteristics (Cronbach et al., 1972, pp. 189–193).

Estimation Design

The size of each facet universe is determined in the estimation design and can be labeled as fixed or infinite random (random). For this study, all three facets were determined to be random as in previous, similar studies. This determination was also based on the assumption that the raters, teachers, and occasions used in this study were selected at random from an indefinitely large universe of raters, teachers, and occasions, or can be considered exchangeable with any of the other raters, teachers, and occasions in the universe (Shavelson & Webb, 1991).

Measurement Design

The measurement design distinguishes facets from those as the object of measurement (also referred to as the differentiation facet), and those that condition the measurement procedure (also referred to as the instrumentation facets). In this study, the differentiation facet was teachers (t), and the instrumentation facets were raters (r) and occasions (o).

Summary

This study continued development of a pilot special education observation tool by using generalizability theory to identify sources and levels of variance to increase measurement precision of the tool. Five raters were trained to use the pilot RESET observation tool to evaluate video observations of special education classroom instruction captured via the Teachscape system. Rater data was analyzed using the EduG v. 6.1 software program to: 1) complete generalizability study analyses to identify sources of variances, and 2) follow up with decision study analyses to determine the strongest levels of reliability in optimal observation conditions (using raters and occasions) (Brennan, 2001; Cronbach et al., 1972; Shavelson & Webb, 1991). A two-facet, partially nested design was used: occasions (lessons) was nested within teachers and crossed with raters {o:t} x r (Shavelson & Webb, 1991, pp. 52–54). Two data sets were created to account for the Rater 7/8 combination, and all generalizability study and decision study analyses were conducted on both data sets.

CHAPTER 4: RESULTS

Introduction

The purpose of this study was to identify sources of variance to increase measurement precision on a pilot observation tool designed to measure special education teacher effectiveness. The primary question guiding the generalizability and decision studies conducted in this paper was: *How many occasions and raters are needed for acceptable levels of reliability when using the pilot RESET observation tool to evaluate special education teachers?*

This study used rater data collected from two data coding sessions in October 2012 and April 2013. Raters evaluated special education classroom instruction video observation data collected during the 2011-2012 and 2012-2013 school years from five school districts located across southern Idaho. Generalizability and decision studies were completed to: 1) analyze the sources of variance, and 2) identify the optimal facet conditions for reliability and maximum precision. Tables 4-1 to 4-4 include the results of this generalizability study, organized by specific items from the RESET observation tool into three subscales: lesson objective (subscale 1), EBP implementation (subscale 2), and whole lesson review (subscale 3). Two data sets were included for analysis in this study: the "complete" data set that consisted of nine teachers (t), three occasions (o), and **five** raters (r) (Data Set A), and the "missing" data set that removed rater 7/8 leaving nine teachers (t), three occasions (o), and **four** raters (r) (Data Set B).

The results of these analyses are organized first by each data set, and then by each of the three subscales. When applicable, collapsed or condensed tables by subscale and/or data set are presented for comparative purposes. The results begin first with ANOVA tables, followed by the G study results, and conclude with D study scenarios organized by raters and occasions.

Sources of Variance

The overall purpose of this study was to identify sources of variance using generalizability theory so that further development and refinement on the pilot RESET observation tool can be made to increase overall precision. Because controls for measurement error and true score ratings are limited, considerable information is lost about rating scores when using traditional interrater reliability measures like kappa (Brennan, 2001; Cronbach et al., 1972; Shavelson & Webb, 1991). Instead, generalizability theory analyses parse rater variability owing to error into facets that are part of any measurement situation (Cronbach et al., 1972).

In this section, the results of the ANOVA analyses are presented (Tables 4-1 and 4-2) and are organized by each data set and then by each subscale, followed with a condensed table of just the variance decomposition for all three subscales for each data set (Table 8). Each ANOVA table is organized by a facet or a facet interaction (the source of variation) and includes the sums of squares (SS), degrees of freedom (df), mean squares (MS), percentage contribution of each source to the total variance (i.e., the sum of the corrected variance components (% of total variance)), and the standard error associated with each variance component (SE).

In Table 4-1, the ANOVA results for Data Set A for all three subscales are presented for each source of variation. The variance component for teachers (σ^2_t) shows the amount of systemic variability between teachers (the object of measurement) in their behavior. The variance component for teachers varies between subscales: 21.3% (lesson objective), 14.8% (EBP implementation), and 19.6% (whole lesson review). Because this source of variation represents the differentiation facet, ideally this number should be the highest source of variation. That is, the differentiation facet indicates the level of variation in the unit being measured (teachers), as opposed to another source that indicates a lack of precision with the RESET observation tool (e.g., residual) or the inconsistency of raters. Thus, variability is concentrated where it should be: teacher's instructional practice. For Data Set A, the teacher source of variation is only the second highest for subscale 1 (second to the residual score). For the other subscales, this source of variation is not the strongest source of variation.

Table 4-1. ANOVA for Data Set A, Subscales 1-3

ANOVA for Data Set A, Subscale 1: Lesson Objective					
Source of Variation	SS	df	MS	% of Total Variance	SE
Teachers (t)	39.978	8	14.8	21.3	0.113
Occasions:Teachers (o:t)	25.000	27	14.9	9.5	0.051
Raters (r)	5.633	4	15	1.8	0.023
Teachers x Raters (t x r)	29.967	32	24.5	9.4	0.054
Residual [(o:t) x r, e]	55.000	108	30.8	58	0.069
Total	155.578	179		100%	

ANOVA for Data Set A, Subscale 2: EBP Implementation					
Source of Variation	SS	df	MS	% of Total Variance	SE
Teachers (t)	222.500	8	27.813	14.8	0.633
Occasions:Teachers (o:t)	158.500	27	5.870	14.9	0.312
Raters (r)	148.967	4	37.242	15	0.599
Teachers x Raters (t x r)	230.333	32	7.198	24.5	0.440
Residual [(o:t) x r, e]	185.500	108	1.718	30.8	0.232
Total	945.800	179		100%	
ANOVA for Data Set A, Subscale 3: Whole Lesson Review					
Source of Variation	SS	df	MS	% of Total Variance	SE
Teachers (t)	321.400	8	40.175	19.6	0.908
Occasions:Teachers (o:t)	180.750	27	6.694	14.1	0.355
Raters (r)	191.033	4	47.758	15.9	0.768
Teachers x Raters (t x r)	266.767	32	8.336	23.6	0.509
Residual [(o:t) x r, e]	199.000	108	1.843	26.8	0.248
Total	1158.950	179		100%	

The variance component for raters (σ^2_r) indicates how much raters differed amongst themselves in the behavior they "saw," averaging over teachers and occasions (Shavelson & Webb, 1991, p. 54). The variance component for raters varies between subscales, 1.8% (lesson objective), 15% (EBP implementation), and 15.9% (whole lesson review) of the total variance for each subscale. Because this source of variation represents the instrumentation facet that has direct control over both the reliability and preciseness of the pilot RESET observation tool, ideally this number should be one of the lowest for each ANOVA subscale analysis. In fact, any source of variation that includes raters (r) should be low as this is the measurement related to how well raters "behave" using the tool. Again, as with teachers (t), this result does well in subscale 1, but for the other two

subscales, this source of variation falls in the middle compared to other sources of variation.

The variance component for occasions is nested within teachers ($\sigma^2_{o,\,to}$), which makes it impossible to separate the occasion main effect from the interaction between teachers and occasions. The variance component for occasions nested within teachers is 9.5% (lesson objective), 14.9% (EBP implementation), and 14.1% (whole lesson review) of the total variance for each subscale. However, because this facet is nested, it is not known whether one occasion produced more behavior than another (occasion main effect), whether the relative standing of teachers differed from one occasion to another (teacher-by-occasion interaction), or both (Shavelson & Webb, 1991).

The variance component for the interaction between teachers and raters (σ^2_{tr}) indicates the relative standing of teachers in terms of how they differed from one rater to another. The variance component for the interaction between teachers and raters is 9.4% (lesson objective), 24.5% (EBP implementation), and 23.6% (whole lesson review) of the total variance for each subscale. As with σ^2_r, because this source of variation includes the instrumentation facet that has partial influence over both the reliability and preciseness of the pilot RESET observation tool (r), it is important that this number be one of the lowest for each ANOVA subscale analysis. And again, as with σ^2_r, this result does well in subscale 1, but for the other two subscales, this source of variation is one of the highest compared to other sources of variation.

Lastly, the interaction between raters and occassions, the three-way intreaction between teachers, raters, and occasions, and unaccounted/unmeasured variation are confounded in this two-facet, partially nested design. The residual component ($\sigma^2_{ro,\,tro,\,e}$)

indicates that for subscale 1, 58% of the total variance is due to these substantial confounded sources of variation. However, for subscales 2 and 3, only 30.8% and 26.8% (respectively) are due to confounded sources of variation, indicating that the other facets do a better job of explaining variance in subscales 2 and 3 than in subscale 1.

In Table 4-2, the ANOVA results for all three subscales for Data Set B are presented for each source of variation. The variance components for teachers (σ^2_t) are 21.4% (lesson objective), 12% (EBP implementation), and 19.2% (whole lesson review). Like Data Set A, σ^2_t is only the second highest (after the residual score) for subscale 1. The variance components for raters (σ^2_r) are 2.9% (lesson objective), 21% (EBP implementation), and 22.3% (whole lesson review) of the total variance for each subscale.

As previously mentioned, in this study, sources of variation that include the rater (r) facet are important because they are directly related to both the reliability and precision of the pilot RESET observation tool. As with Data Set A, this result does well in subscale 1 as the lowest source of variance, but for the other two subscales, it remains in the middle.

Table 4-2. ANOVA for Data Set B, Subscales 1-3

ANOVA for Data Set B, Subscale 1: Lesson Objective					
Source of Variation	SS	df	MS	% of Total Variance	SE
Teachers (t)	33.375	8	4.172	21.4	0.118
Occasions:Teachers (o:t)	19.125	27	0.708	5.2	0.051
Raters (r)	5.611	3	1.870	2.9	0.034
Teachers x Raters (t x r)	22.514	24	0.938	11.7	0.068
Residual [(o:t) x r, e]	42.375	81	0.523	58.8	0.081
Total	123.000	143		100%	

ANOVA for Data Set B, Subscale 2: EBP Implementation					
Source of Variation	SS	df	MS	% of Total Variance	SE
Teachers (t)	183.639	8	22.955	12	0.661
Occasions:Teachers (o:t)	153.438	27	5.683	16.5	0.379
Raters (r)	159.854	3	53.285	21	0.938
Teachers x Raters (t x r)	173.583	24	7.233	22.9	0.506
Residual [(o:t) x r, e]	135.813	81	1.677	27.6	0.260
Total	806.326	143		100%	
ANOVA for Data Set B, Subscale 3: Whole Lesson Review					
Source of Variation	SS	df	MS	% of Total Variance	SE
Teachers (t)	291.750	8	36.469	19.2	1.035
Occasions:Teachers (o:t)	161.938	27	5.998	14.4	0.399
Raters (r)	209.910	3	69.970	22.3	1.231
Teachers x Raters (t x r)	207.528	24	8.647	23.1	0.603
Residual [(o:t) x r, e]	129.813	81	1.603	21	0.249
Total	1000.938	143		100%	

The variance component for occasions ($\sigma^2_{o,\,to}$), are 5.2% (lesson objective), 16.5% (EBP implementation), and 14.4% (whole lesson review). The variance components for the interaction between teachers and raters are 11.7% (lesson objective), 22.9% (EBP implementation), and 23.1% (whole lesson review). As with Data Set A, this component holds up well in subscale 1, but for the other two subscales, it is one of the highest compared to other sources of variation. Lastly, the residual component ($\sigma^2_{ro,\,tro,\,e}$) variance scores are 11.7% (lesson objective), 27.6% (EBP implementation), and 21% (whole lesson review).

For ease of comparison, Table 4-3 presents just the percent of total variance results for all three subscales and both data sets. While some of the differences between

total variance scores for each data set vary very little, it is interesting to note some of the larger differences within the two data sets in Table 4-3.

Table 4-3. Variance Decomposition for RESET Subscales, Datasets A and B

Source of Variation (%)	Lesson Objective		EBP Implementation		Whole Lesson Review	
	A	B	A	B	A	B
Teachers (t)	21.3	21.4	14.8	12	19.6	19.2
Occasions:Teachers (o:t)	9.5	5.2	14.9	16.5	14.1	14.4
Raters (r)	1.8	2.9	15	21	15.9	22.3
Teachers x Raters (t x r)	9.4	11.7	24.5	22.9	23.6	23.1
Residual [(o:t) x r, e]	58	58.8	30.8	27.6	26.8	21
Total	100	100	100	100	100	100

Overall, the trend between the two data sets can be found in the consistent increases and decreases in the sources of variation. For σ^2_t, although slight, the variance decreases in Data Set B across subscales 2 and 3, but for σ^2_r, the variance increases across the three subscales. For $\sigma^2_{ro,\ tro,\ e}$, the variance decreases in Data Set B across subscales 2 and 3. Because the primary difference between these two data sets is the amount of raters (five vs. four), these trends across subscales suggest that as raters increase, the less residual variance is produced (i.e., the larger the facet sample sizes, the more accurate the measurements). Interpretations of these results will be explored in more detail in the Chapter 5: Summary, Recommendations, and Conclusion.

G Study Results

Based on rater data from the October 2012 and April 2013 data coding sessions using the pilot RESET observation tool as the measure, a G study was conducted to analyze sources of error. The G study was completed to determine the variance

components attributable to teachers (t), occasions (o), and raters (r); their two-way interactions; and the combination of the three-way interaction and the measurement error.

Table 4-4. Generalizability Study Error Variance and G Coefficients for Pilot RESET Observation Tool, Data Sets A and B

Source of Variation (% Absolute)	Lesson Objective		EBP Implementation		Whole Lesson Review	
	A	B	A	B	A	B
Occasions:Teachers (o:t)	31.5	15.1	28.3	24.5	27.6	22.1
Raters (r)	4.8	8.4	22.7	31.3	24.9	34.3
Teachers x Raters (t x r)	25.2	33.8	37.3	34	37	35.5
Residual [(o:t) x r, e]	38.5	42.6	11.7	10.3	10.5	8.1
Total Differentiation Variance (t)	0.187	0.191	0.823	0.732	1.349	1.464
Total Relative Error Variance	0.063	0.070	0.567	0.702	0.659	0.815
Standard Deviation	0.43	.13	.91	.91	1.16	1.05
Relative G-Coefficient	0.75	0.73	0.59	0.51	0.67	0.64
Absolute G-Coefficient	0.74	0.71	0.53	0.42	0.61	0.54

As with the ANOVA analyses, items from the pilot RESET observation tool were collapsed into three subscales and compared against two data sets (A and B). In Table 4-4, the results of the G study are reported including the: source of variation (% absolute), total differentiation variance, standard deviation, total relative error variance, relative G coefficient, and absolute G coefficient. The % absolute source of variation reports how the absolute error variance is distributed among the other sources; the information from this result indicates the sources of variance that have the greatest negative effect on the precision of the pilot RESET observation tool (Cardinet et al., 2010, p. 52). Additionally,

these results also inform the design of the follow-up D study as it indicates which facet contributes the most to measurement error.

For Data Set A subscale 1, the occasions (31.5%) and residual (38.5%) facets are the two largest contributors to measurement error, but for Data Set B, subscale 1, the teachers by raters interaction (33.8%) and residual (42.6%) are the largest contributors to measurement error. Additionally, the raters (r) facet decreases in its contribution to measurement error with the rater sample size increase, 8.4% (Data Set B) to 4.8% (Data Set A). This might suggest that increasing the number of raters from four (Data Set B) to five (Data Set A) helps to increase measurement precision. This same pattern can be found across the other two subscales for the raters facet as both subscales have almost a 10-point difference between Data Set A, subscale 2 (22.7%) to Data Set B, subscale 2 (31.3%) and Data Set A, subscale 3 (24.9%) to Data Set B, subscale 3 (34.3%).

The occasions (o:t) facet had a significant decrease as a contributor to measurement error from Data Set A subscale 1 (31.5%) to Data Set B subscale 1 (15.1%), and a less substantial difference for the other two subscales: Data Set A subscale 2 (28.3%) to Data Set B subscale 2 (24.5%), and Data Set A subscale 3 (27.6%) to Data Set B subscale 3 (22.1%). Because this is a nested facet, like the ANOVA source of variance, it is not known whether one occasion produced more behavior than another (occasion main effect), whether the relative standing of teachers differed from one occasion to another (teacher-by-occasion interaction), or both. Regardless, the (o:t) facet contributes a significant amount of error to warrant further exploration of conditions in a follow up decision study (next section).

The differentiation and relative error variances provide insight into whether a weak G coefficient (relative or absolute) is due to high measurement error, or just to minimal differences between the objects measured (Cardinet et al., 2010). These measurements provide a holistic indication of the reliability of the measurement procedure and give a general indication of each of the measurements' precision. As reviewed earlier, there is no agreed upon 'cut-off' score for what might be considered a strong level of reliability vs. a weak level of reliability. For example, Ho and Kane (2013) described a range of different scenarios to achieve reliabilities of .65 or higher in classroom observations, while Cardinet et al. (2010) consider a sample measurement of .78 as "not entirely satisfactory" (p. 53). Thus, as Brennan (2001) maintains, in order to really understand the value of a G coefficient, one must know the level of variance, what is most contributing to error, and to what extent these influences have in a given sample size.

Across all three subscales between each data set, both the relative and absolute G coefficients have lower values for Data Set B than Data Set A. This finding suggests that the rater facet sample size has a considerable influence in the precision of the measurements. Additionally, the coefficients for subscales 2 and 3 might be affected by the differentiation variance (t), with high values reported in both data sets. Because the difference between the differentiation variance and the relative error variance, the lower G coefficient values might be attributable to either measurement error, or minimal differences between the objects measured (Cardinet et al., 2010). Although the reported G coefficients might initially be interpreted as less than desirable, the values do suggest that the measurement was not entirely inadequate, and that with a few modifications to

facet sample sizes, more desirable levels of reliability might be obtained (Brennan, 2001; Cardinet et al., 2010; Hill, Charalambous, & Kraft, 2012; Ho & Kane, 2013; Shavelson & Dempsey, 1975; Shavelson & Webb, 1991; Webb, Shavelson, & Haertel, 2006).

D Study Results

The D study procedure allows for the "what if?" analyses that develop through the interpretation ANOVA and G study results. D studies use information from a G study to design a measurement to reduce error for a particular purpose. The relative G coefficient generally corresponds to higher scores and is recommended for use in relative decision-making (e.g., rewarding teachers for rated excellence), while the absolute G coefficient generally reports lower values and should be used for absolute decisions (e.g., firing teachers for rated unsatisfactory performance). For the D study procedures conducted in this paper, the relative G coefficient was recorded throughout the process of changing facet size characteristics.

Table 4-5 shows the relative G coefficient scores for both data sets across the three subscales. The rater and occasion facets were 'optimized' using different sample sizes to obtain 'optimal' levels of reliability (Cardinet et al., 2010). Table 4-6 reports the relative standard error of measurement scores for both data sets across the three subscales. Figures 4-1 to 4-6 are the graphical representations of the relative standard error of measurement and reliability across raters and occasions by each subscale for both Data Sets A and B. (The graphs for each data set look almost identical, as can be seen by the reported scores in Tables 4-5 and 4-6.)

Table 4-5. Relative G Coefficient for Decision Studies Comparing Occasions and Raters

Relative G-Coefficient	Lesson Objective		EBP Implementation		Whole Lesson Review	
	A	B	A	B	A	B
Occasion 1						
1 Rater	0.22	0.22	0.17	0.15	0.23	0.25
2 Raters	0.33	0.35	0.26	0.22	0.33	0.34
3 Raters	0.40	0.43	0.31	0.27	0.39	0.40
4 Raters	0.45	0.48	0.34	0.29	0.42	0.43
5 Raters	0.48	0.53	0.36	0.31	0.45	0.45
Occasion 2						
1 Rater	0.33	0.33	0.24	0.21	0.31	0.32
2 Raters	0.47	0.48	0.35	0.31	0.43	0.44
3 Raters	0.55	0.57	0.42	0.37	0.50	0.51
4 Raters	0.60	0.62	0.46	0.41	0.55	0.55
5 Raters	0.63	0.66*	0.49	0.44	0.58	0.58
Occasion 3						
1 Rater	0.40	0.39	0.27	0.24	0.35	0.35
2 Raters	0.55	0.55	0.4	0.36	0.48	0.49
3 Raters	0.62	0.64	0.47	0.43	0.58	0.56
4 Raters	0.67*	0.69*	0.52	0.47	0.60	0.61
5 Raters	0.70*	0.73*	0.55	0.50	0.64	0.64
Occasion 4						
1 Rater	0.45	0.44	0.29	0.26	0.37	0.38
2 Raters	0.60	0.60	0.43	0.39	0.51	0.52
3 Raters	0.67*	0.68*	0.50	0.46	0.59	0.60
4 Raters	0.72*	0.73*	0.56	0.51	0.64	0.64
5 Raters	0.75*	0.77*	0.59	0.54	0.67*	0.67*
Occasion 5						
1 Rater	0.48	0.47	0.3	0.28	0.38	0.39
2 Raters	0.63	0.63	0.45	0.41	0.53	0.54
3 Raters	0.70*	0.71*	0.53	0.49	0.61	0.62
4 Raters	0.75*	0.76*	0.58	0.54	0.66*	0.66*
5 Raters	0.78*	0.79*	0.62	0.57	0.69*	0.70*
Occasion 6						
1 Rater	0.51	0.49	0.31	0.29	0.39	0.40
2 Raters	0.66*	0.65*	0.46	0.42	0.54	0.55
3 Raters	0.73*	0.73*	0.54	0.50	0.63	0.63
4 Raters	0.77*	0.77*	0.60	0.56	0.68*	0.68*
5 Raters	0.80*	0.81*	0.64	0.60	0.71*	0.71*

* >= 0.65

Table 4-6. Relative Standard Error of Measurement (SEM) for Decision Studies Comparing Occasions and Raters

Relative SEM	Lesson Objective		EBP Implementation		Whole Lesson Review	
	A	B	A	B	A	B
Occasion 1						
1 Rater	0.82	0.82	1.98	2.02	2.11	2.11
2 Raters	0.62	0.60	1.54	1.60	1.65	1.67
3 Raters	0.53	0.51	1.36	1.42	1.45	1.49
4 Raters	0.48	0.45	1.27	1.33	1.36	1.39
5 Raters	0.45	0.41	1.20	1.27	1.29	1.33
Occasion 2						
1 Rater	0.62	0.62	1.62	1.65	1.74	1.76
2 Raters	0.46	0.45	1.24	1.27	1.33	1.35
3 Raters	0.39	0.38	1.08	1.12	1.15	1.18
4 Raters	0.36	0.34	0.99	1.03	1.06	1.09
5 Raters	0.33	0.31	0.93	0.97	1.00	1.03
Occasion 3						
1 Rater	0.52	0.54	1.49	1.51	1.60	1.63
2 Raters	0.39	0.39	1.12	1.14	1.20	1.23
3 Raters	0.33	0.33	0.96	0.99	1.03	1.06
4 Raters	0.30	0.29	0.87	0.91	0.94	0.97
5 Raters	0.28	0.27	0.82	0.85	0.88	0.91
Occasion 4						
1 Rater	0.48	0.50	1.42	1.43	1.53	1.56
2 Raters	0.36	0.36	1.05	1.07	1.13	1.16
3 Raters	0.30	0.30	0.90	0.92	0.97	1.00
4 Raters	0.27	0.26	0.81	0.84	0.87	0.9
5 Raters	0.25	0.24	0.75	0.78	0.81	0.84
Occasion 5						
1 Rater	0.45	0.47	1.37	1.39	1.48	1.52
2 Raters	0.33	0.34	1.01	1.03	1.09	1.12
3 Raters	0.28	0.28	0.86	0.88	0.93	0.96
4 Raters	0.25	0.25	0.77	0.79	0.83	0.86
5 Raters	0.23	0.23	0.71	0.74	0.77	0.8
Occasion 6						
1 Rater	0.43	0.45	1.34	1.35	1.45	1.49
2 Raters	0.31	0.32	0.98	1.00	1.06	1.09
3 Raters	0.26	0.27	0.83	0.85	0.90	0.93
4 Raters	0.24	0.24	0.74	0.76	0.80	0.83
5 Raters	0.22	0.21	0.69	0.71	0.74	0.77

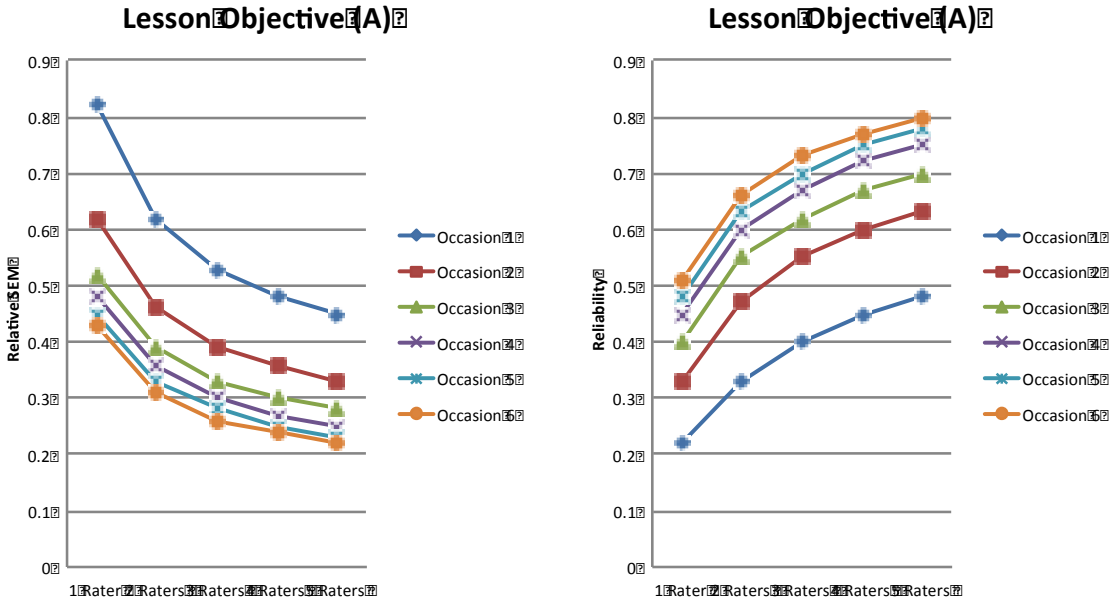

Figure 4-1. Data Set A, Lesson Objective D Study, Raters (r) and Occasions (o), SEM and G Coefficient

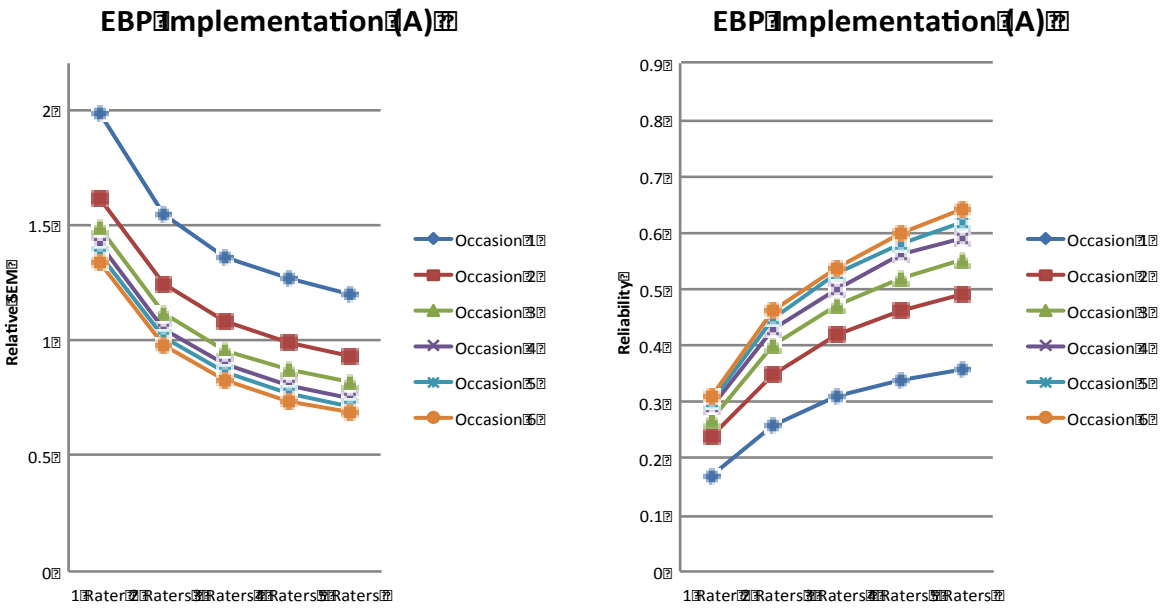

Figure 4-2. Data Set A, EBP Implementation D Study, Raters (r) and Occasions (o), SEM and G Coefficient

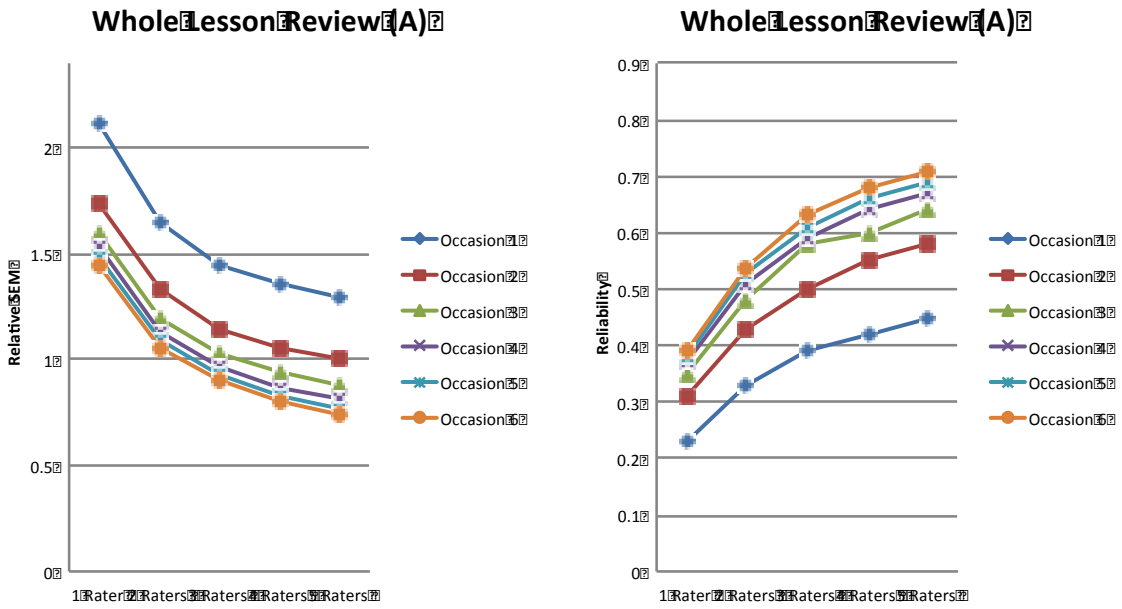

Figure 4-3. Data Set A, Whole Lesson Review D Study, Raters (r) and Occasions (o), SEM and G Coefficient

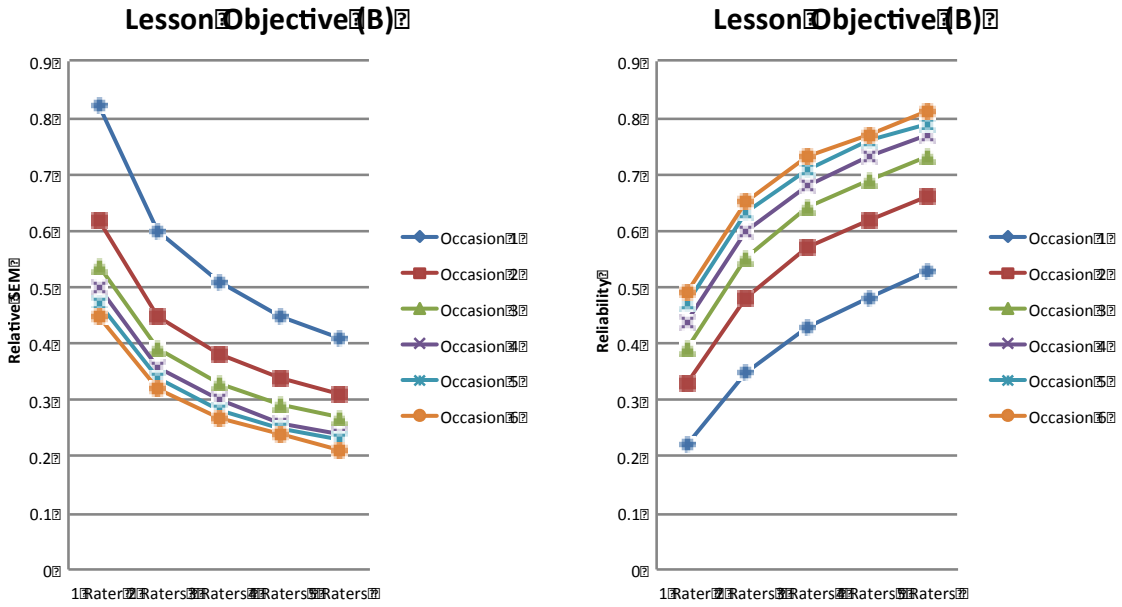

Figure 4-4. Data Set B, Lesson Objective D Study, Raters (r) and Occasions (o), SEM and G Coefficient

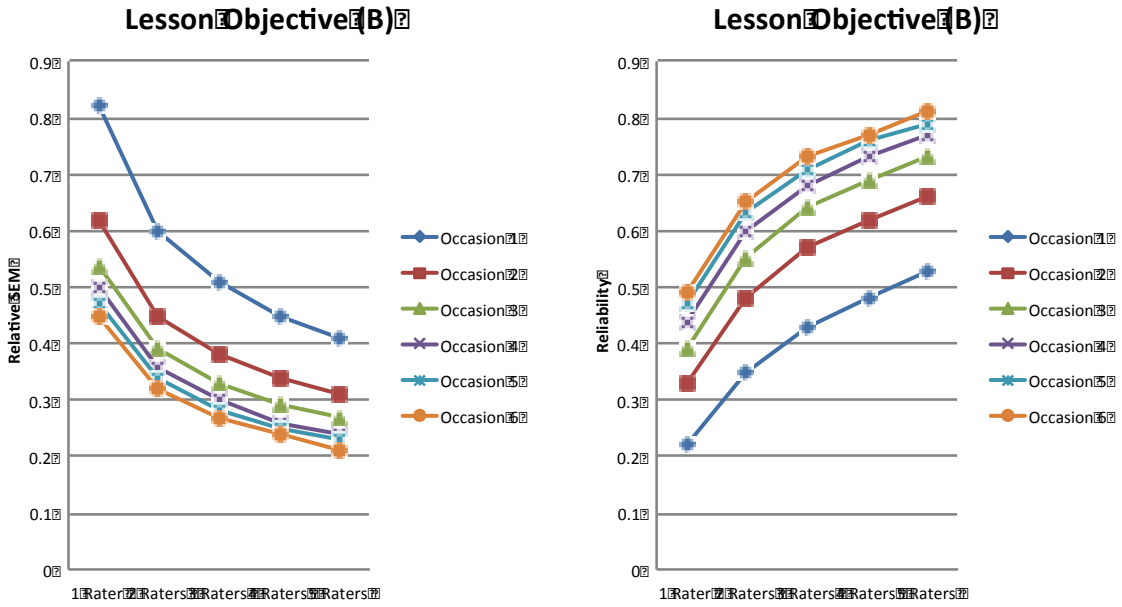

Figure 4-5. Data Set B, EBP Implementation D Study, Raters (r) and Occasions (o), SEM and G Coefficient

Figure 4-6. Data Set B, Whole Lesson Review D Study, Raters (r) and Occasions (o), SEM and G Coefficient

From Figures 4-1 to 4-6, it can be seen that as raters and occasions increase, so too does the relative G coefficient, while the SEM steadily decreases. For all three subscales and both data sets, though there are significant differences in reported values between Occasion 1 and Occasion 2, and somewhat between Occasion 2 and Occasion 3, the gaps between measurements are smaller between Occasions 3 and 6. This suggests that there might be a "happy medium" between empirical reliability and practical application somewhere along the continuum of multiple raters observing 2-4 occasions. Similarly, while there are significant differences for all three subscales from Rater 1 to Rater 3, the increase flattens out from Rater 3 to Rater 5. Like the differences between occasions, there seems to be a practical middle ground somewhere between between 2-4 raters. This finding suggests that real-life applications of the pilot RESET observation tool would not require ideal, research-like settings (e.g., 6-8 observations using 6-8 raters), but will be able to more practically consider finite resources.

In Table 4-5, the reported relative G-coefficient scores for facet conditions are presented, with scores at 0.65 or higher (Ho & Kane, 2013) indicated with an *. For both data sets A and B, subscale 1 indicates 0.65 and higher levels of reliability for three occasions with four raters at 0.67 (A) and 0.69 (B), and five raters at 0.70 (A) and 0.73 (B). However, for subscale 2, the corresponding scores are much lower for four raters, 0.52 (A) and 0.47 (B) and five raters 0.55 (A) and 0.50 (B), and almost equivalent for subscale 3 for four raters 0.60 (A) and 0.61 (B) and five raters, 0.64 (A) and 0.64 (B). Overall, subscale 1 reports higher levels of reliability with fewer occasions (starting with occasion 3) than the other two subscales. Subscale 2 consistently has lower scores than

the other two subscales, and does not report any coefficients higher than 0.65. Subscale 3 reports consistently stronger levels of reliability starting with four occasions.

In Table 4-6, the relative standard error of measurement (SEM) is reported for each of the D studies conducted for both data sets across the three subscales. The relative SEM corresponds to error variance found in classical test theory (Brennan, 2001), and is considered to be a critical piece of information when evaluating the measurement precision of a tool (Cardinet et al., 2010). When interpreting the output of a G study, it is the SEM that informs the user about the size of error affecting the results in the context of relative or absolute measurement (Cardinet et al., 2010; Cronbach et al., 1972). In effect, the SEM quantifies the precision, or lack thereof, of the measuring procedure (Cardinet et al., 2010). As can be seen in both Table 4-6 and Figures 4-1 to 4-6, the SEM steadily decreases as the raters and occasions increase. These results suggest that levels of precision on the RESET tool are much less reliable with fewer raters and occasions. These results also suggest that the level of error decreases as facet sizes increase (i.e., not only is there a steady decrease in SEM scores as the number of occasions and raters increase, there are also differences among subscales between the two data sets).

Summary

Overall, results from the generalizability and decision studies indicate that in order to increase reliability and decrease measurement error, multiple observations across multiple raters must be observed when using the pilot RESET observation tool. Across data sets, the ANOVA and generalizability study analyses reported different results, suggesting that the amount of raters in the sample size can make a difference in determining reliability as evidenced by reported levels of variance. Across subscales,

facets reported as the highest and lowest sources of variance for subscale 1, and subscales 2 and 3, suggest that there might be substantive differences between what each subscale is able to measure at this time. In the decision study analyses, as raters and occasions increase, levels of reliability correspondingly do as well, while the relative standard error of measurement decreases. The findings in this study are in alignment with similar generalizability theory studies on observation tools to measure teacher behavior; and, in order to achieve acceptable levels of relative reliability and error, multiple raters and occasions must be used (Bell et al., 2012; Erlich & Shavelson, 1978; Hill, Charalambous, & Kraft, 2012; Ho & Kane, 2013; Kane & Cantrell, 2013; Medley & Mitzel, 1958; Shavelson & Dempsey, 1975).

In the following chapter, additional discussion of the results is included, along with implications for special education teacher evaluation. Recommendations for future research are discussed, followed with the conclusion to this study.

CHAPTER 5: SUMMARY, RECOMMENDATIONS, AND CONCLUSION

Overview

The purpose of this study was to continue development of the pilot RESET observation tool by identifying sources and levels of variance using generalizability theory to analyze special education teacher evaluation data. From the results of the generalizability studies, decision study analyses were completed to identify optimal numbers of raters and occasions to maintain the highest levels of reliability when using the RESET tool. Results from this study were in alignment with similar studies on observation tools that found multiple observations across multiple raters are needed in order to achieve acceptable levels of reliability and minimum levels of error (Hill, Charalambous, & Kraft, 2012; Ho & Kane, 2013; Kane & Cantrell, 2013; Medley & Mitzel, 1958).

This study sought to answer the following questions:

1. What sources of variance affect reliability across raters on the pilot RESET observation tool?
2. When organized by content subscales, which part of the pilot RESET observation tool demonstrates the strongest and weakest levels of reliability?
3. What are the optimal observation conditions to maximize reliability using the RESET observation tool?

In order to answer these questions, generalizability theory was used to identify contributing sources variance and minimize the largest sources of error, with the ultimate goal of increasing the precision of the pilot RESET observation tool for future studies.

To answer these questions, generalizability theory was used to analyze rater data from the pilot RESET observation tool. From the ANOVA and G study analyses, sources of variance were reported, and using the data from the G studies, decision studies were conducted to identify the optimal levels of reliability to inform future applications and development of the pilot RESET observation tool.

The first question, *What sources of variance affect reliability across raters on the pilot RESET observation tool?* was answered through the results of the ANOVA results and G-studies. Additionally, the first question was answered partially through the two-facet, partially nested, *{o:t} x r* study design, which was selected based on previous, similar studies (Brennan, 2001; Cronbach et al., 1972; Erlich & Shavelson, 1978; Medley & Mitzel, 1958; Shavelson & Webb, 1991). That is, the determination of the study design identified raters, occasions, teachers, and their interactions, as the primary sources of variance.

The results of the ANOVA (Tables 4-1 to 4-3) found generally inconsistent patterns of variance across each content subscale for both data sets A and B. While almost all subscales reported the residual component ($\sigma^2_{ro,\, tro,\, e}$) as the highest source of variation: 58% (A) and 58.8% (B) (Lesson Objective), 30.8% (A) and 27.6% (B) (EBP Implementation), and 26.8% (A) and 21% (B) (Whole Lesson Review), each subscale reported different sources of variance as the second highest component. For Lesson Objective (SS1), the second highest source of variance was teachers (σ^2_t), but for EBP

Implementation (SS2) and Whole Lesson Review (SS3), the second highest source of variation was teachers and raters (σ^2_{tr}).

The second question, *When organized by content subscales, which part of the pilot RESET observation tool demonstrates the strongest and weakest levels of reliability?*, was answered using the relative and absolute G coefficients reported in the G studies conducted for both data sets A and B (Table 4-4). For both the relative and absolute G coefficient values, the highest levels of reliability can be found in SS1 (.71-.75), while SS2 reported the lowest levels of reliability (.42-.59). SS3 remained in the middle between the other two subscales (.54-.67).

The third question, *What are the optimal observation conditions to maximize reliability using the RESET observation tool?*, was answered using the results from the decision studies. And because studies that use generalizability theory seek to ask, "How many instances of which conditions of measurment are needed for acceptably precise measurment?" (Brennan & Lee, 2013, p. 3), as opposed to the more traditional null and alternative hypotheses used in typical quantitative studies, this study was guided by the research question: *How many occasions and raters are needed for acceptable levels of reliability when using the pilot RESET observation tool to evaluate special education teachers?*, which was similarly answered with the results from the decision studies.

However, neither of these questions can be as easily answered as the previous two because each subscale has its own set of relative G coefficient scores reported in the decision studies (i.e., each subscale has its own optimal observation conditions for maximum reliability). Using .65 as a 'cut-off' score, there are different ways to define "optimal observation conditions" for each subscale (using Data Set A). For SS1, the .65

score can be obtained three ways (using the minimum level of observations and raters): three observations, four raters; four observations, three raters; or six observations, two raters. For SS2, a minimum .65 score was not obtained. For SS3, a .65 score can be obtained two ways: four observations, five raters; or six observations, and four raters.

Interpretation of Findings

The overall purpose of this study was to identify how many occasions and raters are needed for acceptable levels of reliability on the pilot RESET observation tool. Generalizability theory was used to identify and measure sources of variance from rater data collected from two separate data coding sessions. The results from this study are in alignment with previous similar studies, but indicate that there is more work to be completed for future development on the pilot RESET observation tool. Thus, there are important points to review in the interpretation of the results.

First, consistent with other studies of teacher observation, multiple observations and raters are needed for more reliable ratings when using the pilot RESET observation tool to evaluate special education teachers. Overall, the use of at least four raters seems to be optimal (with the number of observations varying across subscales). This is also consistent with other generalizability theory studies on teacher observations (Hill, Charalambous, & Kraft, 2012). At the very least, like the MET study results, results indicate that more than one rater and more than one observation are needed for reliable evaluations (Ho & Kane, 2013; Kane & Cantrell, 2013). Across all subscales and both data sets, low levels of measurement reliability and high levels of error were reported when using just one to two of these conditions. This empirically consistent finding

suggests that future development on the RESET tool must plan for the use of multiple observations and raters to obtain acceptable levels of reliability.

In addition, future research on the RESET observation tool must also consider issues related to feasibility of practice. Four observations per school year might be too resource-intensive for schools and districts, and additional research is needed to determine ways to minimize error and increase measurement precision. One anticipated line of research to do this is to systematize the link between teachers and evaluators (raters). This type of study would require observed teachers to identify which instructional practice will be used BEFORE collecting the video observation data, and would improve overall levels of rater reliability.

Secondly, the findings from this study indicate that an overall evaluative judgment of special education teacher performance (SS3) is more reliable than ratings on individual lesson components (SS2), but not as reliable as the determination of a lesson's objective (SS1). However, there are a few possible reasons for this finding. First, the collapse of the evidence-based instructional components into one holistic score might have affected the results of the G study analyses. Because each component is defined through the review of literature specific to the instructional practice, the nuances of differences within specific scores might have been lost in the holistic score used in the G studies. Secondly, the rubric itself might be too restrictive in the determination of a teacher's ability to implement very specific instructional practice characteristics within one lesson. For example, future research will need to address how to distinguish the difference between a teacher's (in)ability to implement a specific instructional practice component when the need is present, and the teacher might not even be aware that an

instructional need even exists. Third, the lower levels of reliability reported in SS2 might be due to the simple fact that instructional practice is an extremely complex activity and is difficult to reduce down to a single numerical score. Fourth, because instructional practice can be a very complex activity, and because SS2 is comprised of the essential building blocks of instructional practice, it leaves itself vulnerable to issues that influence rater disagreement. And given that the occasion of the observation can be one of the greatest sources of error to resolve in observation protocol design, a recent study used multivariate generalizability theory to more precisely measure the influence of occasions on scores (Meyer, Cash, & Mashburn, 2011), suggesting that more complex uses of G theory might be beneficial. Lastly, the pilot RESET observation tool was developed in alignment with Danielson's (2011) assertions that an effective evaluation system should ensure teacher quality and promote professional development. With this in mind, even though the overall judgment of a teacher's practice was found to be more reliable in this study (SS3), it does not really address specific components of instructional practice. The higher levels of reliability found in SS3 might be useful in assisting schools and districts with relative decisions, but it is the feedback found in SS2 that will provide a teacher with targeted, specific feedback to improve components of evidence-based instructional practice.

Lastly, there are some issues overall that might have affected the results of the generalizability and decisions study analyses completed in this paper. Firstly, the use of two data sets over a six-month period may have led to a range of unaccounted sources of variance (e.g., differences in training sessions, different data sets etc.). Secondly, the consistent differences in results between data sets A and B seem to suggest that the

unexpected small rater sample size (five compared to the expected eight) might have influenced the results. The results from the decision studies seem to uphold this as well as the more raters, the higher the levels of reliability. Thirdly, while the impact may have been minimal, raters reported during both sessions that coding errors were occurring during the completion of the evaluation (e.g., a teacher might have indicated a "2" for a particular rating, but when going back to review, found that the score had been changed to a "1"). Again, it is not known how prevalent this type of occurrence was, but it would only have required a few occurrences per rater to negatively impact measures of agreement given the smaller facet sizes.

Recommendations for Future Research

Future studies should focus on further explorations of reliability, preliminary work on validity, and measurements of improved teacher instructional practice after using feedback from the tool. Overall, further research in areas related to rater reliability is needed. Although previous studies suggest that lower numbers are possible, the results from this study suggest that the larger the sample size, the more accurate the results. Additionally, future studies should investigate the use of different raters that might eventually be tasked with evaluating special education teachers using the pilot RESET observation tool (i.e., principals, special education teachers with specific expertise, mentor teachers, district personnel, and university faculty).

Conclusion

The overall purpose of this study was to identify how many occasions and raters are needed for acceptable levels of reliability when using the pilot RESET observation

tool to evaluate special education teachers. Generalizability and decision study analyses were completed to identify and measure sources of variance from rater data collected from two separate data coding sessions. The results from this study are supported by similar results found in previous research, but also suggest that additional work is needed to refine and develop the optimal use of the tool.

The purpose of special education is to provide individualized instruction to students who present with the most intensive of needs. Students served through special education require teachers who are highly skilled in the most effective forms of instructional practice. Unfortunately, the profession has been characterized with high attrition and lack of qualified teachers, instead of one that is defined as an elite group of educators. Although the pilot RESET observation tool is not able to solve any of the systemic problems found in the field, it does attempt to address these problems by measuring what is most important to positively impacting student achievement: effective instruction. For these reasons, ongoing work on the pilot RESET observation tool should continue to focus on improving the precision and reliability of the tool, so that students with disabilities are supported with the levels of professionalism they deserve.

REFERENCES

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*, 1034–1037. doi:10.1126/science.1207998

Ashton, J. (2011). The CEC professional standards: A Foucauldian genealogy of the re/construction of special education. *International Journal of Inclusive Education*, *15*(8), 775-795.

Baker, A. (2013, January 18). More money at risk on teacher evaluations. *New York Times*. New York. Retrieved from http://www.nytimes.com/2013/01/19/nyregion/more-money-at-risk-over-teacher-evaluations.html

Baker, A., & Santora, M. (2013, January 17). No deal on teacher evaluations; City risks losing $450 million. *New York Times*. New York. Retrieved from http://www.nytimes.com/2013/01/18/nyregion/new-york-city-talks-on-teacher-evaluations.html?ref=nyregion

Baker, E. L., Barton, P. E., Darling-Hammond, L., Haertel, E., Ladd, H. F., Linn, R. L., Ravitch, D., et al. (2010). Problems with the use of student test scores to evaluate teachers. *Economic Policy* (Vol. 278, pp. 1–29). Economic Policy Institute. Retrieved from http://epi.3cdn.net/b9667271ee6c154195_t9m6iij8k.pdf

Baker, S. K., Chard, D. J., Ketterlin-Geller, L. R., Apichatabutra, C., & Doabler, C. (2009). Teaching writing to at-risk students : The quality of evidence for self regulated strategy development. *Exceptional Children*, *75*(3), 303–318.

Ball, D. L., & Forzani, F. M. (2011). Teaching skillful teaching. *Educational Leadership*, *68*(4), 40–45.

Ballou, D., Sanders, W., & Wright, P. (2004). Controlling for student background in value-added assessment of teachers. *Journal of Educational and Behavioral Statistics*, *29*(1), 37–65.

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*(2-3), 62–87. doi:10.1080/10627197.2012.715014

Bellini, S., Peters, J. K., Benner, L., & Hopf, A. (2007). A meta-analysis of school-based social skills interventions for children with autism spectrum disorders. *Remedial and Special Education*, *28*(3), 153–162.

Berkeley, S., Scruggs, T. E., & Mastropieri, M. A. (2009). Reading comprehension instruction for students with learning disabilities, 1995-2006: A meta-analysis. *Remedial and Special Education*, *31*(6), 423–436. doi:10.1177/0741932509355988

Betebenner, D. (2009). Norm- and criterion-referenced student growth. *Educational Measurment: Issues and Practices*, *28*(4), 42–51. doi:10.1111/j.1745-3992.2009.00161.x

Billingsley, B. S. (2004). Special education teacher retention and attrition: A critical analysis of the research literature. *The Journal of Special Education*, *38*(1), 39–55. doi:10.1177/00224669040380010401

Bloom, H. S., Hill, C. J., Black, A. R., & Lipsey, M. W. (2008). Performance trajectories and performance gaps as achievement effect-size benchmarks for educational interventions. *Journal of Research on Educational Effectiveness*, *1*(4), 289–328. doi:10.1080/19345740802400072

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

Braun, H. I. (2012). *Using student progress to evaluate teachers: A primer on value-added models*. Princeton, NJ.

Brennan, R. L. (1992). An NCME instructional module on generalizability theory. *Educational Measurement: Issues and Practices*, *11*(4), 27–34.

Brennan, R. L. (2001). *Generalizability theory*. New York: Springer.

Brennan, R. L., & Lee, W.-C. (2013). Generalizability theory and applications. *National Council on Measurement in Education*. San Francisco, CA.

Brennan, R. L., & Prediger, D. J. (1981). Coefficient kappa: Some uses, misuses, and alternatives. *Educational and Psychological Measurement*, *41*, 687–699.

Briggs, D., & Domingue, B. (2011). *Due diligence and the evaluation of teachers: A review of the value-added analysis underlying the effectiveness rankings of Los Angeles Unified School District teachers by the Los Angeles Times*. Boulder, CO. Retrieved from http://nepc.colorado.edu/publication/due-diligence

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

Browder, D. M., Spooner, F., Harris, A. A., & Wakeman, S. (2008). A meta-analysis on teaching mathematics to students with significant cognitive disabilities, (4), 407–432.

Bruckner, C. T., & Yoder, P. (2006). Interpreting kappa in observational research: Baserate matters. *American Journal on Mental Retardation*, *111*(6), 433–41. doi:10.1352/0895-8017(2006)111[433:IKIORB]2.0.CO;2

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: Routledge.

Cardinet, J., Tourneur, Y., & Allal, L. (1976). The symmetry of generalizability theory: Applications to educational measurement. *Journal of Educational Measurement*, *13*(2), 321–328.

Council for Exceptional Children (2009). *What every special educator must know: Ethics, standards and guidelines*. Arlington, VA: Council for Exceptional Children. Retrieved from http://www.cec.sped.org/Content/NavigationMenu/ProfessionalDevelopment/ProfessionalStandards/What_Every_Special_Educator_Should_Know_6th_Ed_revised_2009.pdf

Council for Exceptional Children (2012). *The council for exceptional children's position on special education teacher evaluation*. Arlington, VA. Retrieved from http://www.cec.sped.org/Content/NavigationMenu/PolicyAdvocacy/CECProfessionalPolicies/Position_on_Special_Education_Teacher_Evaluation_Background.pdf

Chard, D. J., Ketterlin-Geller, L. R., Baker, S. K., Doabler, C., & Apichatabutra, C. (2009). Repeated reading interventions for students with learning disabilities: Status of the evidence. *Exceptional Children*, *75*(3), 263–281.

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2011). *The long-term impacts of teachers: Teacher value-added and student outcomes in adulthood (NBER Working Paper 17699)*. Cambridge, MA. Retrieved from http://www.nber.org/papers/w17699

Connelly, V., & Graham, S. (2009). Student teaching and teacher attrition in special education. *Teacher Education and Special Education*, *32*(3), 257–269. doi:10.1177/0888406409339472

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

Cook, B. G., Tankersley, M., & Landrum, T. J. (2009). Determining evidence-based practices in special education. *Exceptional Children*, *75*(3), 365–383.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. New York: Wiley.

Danielson, C. (2007). *Enhancing professional practice: A framework for teaching*. Alexandria, VA.: Association for Supervision and Curriculum Development.

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

Danielson, C. (2013). *The framework for teaching evaluation instrument, 2013 Edition* (2nd ed.). Princeton, NJ: Danielson Group.

Darling-Hammond, L. (2010). *Evaluating teacher effectiveness: How teacher performance assessments can measure and improve teaching* (pp. 1–36).

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

Dexter, D. D., Park, Y. J., & Hughes, C. A. (2011). A Meta-Analytic Review of Graphic Organizers and Science Instruction for Adolescents with Learning Disabilities: Implications for the Intermediate and Secondary Science Classroom. *Learning Disabilities Research & Practice*, *26*(4), 204–213. doi:10.1111/j.1540-5826.2011.00341.x

Doabler, C. T., Fien, H., Nelson-Walker, N. J., & Baker, S. K. (2012). Evaluating Three Elementary Mathematics Programs for Presence of Eight Research-Based Instructional Design Principles. *Learning Disability Quarterly*, *35*(4), 200–211. doi:10.1177/0731948712438557

Ehlert, M., Koedel, C., Parsons, E., & Podgursky, M. (2012). *Selecting growth measures for school and teacher evaluations*. Washington, DC.

Elliott, K. (2012). Charlotte Danielson releases Framework for Teaching evaluation instrument, 2013 edition. San Francisco, CA. Retrieved from http://www.teachscape.com/about/press-and-news/2012/charlotte-danielson-releases-framework-for-teaching-evaluation-instrument-2013-edition.html

Elliott, S. N., & Kurz, A. (2012). Opportunity to learn as a moderating variable in growth. *CCSSO NCSA*. Retrieved from http://www.ncaase.com/docs/NCAASE_OTL_NCME_2012.pdf

Erlich, O., & Shavelson, R. J. (1976). The application of generalizability theory to the study of teaching. Beginning teacher evaluation study.

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*(2), 77–89.

Feinstein, A. R., & Cicchetti, D. V. (1990). High agreement but low kappa: I. The problems of two paradoxes. *Journal of Clinical Epidemiology*, *43*(6), 543–549.

Feng, L., & Sass, T. R. (2010). *What makes special education teachers special? Teacher training and achievement of students with disabilities* (pp. 1–38). Washington DC.

Fixsen, D., Blase, K., Metz, A., & Van Dyke, M. (2013). Statewide implementation of evidence-based programs. *Exceptional Children*, *79*(2), 213–230.

Floden, R. E. (2012). Teacher value added as a measure of program quality: Interpret with caution. *Journal of Teacher Education*, *63*(5), 356–360. doi:10.1177/0022487112454175

Foegen, A., Espin, C. A., Allinder, R. M., & Markell, M. A. (2001). Translating research into practice: Preservice teachers' beliefs about curriculum-based measurement. *The Journal of Special Education*, *34*(4), 226–236.

Fuchs, L. S., & Fuchs, D. (2005). Enhancing mathematical problem solving for students with disabilities. *The Journal of Special Education*, *39*(1), 45–57.

Gersten, R., Chard, D. J., Jayanthi, M., Baker, S. K., Murphy, P., & Flojo, J. (2009). Mathematics instruction for students with learning disabilities: A meta-analysis of instructional components. *Review of Educational Research*, *79*(3), 1202–1242.

Gersten, R., Keating, T., Yovanoff, P., & Harniss, M. K. (2001). Working in special education: Factors that enhance special educators' intent to stay. *Exceptional Children*, *67*(4), 549–567.

Gersten, R., Jordan, N. C., & Flojo, J. R. (2005). Early identification and interventions for students with mathematics difficulties. *Journal of Learning Disabilities*, 38(4), 293-304.

Gersten, R., & Smith-Johnson, J. (2001). Reflections on the research to practice gap. *Teacher Education and Special Education*, *24*(4), 356–361. doi:10.1177/088840640102400409

Gersten, R., Vaughn, S., Deshler, D., & Schiller, E. (1997). What we know about using research findings: Implications for improving special education practice. *Journal of Learning Disabilities*, *30*(5), 466–76. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/9293227

Goe, L., Bell, C., & Little, O. (2008). *Approaches to evaluating teacher effectiveness: A research synthesis* (pp. 1–103). Washington DC.

Goe, L., Biggers, K., & Croft, A. (2012). *Linking teacher evaluation to professional development: Focusing on improving teaching and learning*. Retrieved from http://www.tqsource.org/publications/LinkingTeacherEval.pdf

Goe, L., & Croft, A. (2009). *Methods of evaluating teacher effectiveness*. Retrieved from http://www.tqsource.org/publications/RestoPractice_EvaluatingTeacherEffectivenes s.pdf

Goe, L., & Holdheide, L. (2011). *Measuring teachers' contributions to student learning growth for nontested grades and subjects* (pp. 1–32). Washington DC. Retrieved from http://www.tqsource.org/publications/MeasuringTeachersContributions.pdf

Goodman, S., & Turner, L. (2010). Teacher incentive pay and educational outcomes: Evidence from the NYC Bonus Program. In H. K. School (Ed.), *PEPG Conference*

(pp. 1–37). Cambridge, MA. Retrieved from http://www.hks.harvard.edu/pepg/MeritPayPapers/goodman_turner_10-07.pdf

Graham, S. (2009). Special issue: Evidence-based practices for reading, math, writing, and behavior. *Exceptional Children*, *75*(3).

Greenwood, C. R., Horton, B. T., & Utley, C. A. (2002). Academic engagement: Current perspectives on research and practice. *School Psychology Review*, *31*(3), 328–349.

Hanushek, E. A., & Rivkin, S. G. (2010a). Generalizations about using value-added measures of teacher quality. *American Economic Review*, *100*(2), 267–271. doi:10.1257/aer.100.2.267

Hanushek, E. A., & Rivkin, S. G. (2010b). *Using value-added measures of teacher quality* (pp. 1–6). Washington DC.

Heck, R. H. (2007). Examining the relationship between teacher quality as an organizational property of schools and students' achievement and growth rates. *Educational Administration Quarterly*, *43*(4), 399–432. doi:10.1177/0013161X07306452

Hendrickson, A., & Yin, P. (2010). Generalizability theory. In G. R. Hancock & R. O. Mueller (Eds.), *The reviewer's guide to quantitative methods in the social sciences* (pp. 115–122). New York: Routledge. Retrieved from http://public.eblib.com/EBLPublic/PublicView.do?ptiID=481046

Hill, H. C., Blunk, M. L., Charalambous, C. Y., Lewis, J. M., Phelps, G. C., Sleep, L., & Ball, D. L. (2008). Mathematical knowledge for teaching and the mathematical

quality of instruction: An exploratory study. *Cognition and Instruction*, *26*(4), 430–511. doi:10.1080/07370000802177235

Hill, H. C., Charalambous, C. Y., Blazar, D., McGinn, D., Kraft, M. A., Beisiegel, M., Humez, A., et al. (2012). Validating arguments for observational instruments: Attending to multiple sources of variation. *Educational Assessment*, *17*(2-3), 88–106. doi:10.1080/10627197.2012.715019

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*(2), 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. (2012). State considerations in designing and implementing evaluation systems that include teachers of students with disabilities. *Office of Special Education Programs Project Director's Conference*. Washington DC.

Holdheide, L., Browder, D., Warren, S., Buzick, H., & Jones, N. (2012). *Summary of "using student growth to evaluate educators of students with disabilities: Issues, challenges, and next steps"* (pp. 1–36). Retrieved from http://www.tqsource.org/pdfs/TQ_Forum_SummaryUsing_Student_Growth.pdf

Holdheide, L., Goe, L., Croft, A., & Reschly, D. J. (2010). *Challenges in evaluating special education teachers and english language learner specialists* (pp. 1–40). Washington DC.

Horner, R. H., Carr, E. G., Halle, J., McGee, G., Odom, S., & Wolery, M. (2005). The use of single-subject research to identify evidence-based practice in special education. *Exceptional Children*, *71*(2), 165–180.

Johnson, E. S., & Semmelroth, C. L. (2011). *Special education teacher evaluation: Issues, obstacles, potential directions*.

Johnson, E. S., & Semmelroth, C. L. (2012). Examining interrater agreement analyses of a pilot special education observation tool. *Journal of Special Education Apprenticeship*, *1*(4).

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 measures 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., & Darling-Hammond, L. (2012, June 24). Should student test scores be used to evaluate teachers? *Wall Street Journal*. Retrieved from

http://online.wsj.com/article/SB10001424052702304723304577366023832205042.html?mod=googlenews_wsj

Kane, T. J., & Staiger, D. O. (2012). *Gathering feedback for teaching: Combining high-quality observations with student surveys and achievement gains* (pp. 1–68). Retrieved from http://www.metproject.org/downloads/MET_Gathering_Feedback_Research_Paper.pdf

Karvonen, M., Wakeman, S., Moody, S., & Flowers, C. (2012). Building blocks: Cross-grade progressions in alternate assessments based on alternate achievement standards (AA-AAS). *National Council on Measurement in Education* (pp. 1–27). Vancouver, BC.

Konstantopoulos, S., & Chung, V. (2010). The persistence of teacher effects in elementary grades. *American Educational Research Journal*, *48*(2), 361–386. doi:10.3102/0002831210382888

Kretlow, A. G., & Bartholomew, C. C. (2010). Using coaching to improve the fidelity of evidence-based practices: A review of studies. *Teacher Education and Special Education*, *33*(4), 279–299. doi:10.1177/0888406410371643

Landis, J. R., & Koch, G. G. (1977). An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. *Biometrics*, *33*(2), 363–374.

Lewis, W. D., & Young, T. V. (2013). The Politics of Accountability: Teacher Education Policy. *Educational Policy*, *27*(2), 190–216. doi:10.1177/0895904812472725

Lipscomb, S., Teh, B., Gill, B., Chiang, H., & Owens, A. (2010). *Teacher and principal value-added: Research findings and implementation practices*. Cambridge, MA. Retrieved from http://www.mathematica-mpr.com/publications/PDFs/education/teacherprin_valueadded.pdf

Littrell, P. C., Billingsley, B. S., & Cross, L. H. (1994). The effects of principal support on special and general educators' stress, job satisfaction, school commitment, health, and intent to stay in teaching. *Remedial and Special Education*, *15*(5), 297–310.

Lohr, S. L. (2012). The value Deming's ideas can add to educational evaluation. *Statistics, Politics, and Policy*, *3*(2). doi:10.1515/2151-7509.1057

Mariano, L. T., McCaffrey, D. F., & Lockwood, J. R. (2010). A model for teacher effects from longitudinal data without assuming vertical scaling. *Journal of Educational and Behavioral Statistics*, *35*(3), 253–279. doi:10.3102/1076998609346967

Martineau, J. A. (2006). Distorting value added: The use of longitudinal, vertically scaled student achievement data for growth-based, value-added accountability. *Journal of Educational and Behavioral Statistics*, *31*(1), 35–62.

Martineau, J. A., Paek, P., Keene, J., & Hirsch, T. (2007). Integrated, comprehensive alignment as a foundation for measuring student progress. *Educational Measurement: Issues and Practices*, *Spring*, 28–35.

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

McGuinn, P. (2012). *The state of teacher evaluation reform* (pp. 1–55). Washington DC. Retrieved from http://www.americanprogress.org/wp-content/uploads/2012/11/McGuinn_TheStateofEvaluation-1.pdf

McLeskey, J. (2011). Supporting improved practice for special education teachers. *Journal of Special Education Leadership*, *24*(1), 26–36.

McLeskey, J., Tyler, N. C., & Flippin, S. S. (2004). The supply of and demand for special education teachers : A review of research regarding the chronic shortage of special education teachers. *The Journal of Special Education*, *38*(1), 5–21.

Medley, D. M., & Mitzel, H. E. (1958). Application of analysis of variance to the estimation of the reliability of observations of teachers' classroom behaviors. *The Journal of Experimental Education*, *27*(1), 23–35.

Mehta, J. (2013). How paradigms create politics: The transformation of American educational policy, 1980-2001. *American Educational Research Journal*, *50*(2), 285–324. doi:10.3102/0002831212471417

Meyer, J. P., Cash, A. H., & Mashburn, A. (2011). Occasions and the reliability of classroom observations: Alternative conceptualizations and methods of analysis. *Educational Assessment*, *16*(4), 227–243. doi:10.1080/10627197.2011.638884

Mihaly, K., Mccaffrey, D. F., Staiger, D. O., & Lockwood, J. R. (2013). *A composite estimator of effective teaching* (pp. 1–51).

Mislevy, R. J., & Haertel, G. D. (2006). Implications of evidence-centered design for educational testing. *Educational Measurement: Issues and Practices*. *25*(4), 6-20.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives*, *1*(1), 3–62.

Moore, C. M. (2012). The role of school environment in teacher dissatisfaction among U.S. public school teachers. *SAGE Open*, *2*(1), 1–16. doi:10.1177/2158244012438888

Morgan, P. L., Frisco, M. L., Farkas, G., & Hibel, J. (2008). A propensity score matching analysis of the effects of special education services. *The Journal of Special Education*, *43*(4), 236–254. doi:10.1177/0022466908323007

Moscoso, S., Tello, F., & López, J. (2006). Using generalizability theory to assess the validity of the evaluation process. *Quality and Quantity*, *40*(3), 315-329.

Murphy, P., & Rainey, L. (2012). *Modernizing the state education agency: Different paths toward performance management*. Seattle, WA. Retrieved from http://www.crpe.org/sites/default/files/pub_states_ModernizingSEAs_sept12.pdf

National Autism Center. (2009). National standards report. Randolph, MA: National Autism Center.

National Council on Teacher Quality. (2011). *State of the states: Trends and early lessons on teacher evaluation and effectiveness policies*. Washington DC.

National Council on Teacher Quality. (2012). *State of the states 2012: Teacher effectiveness policies*. *National Standards Report*. (2009). Washington DC.

Newman, L., Wagner, M., Knokey, A.-M., Marder, C., Nagle, K., Shaver, D., Wei, X., et al. (2011). *The post-high school outcomes of young adults with disabilities up to 8 years after high school: A Report from the National Longitudinal Transition Study-2 (NLTS2)* (Vol. 2, pp. 1–29). Washington DC.

Newton, X. A., Darling-Hammond, L., Haertel, E., & Thomas, E. (2010). Value-added modeling of teacher effectiveness: An exploration of stability across models and contexts. *Education Policy Analysis Archives*, *18*(23).

Nougaret, A. A., Scruggs, T. E., & Mastropieri, M. A. (2005). Does teacher education produce better special education teachers? *Exceptional Children*, *71*(3), 217–229.

Odom, S. L. (2009). The tie that binds: Evidence-based practice, implementation science, and outcomes for children. *Topics in Early Childhood Special Education*, *29*(1), 53–61.

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

Odom, S. L., Brown, W. H., Frey, T., Karasu, N., Smith-Canter, L. L., & Strain, P. S. (2003). Evidence-based practices for young children with autism: Contributions for

single-subject design research. *Focus on Autism and Other Developmental Disabilities*, *18*(3), 166–175. doi:10.1177/10883576030180030401

Odom, S. L., Collet-Klingenberg, L., Rogers, S. J., & Hatton, D. D. (2010). Evidence-based practices in interventions for children and youth with Autism Spectrum Disorders. *Preventing School Failure: Alternative Education for Children and Youth*, *54*(4), 275–282. doi:10.1080/10459881003785506

Odom, S. L., Cox, A. W., & Brock, M. E. (2013). Implementation science, professional development, and Autism Spectrum Disorders. *Exceptional Children*, *79*(2), 233–251.

Parker, R. I., Vannest, K. J., & Davis, J. L. (January 01, 2011). Effect size in single-case research: a review of nine nonoverlap techniques. *Behavior Modification,* 35(4), 303-22.

Partee, G. L. (2012). *Using multiple evaluation measures to improve teacher effectiveness: State strategies from Round 2 of No Child Left Behind Act Waivers*. Washington, DC. Retrieved from http://www.americanprogress.org/wp-content/uploads/2012/12/MultipleMeasures-2.pdf

Prewitt, K., Schwandt, T. A., & Straf, M. L. (2012). *Using science as evidence in public policy center*. Washington, DC: The National Academies Press.

Prince, C. D., Schuermann, P. J., Guthrie, J. W., Witham, P. J., Milanowski, A. T., & Thorn, C. A. (2009). *The other 69 percent: Fairly rewarding the performacne of*

teachers of nontested subjects and grades. Washington, DC. Retrieved from http://www.cecr.ed.gov/guides/other69Percent.pdf

Ravitch, D. (2010). *The death and life of the great American school system: How testing and choice are undermining education*. New York: Basic Books.

Riley, B. (2012). *Waive to the top: The dangers of legislating education policy from the executive branch*. Washington, DC. Retrieved from http://heartland.org/sites/default/files/waive-to-the-top.pdf

Roberts, G., Torgesen, J. K., Boardman, A., & Scammacca, N. (2008). Evidence-based strategies for reading instruction of older students with learning disabilities. *Learning Disabilities Research & Practice*, *23*(2), 63–69. doi:10.1111/j.1540-5826.2008.00264.x

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

Rockoff, J. E., & Speroni, C. (2010). Subjective and objective evaluations of teacher effectiveness. *American Economic Review*, *100*(May), 261–266.

Rothstein, J. (2010). Teacher quality in educational production: tracking, decay, and student achievement*. *Quarterly Journal of Economics*, *125*(1), 175–214.

Rothstein, J., & Mathis, W. J. (2013). *Review of two culminating reports from the MET project*. Retrieved from http://nepc.colorado.edu/files/ttr-final-met-rothstein.pdf

Russ, S., Chiang, B., Rylance, B. J., & Bongers, J. (2001). Caseload in special education : An integration of research findings. *Exceptional Children*, *67*(2), 161–172.

Santoro, D. A. (2011). Good teaching in difficult times: Demoralization in the pursuit of good work. *American Journal of Education*, *118*(1), 1–23.

Sawchuk, S. (2012). "Value added" measures at secondary level questioned. *Education Week*, *32*(6). Retrieved from http://www.edweek.org/ew/articles/2012/10/24/09tracking_ep.h32.html?tkn=ZZCFykFSINJI5m/w8+2s7GW0dBIB72PTOEza&cmp=clp-sb-ascd&print=1

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*(6), 437–449. doi:10.1177/0741932508327465

Seidel, T., & Shavelson, R. J. (2007). Teaching effectiveness research in the past decade: The role of theory and research design in disentangling meta-analysis results. *Review of Educational Research*, *77*(4), 454–499. doi:10.3102/0034654307310317

Semmelroth, C. L., Johnson, E. S., & Allred, K. (in press). Special educator evaluation: Cautions, concerns and considerations. *Journal of the American Academy of Special Education Professionals*.

Shavelson, R. J., & Dempsey, N. (1975). *Generalizability of measures of teacher effectiveness and teaching process (Technical Report #75-4-2, Beginning Teacher*

Evaluation Study). San Francisco, CA: Far West Laboratory for Educational Research and Development.

Shavelson, R. J., & Dempsey-Atwood, N. (1976). Generalizability of measures of teaching behavior. *Review of Educational Research*, *46*(4), 553–611.

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

Sindelar, P. T., Brownell, M. T., & Billingsley, B. (2010). Special education teacher education research: Current status and future directions. *Teacher Education and Special Education*, *33*(1), 8–24. doi:10.1177/0888406409358593

Slavin, R. E., Lake, C., Davis, S., & Madden, N. A. (2009). Effective programs for struggling readers : A best-evidence synthesis. *Best Evidence Encyclopedia*.

Slavin, R., & Madden, N. a. (2011). Measures inherent to treatments in program effectiveness reviews. *Journal of Research on Educational Effectiveness*, *4*(4), 370–380. doi:10.1080/19345747.2011.558986

Smith, G. J., Schmidt, M. M., Edelen-Smith, P. J., & Cook, B. G. (2013). Pasteur's Quadrant as the bridge linking rigor with relevance. *Exceptional Children*, *79*(2), 147–161.

Smolkowski, K., & Gunn, B. (2012). Reliability and validity of the Classroom Observations of Student–Teacher Interactions (COSTI) for kindergarten reading instruction. *Early Childhood Research Quarterly*, *27*(2), 316–328. doi:10.1016/j.ecresq.2011.09.004

Spooner, F., Algozzine, B., Wood, C. L., & Hicks, S. C. (2010). What we know and need to know about teacher education and special education. *Teacher Education and Special Education*, *33*(1), 44–54. doi:10.1177/0888406409356184

Spooner, F., Knight, V. F., Browder, D. M., & Smith, B. R. (2012). Evidence-based practice for teaching academics to students with severe developmental disabilities. *Remedial and Special Education*, *33*(6), 374–387. doi:10.1177/0741932511421634

Stempien, L. R., & Loeb, R. C. (2002). Differences in job satisfaction between general education and special education teachers: Implications for retention. *Remedial and Special Education*, *23*(5), 258–267. doi:10.1177/07419325020230050101

Students Come First. (2011). Retrieved July 11, 2012, from http://www.studentscomefirst.org/

Swanson, H. Lee, Sachse-Lee, C. (2000). A meta-analysis of single-subject-design intervention research for students with LD. *Journal of Learning Disabilities*, *33*(2), 114–136.

Test, D. W., Richter, S., Knight, V., & Spooner, F. (2011). A comprehensive review and meta-analysis of the social stories literature. *Focus on Autism and Other Developmental Disabilities*, *26*(1), 49-62.

Tindal, G., Yovanoff, P., & Geller, J. P. (2010). Generalizability theory applied to reading assessments for students with significant cognitive disabilities. *The Journal of Special Education*, *44*(1), 3–17. doi:10.1177/0022466908323008

Tyler, N. C., Yzquierdo, Z., Lopez-Reyna, N., & Flippin, S. S. (2004). Cultural and linguistic diversity and the special education workforce: A critical overview. *The Journal of Special Education*, *38*(1), 22–38.

van den Heuvel, J. R., Hansen, M., & Ilangakoon, C. (2012). Examining student growth on four states' alternate assessments: Measurements of success. *National Council on Measurement in Education* (pp. 1–77). Vancouver, BC.

Vannest, K. J., & Hagan-Burke, S. (2009). Teacher time use in special education. *Remedial and Special Education*, *31*(2), 126–142. doi:10.1177/0741932508327459

Vannest, K. J., Hagan-Burke, S., Parker, R. I., & Soares, D. A. (2011). Special education teacher time use in four types of programs. *The Journal of Educational Research*, *104*(4), 219–230. doi:10.1080/00220671003709898

Voight, A., Shinn, M., & Nation, M. (2012). The longitudinal effects of residential mobility on the aademic achievement of urban elementary and middle school students. *Educational Researcher*, *41*(9), 385–392. doi:10.3102/0013189X12442239

Watanabe, T. (2013, January 19). L.A. teachers union members OK new evaluation method. *Los Angeles Times Times*. Los Angeles. Retrieved from http://articles.latimes.com/2013/jan/19/local/la-me-utla-evals-20130120

Webb, N. M., Shavelson, R. J., & Haertel, E. H. (2006). Reliability coefficients and generalizability theory. *Handbook of Statistics*, *26*. doi:10.1016/S0169-7161(06)26004-8

Wehmeyer, M. L., Palmer, S. B., Shogren, K., Williams-Diehm, K., & Soukup, J. H. (2010). Establishing a Causal Relationship Between Intervention to Promote Self-Determination and Enhanced Student Self-Determination. *The Journal of Special Education*, *46*(4), 195–210. doi:10.1177/0022466910392377

Wehmeyer, Michael L., & Field, S. L. (2007). *Self-determination: Instruction and assessment strategies*. Thousand Oaks, CA: Corwin Press.

Winters, M. A., & Cowen, J. M. (2012). Grading New York: Accountability and student proficiency in America's largest school district. *Educational Evaluation and Policy Analysis*, *34*(3), 313–327. doi:10.3102/0162373712440039

U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, What Works Clearinghouse (2011). *What Works Clearinghouse: Procedures and standards handbook (version 1.2.* Retrieved from http://ies.ed.gov/ncee/wwc/DocumentSum.aspx?sid=19

U.S. Department of Education (2012a). *ESEA flexibility*. Retrieved from http://www2.ed.gov/policy/elsec/guid/esea-flexibility/index.html

U.S. Department of Education (2012b). *Race to the top fund*. Retrieved from http://www2.ed.gov/programs/racetothetop/index.html

APPENDIX A

“Subscale 1: Lesson Objective” - Excerpt of Rubric

from RESET Observation Tool User Manual

"Subscale 1: Lesson Objective" - Excerpt of rubric

from RESET Observation Tool User Manual

Question 2.3: Lesson and Component Alignment		
Question	**Question Type/Options**	**Criteria**
Q2.3 Is component objective aligned with the larger lesson objective?	Yes	Select this choice if the component objective is aligned with the lesson objective identified in the Lesson Objective. There should be no ambiguity how this objective aligns with the larger lesson objective.
	Partially	Select this choice if the component objective is partially aligned with the identified between the lesson and component objectives—there may still be some ambiguity how the lesson and component objective are aligned, but some relationship between the two can be observed.
	No/Inconclusive	Select this choice if the component objective is not aligned with the lesson objective identified in the Lesson Objective. OR Select this choice if it is unknown if the component objective is aligned with the lesson objective identified in the Lesson Objective. The observer should select this choice if either the lesson objective or component objectives were unidentifiable.

APPENDIX B

"Subscale 2: EBP Implementation" - Excerpt of Rubric

from RESET Observation Tool User Manual

"Subscale 2: EBP Implementation" - Excerpt of Rubric

from RESET Observation Tool User Manual

Explicit, Direct Instruction Component: Organized Instruction	
0	**1**
• The instructional purpose of the lesson is not presented, or inappropriate to students. • The teacher's spoken or written language is not clear and concise, or is inappropriate for the age, ability or culture of the students. • Students indicate through their questions that they are confused about the learning task. • The lesson is not tied to any previous learning.	• The teacher's attempt to explain the instructional purpose has only limited success, and/or directions and procedures must be clarified after initial student confusion. • The teacher's explanation of the content may contain minor errors; some portions are clear; other portions are difficult to follow. • The teacher's explanation of the content consists of a monologue or is purely procedural, with minimal academic participation/engagement from students. • The previous lesson is referenced, but no additional practice is provided.
2	**3**
• The teacher clearly communicates the instructional purpose of the lesson, including where it is logically situated within broader learning and the "big idea," and explains procedures and directions clearly. • Teacher's spoken and written language is clear and correct and uses vocabulary appropriate to the students' ages, abilities and interests. • Student academic engagement time with the learning task is maximized. • The lesson clearly ties to previous lessons by providing some form of additional practice, and the teacher has provided some form of cumulative review or applications.	• The teacher's spoken and written language is expressive, and the teacher finds opportunities to extend students' vocabularies. • The amount of new information presented is appropriate so that mastery could probably be achieved within class time, and efforts to arrange students and student materials to make the most effective use of class time are evident. • Pre-arrangement of materials is indicated through teacher preparation, classroom environment, student routine, etc.

Explicit, Direct Instruction Component: Sequenced Instruction	
0	**1**
• There is no sequenced instruction in lesson organization, pace or content, or the instructional sequence of the lesson is unclear. • The organization or planning of the sequenced instruction does not meet the needs of all learners in the classroom, or is designed to meet the needs of just one student group/level. • The teacher may provide an activity (e.g. worksheet) but the purpose of the activity is unclear, and/or there is no explanation provided how the activity fits into a larger sequence of instruction. • Students may indicate through their classroom behaviors that they are engaged, but this is a function of classroom routine, and not of sequenced instructional planning.	• There is an instructional sequence in the lesson, but the sequence is a result of a curriculum script and does not indicate that there was any previous or additional planning from the teacher. • The teacher might have a distinct style or character, and the lesson might have a loosely identified beginning, middle and end, but there are no consistent lesson features that can be identified, e.g. modeling, highlighting, feedback, review, or practice application. • The lesson is loosely structured with few opportunities for practice. • The teacher uses modeling and/or explains the generalizability of skills, but these efforts do not fit into an identifiable sequence of instruction.
2	**3**
• The instructional sequence is highly structured and provides students with multiple opportunities for successful practice, and paced at a brisk tempo. • The teacher utilizes direct instruction features like modeling, highlighting, feedback, review, and opportunities for student practice in an organized and deliberate way. • The lesson utilizes the classroom curriculum as a part or a form of support of the sequenced instruction, but the teacher is easily identifiable as the "captain" of the lesson and the applied sequenced instruction features.	• The instructional sequence is seamlessly and briskly paced. • The teacher smoothly guides students from initial practice to generalized skill training (if applicable). • The lesson utilizes the classroom curriculum in a minor role. The focus of the teacher is clearly on students and student success. • The lesson is effectively sequenced so that it maximizes the similarities of different units, and the arrangement of learning units is "exploited" so that they are related in some way.

Explicit, Direct Instruction Component: Scaffolding	
0	**1**
• The teacher does not provide any support to students to move to the next step. • Classroom instruction is increasingly challenging with no plan to help bring student performance to the higher level of performance.	• The teacher has not identified difficulties but provides some support as challenges arise. • The teacher provides ways to support learning but they are not presented in a strategic, structured or systematic way – rather they are very task specific and 'on the fly'. • Scaffolding is overly teacher-directed with no opportunity for transfer to the student.
2	**3**
• The teacher has identified some but not all difficulties that might be encountered and provides responsive (but not proactive) support. • The teacher provides strategies but they might not be consistent or structured with the rest of the learning environment. • The teacher develops some transfer of control, but the transition is not integrated into the process.	• 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 – for example, using a graphic organizer to support comprehension of a reading passage, or using a calculator when moving to multi-step word problem solving. • Activities are provided within a structured learning environment – not as an 'add on' or an 'after thought' – but provided intentionally to help move students to new level of learning. • Scaffolding is presented to provide a gradual transfer of control to the student for the learning activity.

Explicit, Direct Instruction Component: Practice and Review	
0	**1**
• The teacher does not monitor student learning. • The teacher does not check for understanding, and does not review instruction.	• The teacher provides inconsistent corrective feedback. • The teacher inconsistently checks for understanding, and reviews of instruction. The checks and

• The teacher does not provide immediate, corrective feedback when presented with a student's incorrect response. • Students are not provided with any opportunities to engage in self-assessment or progress monitoring. • Students are not aware of the criteria that they will be evaluated with and/or the performance standards that they are expected to achieve. • Students are not provided opportunities for practice, or are provided opportunities for practice in areas unrelated to the teacher's lesson.	reviews seem unplanned or unorganized. • Students know some of the criteria that they will be evaluated with and/or the performance standards they are expected to achieve. • There are few opportunities provided to generalize new skills. • Students are provided with few opportunities for practice.
2	**3**
• The teacher provides consistent corrective feedback, and has identified some of the areas students might have difficulties. • The teacher consistently checks for understanding. The teacher reviews instruction in ways that seem obviously planned and organized. • Most students are aware of the criteria that they will be evaluated with and/or the performance standards they are expected to achieve. • Most students are provided opportunities for practice and to generalize new skills. However, some of these opportunities might be developed within the moment, some of these may be planned within a structured environment.	• The teacher provides corrective feedback, frequent checks for understanding, and periodic reviews of instruction that integrates knowledge within a structured learning environment. • All students are provided opportunities to generalize new skills, and receive individual attention when necessary. • All student practice activities and exercises for are designed so that new information/skills are clear and manageable for students. • All students are aware of the evaluations being used to tests their mastery and/or are aware of the performance benchmarks they are expected to achieve.

"Whole Group Instruction" Teaching Component: Individualized Instruction	
0	**1**
• The teacher does not individualize instruction, nor appear to have made any instructional arrangements to account for student differences. • The teacher is non-responsive to student needs.	• The teacher utilizes some level of individualization or differentiation, e.g. grouping techniques, or use of different materials, but the teacher does not have observable instructional strategies in place.
2	**3**
• Individualized instructional strategies are evident through teacher practices that compensate for individual student needs. The teacher has organized and planned for individualized instruction. • The teacher makes responsive adjustments to instruction based on observations of student response and performance.	• Instructional scope and sequence is individualized. • Individualized learning objectives are sequenced, implemented and evaluated.

"Whole Group Instruction" Teaching Component: Skill Development	
0	**1**
• Teacher instruction does not include any connection to previous learning. For example, if a student completes a worksheet without any connection to a larger lesson objective, previous learning, etc., this would not count as skill development. • The teacher is unresponsive to student efforts to promote skills related to maintenance and generalization, or self-determination.	• The teacher includes some level of facilitation or generalization of skill development, but the instruction is loosely organized, and/or somewhat connected to a larger learning objective. • The teacher incorporates concepts related to self-determination, but the connections to the larger lesson, or the lesson is not structured or organized.
2	**3**
• 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 lesson is built into a larger learning objective and/or the	• The teacher integrates the development of affective, social, and life skills within academic curricula. • Instruction includes development of critical-thinking and problem-solving skills that promote self-awareness, self-management, self-control, self-reliance and self-esteem.

sequence of skill instruction is part of a larger process.	

"Whole Group Instruction" Teaching Component: Student Engagement	
0	**1**
• The teacher provides little to no opportunities for guided and independent practice for students. • The teacher provides little to no opportunities for students to participate in classroom activities.	• The teacher provides for some level of student participation or student practice, but the activities not individualized or not appropriate for individual student needs. • •
2	**3**
• The teacher has planned for multiple opportunities for student participation or student practice. The class activities are individualized. • Materials and time have been effectively managed and planned to promote high levels of academic student engagement for most students. • The teacher promotes some levels of self-independence and self-determination.	• The teacher provides for individualized opportunities for 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.

"Whole Group Instruction" Teaching Component: Feedback and Assessment	
0	**1**
• The teacher provides little to no instructional feedback. • The teacher does not use any type of assessment to inform instruction. • The teacher ineffectively manages students' behaviors and the classroom environment, resulting in lost instructional time, OR • The teacher uses feedback to redirect students and provide interventions when necessary, but provides little to no instructional feedback. • The teacher administers a whole group instructional assessment, but	• The teacher uses feedback to redirect students, and provides interventions when necessary or provides reactive (not pre-planned) instructional feedback. • The teacher administers a whole group instructional assessment with a basic explanation why students have to take it, or with a simple explanation how it ties into the larger learning objective, but the assessment is not individualized or not designed for specific student needs.

the purpose is unclear and/or the teacher has not provided any explanation how the assessment is tied to the larger learning objective.	
2	**3**
• The teacher effectively uses individualized feedback to praise and prompt students through the instructional process most of the time. • The teacher administers different individualized or small group instructional assessments, and provides clear explanations regarding the instructional purpose.	• Formal and informal assessments of behavior, learning, achievement and environments are used to inform instruction as evidenced by highly-individualized, organized instruction. • Feedback is used to promote learning, as well as redirect and intervene as necessary. • The teacher uses effective questioning techniques that challenge students either at the individual or whole group level.

Discrete Trial Teaching Component: Antecedent	
0	**1**
• The teacher does not provide an antecedent. • The teacher provides an antecedent, but it is inappropriate in its delivery or request for student level. • The teacher provides an antecedent, but the student is inconsistently attentive or not ready for instruction. • The teacher provides an antecedent, but when the student responds incorrectly (or not at all) the teacher does not provide a prompt. • The teacher provides an antecedent, but it is characterized with "patter," for example, if a student is being taught to discriminate a red circle from a blue one, the teacher should say, "Touch the red circle." rather than say "Let's see what a smart little student you are by showing me the difference between a red and a blue circle by touching the red one instead of the blue one."	• The teacher provides an antecedent, but it is not consistent in its delivery or succinctness. • The teacher provides an antecedent, but delivers it with little emphasis or intonation (when applicable). • The teacher provides an antecedent and prompting when needed, but the prompt is characterized by being reflexive, intrusive, and/or not specific to the antecedent OR • The teacher provides an antecedent and prompt but neither is delivered consistently.

• The environment has not been prepared for discrete trial teaching and the student is visibly distracted.	
2	**3**
• The antecedent is stated clearly, and succinctly. The antecedent is minimal in options and communicates exactly what is expected of the student. • The teacher provides an appropriate antecedent and prompt, but does not consistently fade the prompt. • The learning environment appears to be removed of distractions, and the student is attentive.	• The teacher provides an appropriate antecedent, varies instruction, and fades prompts as needed. • The learning environment appears to be removed of distractions, and the student is thoroughly attentive. • When teaching a new response, the teacher emphasizes certain words or phrases in an instruction, e.g., by altering loudness or intonation. • The teacher uses a variety of prompts, i.e. verbal, modeling, gesturing, or physical guidance, to support a student's response.

Discrete Trial Teaching Component: Response	
0	**1**
• The target response is not specified, or the target response is inappropriate for student. • The response is not observable. • There is ambiguity about whether or not the correct response has occurred.	• The teacher has specified a target response, but it is either inconsistent or lacks specificity. For example, if asked to describe the actions of a horse using the question, "What is the horse doing?" a correct student response is not clearly identified (The teacher must provide the student with information that a correct answer would include "running," "galloping," "trotting.") • There is some ambiguity about whether or not the response has occurred.
2	**3**
• The teacher has selected a target response that the learner can achieve. • There is little to no ambiguity about whether or not the response has occurred. • The target response is mostly defined	• The teacher has created a structured learning environment that utilizes a variety of nonverbal and verbal responses, and the teacher has anticipated and planned for possible incorrect student responses. • The target response is defined in

in observable terms and it appears that the teacher is utilizing some type of measurement system.	observable terms, and there is no ambiguity whether or not the response has occurred.

Discrete Trial Teaching Component: Consequence	
0	**1**
• There is no reinforcer. • The reinforcer is inconsistent, unpredictable, or delayed to the point of ineffectiveness. • The student is unresponsive or doesn't appear to care about the reinforcer; the student does not like the reinforcer.	• The effectiveness of the reinforcer is questionable-the student's response to the reinforcer is inconsistent. • The teacher responds with a reinforcer inconsistently—the teacher leaves some student responses without an appropriate response.
2	**3**
• The teacher consistently responds with a reinforcer. • The student attends to the reinforcer; the reinforcer appears to be effective for the student. • The teacher includes descriptive praise statements with reinforcers.	• The teacher provides the reinforcer contingently, immediately, and continuously. • The teacher has identified a reinforcer that the student responds to, and has prepared for alternate reinforcers in case the student's preferences change. • The teacher includes descriptive praise statements with reinforcers.

Discrete Trial Teaching Component: Intertrial Interval (ITI)	
0	**1**
• The teacher does not provide for any ITIs • The teacher pauses in between discrete trials as a result of poor planning or management, and not as a structured ITI. • The teacher uses ITI as a transition, e.g. relating the ITI to the next scheduled trial or instruction is included during the ITI.	• The teacher inconsistently applies the use of an ITI between the end of one trial and the beginning of another. • The ITIs are inconsistent in length of time. • The teacher uses an ITI, but it is mismanaged and results in undesirable responses such as fidgeting, whining, or crying.
2	**3**
• The ITI is unrelated to the next	• The teacher effectively uses the ITI as

scheduled trial and does not contain any instruction. • The teacher effectively uses the ITI as a pause between the previous and the preceding trials.	a pause between the previous and the preceding trials. • The teacher maintains an instructional momentum that is easily observable.

APPENDIX C

“Subscale 3: Whole Lesson Summary” - Excerpt of Rubric

from RESET Observation Tool User Manual

"Subscale 3: Whole Lesson Summary" - Excerpt of Rubric

from RESET Observation Tool User Manual

Question 6.2: Whole lesson effective use of time	
0	**1**
• The teacher did not use the time effectively for the lesson objective.	• While some or many of the components might have some instructional merit for some students, the teacher did not contextualize or anchor the components in a larger learning or lesson objective.
2	**3**
• The teacher mostly used the time effectively for the whole lesson, and utilized each lesson component as learning activities that contributed to the lesson as a whole. The whole lesson was pre-planned , structured and sequenced.	• The teacher used the time effectively from beginning to end, and the whole lesson was highly-structured and organized. The teacher also maintained instructional flexibility and efficiency according to student response.

Question 6.3: Whole lesson summary: Does the teacher appear to have a solid understanding of the content?	
0	**1**
• The teacher makes significant errors in lesson content. • The teacher provides incorrect information to students about the content. • When questioned by students, the teacher is unable to respond, or provides incorrect information.	• The teacher presents information in small, disconnected ways that does not create any connections to larger learning. • The teacher presents information so broadly it does not allow a student to develop generalizable skills. •
2	**3**
• The teacher presents information in an organized and structured way that allows most students to make connections between lesson and component objectives. • The teacher individualizes content to meet the needs of most students.	• All students are given opportunities to make connections at many different levels and settings across environments and settings.

Question 6.3: Whole lesson summary: Does the teacher implement effective instructional practices?	
0	**1**
• The teacher did not implement any effective instructional practices.	• The teacher has some elements of effective instructional practice, but the components are disconnected.
2	**3**
• The teacher implements effective instructional practice in mostly organized and structured ways. • The teacher individualizes instruction according to most student needs and response.	• The teacher implements highly-organized and sequenced individualized instruction that promotes positive learning for all students. In addition to the lesson instruction, the instruction enhances deeper learning for students in the areas of critical thinking, problem-solving, and performance skills.

Question 6.3: Whole lesson summary: Does the teacher effectively respond to student needs?	
0	**1**
• The teacher does not individualize instruction. • The teacher does not adapt or modify instruction according to student response.	• The teacher demonstrates a basic level of individualized instruction, e.g. grouping, assigning students to support staff, etc., but the instructional practice is not organized or cohesive. • The teacher partially modifies instruction according to some student response.
2	**3**
• The teacher is aware of the impact of learners' academic and social abilities, attitudes, and interests, and effectively plans individualized instruction accordingly. The teacher meets the instructional needs for most students.	• The teacher acknowledges and plans for all individualized learners, including those from culturally diverse backgrounds, and is prepared with strategies for addressing these differences.

APPENDIX D

Teachscape Video Capture Screenshot

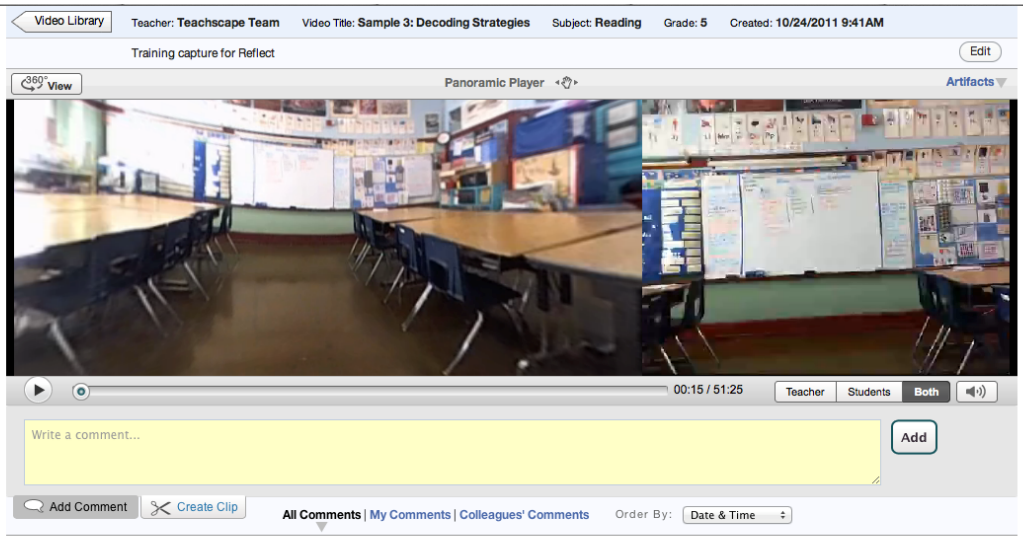

Figure D-1. Teachscape Video Capture Screenshot