DEVELOPING READING INSTRUCTION OBSERVATION PROTOCOLS FOR SPECIAL EDUCATION TEACHERS

by

Laura Ann Moylan

A dissertation

submitted in partial fulfillment

of the requirements for the degree of

Doctor of Education in Curriculum and Instruction

Boise State University

May 2021

BOISE STATE UNIVERSITY GRADUATE COLLEGE

DEFENSE COMMITTEE AND FINAL READING APPROVALS

of the dissertation submitted by

Laura Ann Moylan

Dissertation Title: Developing Reading Instruction Observation Protocols for Special Education Teachers

Date of Final Oral Examination: 18 March 2021

The following individuals read and discussed the dissertation submitted by student Laura Ann Moylan, and they evaluated the student's presentation and response to questions during the final oral examination. They found that the student passed the final oral examination.

Evelyn S. Johnson, Ed.D.	Chair, Supervisory Committee
Keith Thiede, Ph.D.	Member, Supervisory Committee
Sara Hagenah, Ph.D.	Member, Supervisory Committee

The final reading approval of the dissertation was granted by Evelyn S. Johnson, Ed.D., Chair of the Supervisory Committee. The dissertation was approved by the Graduate College.

ACKNOWLEDGMENTS

I would like to first acknowledge my advisor and committee chair Dr. Evelyn Johnson who has provided me with numerous opportunities for professional growth and then come along side with consistent encouragement and support. I am especially grateful for the opportunity to engage in her research as a member of the RESET project team, which then led to my pursuit of a doctoral degree. She is a model of self-determination and resilience, demonstrating to anyone fortunate enough to work with her the value of hard work and confidence in your vision. I would like to thank her for her willingness to stick with me throughout this process, patiently pressing me forward and then pushing when needed.

I would like to thank the additional members of my committee Dr. Keith Thiede and Dr. Sara Hagenah for all I have learned from them and for their guidance and feedback on this work.

Finally, I would like to express gratitude for my family. First, my husband Brent who for the past 30 years has supported and encouraged me throughout my career as an educator and my efforts to continue my education. He is patient, kind, and always solid in his belief that I will be successful. My daughter Sophie inspires me every day to be a better person, to be strong, and to believe in myself. I am extremely grateful to them both for their confidence in me and for the life we have together that has made it possible for me to pursue my interests and dreams.

ABSTRACT

Considerable resources have been invested in identifying effective reading instruction methods for students with disabilities. Unfortunately, students are not routinely receiving instruction aligned with these practices, impacting their ability to reach their potential. To improve reading instruction, teachers need to receive observation feedback and evaluations reflecting instructional practices shown to be effective. One way to ensure teachers are provided with feedback consistent with evidence based reading instruction is to develop observation protocols aligned to these practices. This dissertation addresses this problem with three distinct, yet interconnected, articles detailing the development of reading instruction observation protocols designed to provide accurate teacher evaluations and feedback to improve reading instruction for students with disabilities. Each protocol is part of the larger Recognizing Effective Special Education Teachers (RESET) observation system. The first article explains the framework that was applied to develop both the observation system and an explicit instruction observation protocol. The second and third articles describe the development of a comprehension and a decoding instruction observation protocol. Development included a comprehensive review of literature and rigorous testing. Results indicate the explicit instruction, comprehension, and decoding instruction protocols will provide reliable evaluations of a teacher's ability to implement instruction consistent with practices most effective for students with disabilities. Implications for practice and further research are discussed.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

CDLP	Comprehensive Decoding Lesson Protocol
EBP(s)	Evidence based practice(s)
ECD	Evidence Centered Design
EI	Explicit Instruction
MFRM	many-facet Rasch measurement
PLD(s)	performance level descriptor(s)
RESET	Recognizing Effective Special Education Teachers
SET	special education teacher
SWD	students with disabilities

CHAPTER ONE

Introduction

This dissertation consists of three articles representing a connected body of work. While each article stands alone, there is commonality linking all three articles. Specifically, the chapters in this dissertation are connected by the theme of developing a special education teacher observation system, Recognizing Effective Special Education Teachers (RESET; Johnson et al., 2018). Each of the chapters build upon one another by describing specific stages of the observation system and protocol development. Chapters Two, Three, and Four include articles written for publication in education journals. Each chapter contains detailed introductions to the article, including abstracts providing context for each article. The remainder of this chapter lays the foundation for the importance of and purpose behind the following chapters' articles, as well as this dissertation as a whole.

Chapter Two, *Using Evidence-Centered Design to Create a Special Educator Observation System*, explains how the Evidence-Centered Design framework was applied to ensure the thoughtful and rigorous development of RESET. Within this chapter, two studies are described. The first study describes the processes undertaken to create an initial set of performance-level descriptors for the RESET Explicit Instruction observation protocol. A team of raters independently scored a set of video recorded lessons using the Explicit Instruction protocol. Along with their scores, they also provided time stamped evidence and explanations to support their scoring decisions.

Using these data, a set of performance level descriptors was developed for each item on the protocol. The second study described in this article details the procedures used to analyze the reliability of the Explicit Instruction protocol. Raters used the fully developed Explicit Instruction protocol to evaluate a set of video recorded lessons. Using Many-facet Rasch measurement (MFRM), we were able to assess the reliability of the protocol and review how the teacher, item, rater, and lesson facets functioned. Results show the item, teacher, rater, and lesson facets achieved high psychometric quality, indicating the instrument will provide reliable evaluations of a teacher's ability to implement effective explicit instruction. The development and testing processes described in this chapter are replicated across future studies.

Chapter Three contains an article titled *Developing a Comprehension Instruction Observation Rubric for Special Education Teachers*. Using the framework and processes described in Chapter Two, a Reading for Meaning observation protocol detailing the elements of evidence-based comprehension instruction was developed and the psychometric properties were tested using MFRM. The process for developing the Reading for Meaning protocol began with a review of the research on comprehension instruction for students with disabilities (hereafter abbreviated as SWD). The research was synthesized into a set of components and items representing the key elements of effective reading comprehension instruction. Items indicating full implementation were written first. In order to develop accurate performance level descriptors, raters with expertise in reading instruction used the protocol while observing video recorded comprehension instruction. These raters provided scores indicating degrees of implementation and also provided time stamped evidence and notes explaining their

scoring. This information was then analyzed and translated into performance level descriptors for each item.

Using the fully developed Reading for Meaning rubric, trained raters watched video recorded lessons and scored each item on the protocol as 'implemented', partially implemented' or 'not implemented'. MFRM analysis indicated high psychometric quality for item, teacher, rater, and lesson facets suggesting the Reading for Meaning protocol will provide reliable evaluations of a teacher's ability to implement effective comprehension instruction for SWD.

Finally, Chapter Four contains a manuscript titled *Developing a Comprehensive Decoding Instruction Observation Protocol for Special Education Teachers*. The purpose for the study described in this paper was: 1) to examine the psychometric quality of the Comprehensive Decoding Lesson Protocol through MFRM analysis and 2) to analyze teachers' performance on the implementation of effective decoding instruction. The Comprehensive Decoding Lesson Protocol (hereafter abbreviated as CDLP) was designed to evaluate and support the implementation of systematic and explicit phonics instruction. The CDLP items and components were developed through an extensive review of the research on decoding instruction for SWD. Drafted items underwent multiple revisions as they were reviewed by content experts and internally tested using video recorded instruction. Once a complete set of items and performance level descriptors was completed the protocol was tested for reliability.

Patterned after the prior studies, trained raters scored video recorded reading lessons identified as decoding instruction using the CDLP. Rater scoring data was analyzed using MFRM and indicated strong psychometric properties for item, teacher,

rater, and lesson facets. These results suggest the CDLP will provide reliable evaluations of a teacher's ability to implement decoding instruction for SWD consistent with the effective instructional practices described in the research. Further analysis was conducted to examine the degree of implementation for each item on the protocol. Results of this analysis indicated low levels of proficient decoding instruction implemented by this sample of teachers as a whole. Using this data, it would be possible to provide targeted feedback to support improved implementation and to positively reinforce incidences of proficient implementation. Implications for both practice and continued research are discussed.

Student Reading Achievement

Acquiring the ability to read is fundamental to learning, success in school, and future engagement in the work-force and is therefore a primary focus in our education system. The implications of low levels of literacy extend into adulthood and impact both the health and economic stability of individuals and society as a whole (Miller et al., 2010). During the 2018-19 school year over seven million public school students received special education services, or 14% of public-school enrollment (National Center for Educational Statistics, 2020). SWD tend to have significant achievement gaps in reading compared with their peers in general-education, with substantial numbers of SWD performing below proficiency on state and national measures of reading (Judge & Bell, 2010; NCES, 2019; Schulte et al., 2016). The average reading achievement gap between SWD and students without disabilities has been reported to be as high as 1.17 SD, or the equivalent of 3.3 years of growth (Gilmour et al., 2018). One possible

contributor to low levels of reading proficiency for SWD may be the content, quality, and intensity of instruction students are receiving.

Observations for Reading Instruction and Research to Practice Gap

Over several decades considerable resources have been dedicated to the development and understanding of effective reading instruction for SWD (Ehri, 2004; Lane, 2014; NICHD, 2000). However, observational studies of instructional practice have identified a lack of consistent and effective implementation of evidence-based practices (hereafter abbreviated as EBP), particularly in settings focused on providing reading instruction for SWD (Moody et al., 2000; Swanson, 2008; Vaughn & Wanzek, 2014). Observational studies have consistently concluded reading instruction is frequently lacking critical components and the quality and intensity is inadequate to meet the instructional needs of SWD (Vaughn & Wanzek, 2014). Despite the extensive evidence supporting explicit, systematic decoding instruction as a critical component of reading instruction and intervention (Blachman et al., 2004; Denton et al., 2013; Ehri et al., 2001; Lovett et al., 2000; Torgesen et al., 2001), SWD are spending little time engaged in effective phonemic awareness and phonics instruction (Ciullo et al., 2016; Moody et al., 2000; Swanson, 2008). Comprehension instruction is infrequently observed across observational studies, or when observed, described as inadequate, lacking in strategy instruction, and primarily consisting of asking literal questions or students completing independent work (Cuillo et al., 2016; Klingner et al., 2010; Swanson, 2008; Swanson & Vaughn, 2010; Vaughn & Wansek, 2014). Further, students are spending limited amounts of time engaging with connected text (Kent et al., 2012; Swanson, 2008; Vaughn & Wanzek, 2014), with one synthesis indicating as little as 3-13 minutes spent reading

aloud and 6-10 minutes reading silently across educational settings (Vaughn & Wanzek, 2014). Overall, students have been observed spending excessive amounts of time during dedicated reading instruction on non-literacy related or passive activities (Kent et al., 2012; Vaughn & Wanzek, 2014). Additionally, concerns have been raised about the lack and depth of content knowledge among teachers providing reading instruction, potentially inhibiting their ability to explain concepts effectively, select appropriate examples, be diagnostic, and provide targeted feedback to students (Moats & Foreman, 2003; Moats, 2009; Washburn et al., 2011). While evidence strongly supports explicit, systematic instruction focused on the critical components of phonemic awareness, phonics, vocabulary, and comprehension as essential for students to reach their potential (NICHD, 2000); there appears to be a disconnect between what we know and what is consistently happening in classrooms.

Creating a teacher observation system aligned to the specific instruction practices found to improve reading performance for SWD is one way to provide teachers with clear guidance, improve instruction, and ultimately improve outcomes for students. When teachers are objectively evaluated and supported to improve instruction, accuracy in the implementation of EBP increases and there is a positive impact on student growth (Biancarosa et al., 2010; Fallon et al., 2015; Taylor & Tyler, 2012).

Teacher Observation Systems

If we are to close the research to practice gap and ensure SWD consistently receive high quality instruction, special education teachers must have sufficient knowledge of EBPs along with the skills and systematic support to sustain fidelity of implementation overtime (McLeskey & Billingsley, 2008). Teacher observation systems

have the potential for producing positive changes in practice and supporting sustained implementation when they integrate the improvement of teacher knowledge with ongoing opportunities for practice and feedback (Fallon et al., 2015; Snyder et al., 2015; Schles & Robertson, 2019; Solomon et al., 2012). To be an effective framework for supporting cumulative learning and improving instruction observation systems must provide teachers with feedback and guidance that is accurate, actionable, and subject-specific (Hill & Grossman, 2013). Unfortunately, current observation systems have been designed to be expedient, are generic in nature, and therefore limited in their utility to provide content specific and actionable feedback (Blazar et al., 2017; Hill & Grossman, 2013).

Observation systems designed to effectively measure the complexities of instruction and provide accurate, reliable ratings and feedback require rigorous development and evaluation (Hill & Grossman, 2013). Systems put in place without deliberate construction and assessment may not effectively detect variations in practice or may lack accuracy and clarity in expectations for performance, resulting in inappropriate decisions and feedback unlikely to result in the desired improvement (Hall, 2014). The Evidence-Centered Design framework (ECD) was applied to the development of the RESET observation system as a way to ensure thoughtful and rigorous development (Johnson et al., 2018). The ECD framework is a construct-centered approach to assessment development used to ensure the collection and interpretation of evidence and the stages of development remain consistent with the underlying construct the assessment is intended to measure (Mislevy et al., 2003).

The RESET protocols are high-inference observation instruments designed to capture the critical elements of effective instruction for SWD. The interpretation of such

complex, multi-dimensional practices and determinations of proficiency require high levels of expertise along with a shared understanding of practices and the language used to describe them. The instructional dimensions of observation protocols have been reported to be the most challenging for raters to score reliably (Gitomer et al., 2014; Ho & Kane, 2013). Even after increased training, feedback, and calibration exercises, rater reliability continues to be a persistent challenge with raters accounting for between 25 and as much as 70% of variance in scores assigned to the same lesson (Casabianca et al., 2015).

In research designs where a team of trained raters observes multiple teachers and lessons, statistical adjustments are able to account for rater differences. However, in practice teachers are typically observed by only one rater. Therefore, it is important to consider the implications of the low level of perfect agreement across different raters, take steps to minimize rater differences across teacher observations, and include subject matter experts in the teacher observation process. A lack of agreement may indicate the absence of a shared understanding about important constructs and the evidence required to indicate proficient implementation. Therefore, taking steps to develop understanding and agreement about the practices being evaluated and to provide examples of proficiency is recommended (Gitomer et al., 2014). This recommendation further emphasizes the need for well-designed observation protocols that can support the development of common understandings.

Evidence Based Reading Instruction for SWD

The goal of reading is to construct meaning from written text. This is a complicated endeavor requiring an understanding of the alphabetic principle, the skills to

identify words and connect them to meaning, and the ability to draw on multiple cognitive skills and processes (Cain et al., 2004; Oakhill & Cain, 2007). In 2000 the National Reading Panel published its widely read report identifying five key areas of reading instruction phonological awareness, phonics, vocabulary, comprehension, and fluency (NICHD, 2000). Over the past two decades the field of education has dedicated significant resources toward the effort of understanding how to effectively teach reading with a continued focus on these key areas (Castles et al., 2018). The observation protocols discussed in this dissertation focus on two of these critical components, comprehensive decoding instruction for SWD and comprehension instruction for SWD.

The Alphabetic Principle and Word Reading Instruction for SWD

Understanding that letters and letter patterns represent specific sounds in language and that these relationships are systematic and transferable is essential to learning to read (Blachman et al., 2004; Ehri et al., 2001; Foorman et al., 2003; Steacy et al., 2016). The development of this fundamental understanding referred to as the alphabetic principle requires intentional instruction making this a critical component of effective reading instruction and intervention for SWD (Blachman et al, 2004; Castles et al., 2018; Ehri et al., 2001; Forman et al., 2003; Lovett et al., 2000; Steacy et al., 2016; Torgesen et al., 2001). This awareness provides the foundation for developing accurate and fluent word reading, a skill integral to comprehending text and a key predictor of comprehension ability (Cain et al., 2004; Denton & Al Otaiba, 2011, Ehri et al., 2001; Kang & Shin, 2019).

Studies of word reading intervention for SWD support both synthetic (mapping phonemes to graphemes and blending to decode words) and analytic (recognizing larger

parts and patterns such as onset, rimes, syllables) approaches to word reading instruction (Castles et al., 2018; Denton & Al Otaiba, 2011; Ehri et al., 2001; Lovett et al., 2000; Steacy et al., 2016; Wanzek & Vaughn, 2007). Whether a synthetic, analytic, or a combination of both approaches is applied, the acquisition of word reading skills requires instruction in strategies emphasizing phonological (sound) and orthographic (written) connections (Blachman et al., 2004; Denton et al., 2013; Ehri et al., 2001; Lovett et al., 2000; Torgesen et al., 2001). In order for SWD to develop proficiency with word reading skills and strategies instruction must be intensive, systematic, and highly explicit, providing students with extended opportunities to practice (Blachman et al., 2004; Denton & Al Otaiba, 2011; Ehri, 2014). Systematic instruction teaches phoneme-grapheme correspondence and word reading skills in an ordered manner where skills build upon one another logically, providing students with the necessary prerequisite skills to be successful (Brady, 2011; Ehri et al., 2001). Routines implemented systematically within and across lessons support efficiency and lead to more fluid and focused lessons (Archer & Hughes, 2011). Instruction that is explicit provides students with the scaffolding necessary to acquire complex skills and strategies (Hughes et al., 2017) and was identified as one of 22 high leverage practices for SWD (McLeskey et al., 2017).

Performance in reading, spelling, and writing is enhanced when decoding instruction is integrated with encoding instruction explicitly reinforcing phoneme-grapheme relationships (Denton et al., 2013; Weiser, 2013). Students practice decoding skills when they blend sounds made by letters or letter groups into words. When encoding, students use their knowledge of phonemic awareness and phoneme-grapheme correspondences to transform speech into print (Moats & Hall, 2010; Weiser & Mathes,

2011). Effective encoding instruction may include writing words as well as building words or manipulating the sounds and letters in words using tiles or plastic letters (Weiser, 2013).

The importance of practicing and applying word reading skills and strategies in the context of connected text cannot be overstated. Providing students with frequent opportunities to successfully practice and apply word reading skills in the context of connected text consistently results in improved reading performance (Blachman et al., 2004; Denton et al., 2010; Jenkins et al., 2004; Mathes et al., 2005). It cannot be assumed students will naturally generalize decoding skills taught in isolation to text reading, making teacher guided practice with feedback and appropriate scaffolding an essential component of decoding instruction (Denton & Al Otaiba, 2011). This guided practice with connected text also provides the opportunity to reinforce the purpose for reading, which is reading for meaning, through intentional questioning and discussion appropriate to the text.

Comprehension - Reading for Meaning

Accurate and efficient word reading skills are essential to comprehension, but do not ensure comprehension will occur (Oakhill et al., 2003). Text comprehension requires the orchestration of multiple cognitive skills and processes, interacting with the individual's background knowledge and the content of the text (Cain, 2010; Compton et al., 2014; Kintsch, 2004; Perfetti & Stafura, 2014). Three constructs underlying comprehension of text are identified by Perfetti & Stafura (2014) in their Reading Systems Framework. The first construct is knowledge; more specifically, linguistic knowledge, orthographic knowledge and general knowledge of the world and of text

structures. The second construct involves the reading processes the reader engages in as they activate their knowledge to successfully read words, assign meaning, make inferences, and monitor their understanding. Finally, the third construct encompasses the cognitive processes involved in reading for meaning, which include executive functioning skills such as working memory, cognitive flexibility, and inhibition. These three constructs and the multiple components within each must interact and function with one another for the reader to successfully construct meaning as they engage with text (Perfetti & Stafura, 2014).

Orthographic knowledge supports word recognition, discussed in the prior section and is essential, but not sufficient for comprehension to occur. Linguistic knowledge, a critical factor in comprehending what we read, includes understanding words and having the ability to integrate their meaning into a mental model of the text (Perfetti & Stafura, 2014). Struggling readers benefit from intentional instruction and practice in building and enriching vocabulary knowledge (Bryant et al., 2003; Elleman et al., 2009). Successful approaches for developing vocabulary knowledge include direct instruction, cognitive strategy instruction, the use of mnemonics, and activity based approaches (Jitendra et al., 2004).

The reader's knowledge of text content and text structure influence their ability to effectively attend to and integrate important information, make inferences, and accurately construct meaning (Elleman & Compton, 2017; Kendeou & Van Den Broek, 2007). Poor comprehenders need support in developing their knowledge as it relates to the specific text and scaffolding to support the processes of recalling and integrating relevant information (Cain et al., 2004; Compton et al., 2014). Both general world knowledge and

prior knowledge of the passage topic are associated with stronger performance on text specific comprehension measures (Compton et al., 2013). Additionally, increasing SWD's knowledge of text structure leads to increased attention to critical elements, improves the ability to recall information and has been shown to significantly improve SWD's ability to comprehend narrative and expository texts (Alves et al., 2015; Gajria et al., 2007; Kaldenberg et al., 2015; Mason & Hedin, 2011; Stetter & Hughes, 2010; Williams, 2005). Experts in the field encourage increased attention to building knowledge relevant to the text and to text structure as part of effective intervention (Compton et al., 2014).

The ability to make inferences is essential to comprehension (Oakhill & Cain, 2007; Elleman, 2017). Making inferences refers to the processes a reader engages in to make meaningful connections between information contained within the text and also between information in the text and the reader's background knowledge (Hall & Barnes, 2017). SWD tend to have difficulty with inference making even with sufficient background knowledge and often fail to engage in this critical process altogether (Barth et al., 2015; Cain et al., 2001). Students have shown improvement through interventions designed to explicitly teach inference making processes (Elleman, 2017; Hall, 2016; Yuill & Oakhill, 1988). Features of successful inference interventions include teaching students to locate relevant information using cues in the text, introducing structured and purposeful methods for engaging background knowledge, scaffolding the process of integrating information within and across texts, and the use of advanced organizers (Elleman, 2017).

Explicitly teaching SWD strategies designed to support comprehension processes, develop student's ability to monitoring their understanding, and to use strategies flexibly across various texts improves the students' ability to actively engage with the text and effectively extract meaning (Gersten et al., 2001; Gajria et al., 2007; NICHD, 2000). Including graphic organizers and other content enhancement tools in intervention provides a framework and helps students attend to, organize, and retrieve important information (Ciullo et al., 2016; Dexter & Hughes, 2011). Purposeful instruction in strategies that increase self-monitoring and develop the skills of summarization and the identification of main idea are highly effective in improving comprehension for SWD (Gajria et al., 2007; Kim et al., 2012; NICHD, 2000; Solis et al., 2012). In conjunction with self-monitoring, students benefit from learning to support their understanding by using the text as a resource to clarify meaning or locate important information (Gardill & Jitendra, 1999; Mason 2013; Vaughn et al., 2001). Approaches to comprehension instruction integrating multiple strategies across stages of reading are supported in the comprehension instruction research (Boardman et al., 2016; Kim et al., 2012; Scammacca et al., 2016; Wanzek et al., 2016). Multicomponent intervention may include strategies scaffolding what the reader does throughout stages of the reading process such as before, during, and after reading (Boardman et al., 2016; Klingner et al., 2010; Mason, 2013).

Effective questioning strategies are an important component of intervention focused on understanding content and monitoring understanding. Thoughtfully implemented questioning leads students to attend more carefully and think systematically about what they are reading (Berkeley et al., 2010). Questioning before, during, and after reading is most effective when it is purposefully designed to encourage active

engagement and reflection, focuses the student on integrating and connecting information, and encourages the construction of meaning (McKeown et al., 2009). When designing effective intervention, teachers should carefully consider their approaches to questioning and include both thoughtful teacher directed questioning and scaffolding to support students in developing independent self-questioning strategies (Joseph et al., 2016).

References

Alves, K. D., Kennedy, M. J., Brown, T. S., & Solis, M. (2015). Story Grammar Instruction with Third and Fifth Grade Students with Learning Disabilities and Other Struggling Readers. *Learning Disabilities--A Contemporary Journal*, *13*(1).

Archer, A. L., & Hughes, C. A. (2011). Explicit Instruction: Effective and efficient teaching (what works for special-needs Learners). *Journal of Special Education*, *36*(4), 186-205.

Barth, A. E., Barnes, M., Francis, D., Vaughn, S., & York, M. (2015). Inferential processing among adequate and struggling adolescent comprehenders and relations to reading comprehension. *Reading and writing*, *28*(5), 587-609.

Berkeley, S., Scruggs, T. E., & Mastropieri, M. A. (2010). Reading comprehension instruction for students with learning disabilities, 1995—2006: A meta-analysis. *Remedial and Special Education*, *31*(6), 423-436.

Biancarosa, G., Bryk, A. S., & Dexter, E. R. (2010). Assessing the value-added effects of literacy collaborative professional development on student learning. *The elementary school journal*, *111*(1), 7-34.

Blachman, B. A., Schatschneider, C., Fletcher, J. M., Francis, D. J., Clonan, S. M., Shaywitz, B. A., & Shaywitz, S. E. (2004). Effects of intensive reading remediation for second and third graders and a 1-year follow-up. *Journal of Educational Psychology*, *96*(3), 444.

Blazar, D., Braslow, D., Charalambous, C. Y., & Hill, H. C. (2017). Attending to general and mathematics-specific dimensions of teaching: Exploring factors across two observation instruments. *Educational Assessment*, *22*(2), 71-94.

Boardman, A. G., Vaughn, S., Buckley, P., Reutebuch, C., Roberts, G., & Klingner, J. (2016). Collaborative strategic reading for students with learning disabilities in upper elementary classrooms. *Exceptional Children*, *82*(4), 409-427.

Brady, S. A. (2011). Efficacy of phonics teaching for reading outcomes: Indications from post-NRP research.

Bryant, D. P., Goodwin, M., Bryant, B. R., & Higgins, K. (2003). Vocabulary instruction for students with learning disabilities: A review of the research. *Learning Disability Quarterly*, *26*(2), 117-128.

Cain, K. (2010). Reading development and difficulties (Vol. 8). New York, NY: John Wiley & Sons.

Cain, K., Oakhill, J. V., Barnes, M. A., & Bryant, P. E. (2001). Comprehension skill, inference-making ability, and their relation to knowledge. *Memory & cognition*, *29*(6), 850-859.

Cain, K., Oakhill, J., & Bryant, P. (2004). Children's reading comprehension ability: Concurrent prediction by working memory, verbal ability, and component skills. *Journal of educational psychology*, *96*(1), 31.

Casabianca, J. M., Lockwood, J. R., & McCaffrey, D. F. (2015). Trends in classroom observation scores. *Educational and Psychological Measurement*, *75*(2), 311-337.

Castles, A., Rastle, K., & Nation, K. (2018). Ending the reading wars: Reading acquisition from novice to expert. *Psychological Science in the Public Interest*, *19*(1), 5-51.

Ciullo, S., Lembke, E. S., Carlisle, A., Thomas, C. N., Goodwin, M., & Judd, L. (2016). Implementation of evidence-based literacy practices in middle school response to intervention: An observation study. *Learning Disability Quarterly*, *39*(1), 44-57.

Ciullo, S., Lo, Y. L. S., Wanzek, J., & Reed, D. K. (2016). A synthesis of research on informational text reading interventions for elementary students with learning disabilities. *Journal of Learning Disabilities*, *49*(3), 257-271.

Compton, D. L., Miller, A. C., Elleman, A. M., & Steacy, L. M. (2014). Have we forsaken reading theory in the name of "quick fix" interventions for children with reading disability?. *Scientific Studies of Reading*, *18*(1), 55-73.

Compton, D. L., Miller, A. C., Gilbert, J. K., & Steacy, L. M. (2013). What can be learned about the reading comprehension of poor readers through the use of advanced statistical modeling techniques. *Unraveling the behavioral, neurobiological, & genetic components of reading comprehension*, 135-147.

Denton, C. A., & Al Otaiba, S. (2011). Teaching word identification to students with reading difficulties and disabilities. *Focus on exceptional children*, *2011*, 254245149.

Denton, C. A., Nimon, K., Mathes, P. G., Swanson, E. A., Kethley, C., Kurz, T. B., & Shih, M. (2010). Effectiveness of a supplemental early reading intervention scaled up in multiple schools. *Exceptional Children*, *76*(4), 394-416.

Denton, C. A., Tolar, T. D., Fletcher, J. M., Barth, A. E., Vaughn, S., & Francis, D. J. (2013). Effects of tier 3 intervention for students with persistent reading difficulties and characteristics of inadequate responders. *Journal of educational psychology*, *105*(3), 633.

Dexter, D. D., & Hughes, C. A. (2011). Graphic organizers and students with learning disabilities: A meta-analysis. *Learning Disability Quarterly*, *34*(1), 51-72.

Ehri, L. C. (2014). Orthographic mapping in the acquisition of sight word reading, spelling memory, and vocabulary learning. *Scientific Studies of Reading*, *18*(1), 5-21.

Ehri, L. C. (2004). Teaching Phonemic Awareness and Phonics: An Explanation of the National Reading Panel Meta-Analyses.

Ehri, L. C., Nunes, S. R., Stahl, S. A., & Willows, D. M. (2001). Systematic phonics instruction helps students learn to read: Evidence from the National Reading Panel's meta-analysis. *Review of educational research*, *71*(3), 393-447.

Elleman, A. M. (2017). Examining the impact of inference instruction on the literal and inferential comprehension of skilled and less skilled readers: A meta-analytic review. *Journal of Educational Psychology*, *109*(6), 761.

Elleman, A. M., & Compton, D. L. (2017). Beyond comprehension strategy instruction: What's next?. *Language, Speech, and Hearing Services in Schools*, *48*(2), 84-91.

Elleman, A. M., Lindo, E. J., Morphy, P., & Compton, D. L. (2009). The impact of vocabulary instruction on passage-level comprehension of school-age children: A meta-analysis. *Journal of Research on Educational Effectiveness*, *2*(1), 1-44.

Fallon, L. M., Collier-Meek, M. A., Maggin, D. M., Sanetti, L. M., & Johnson, A. H. (2015). Is performance feedback for educators an evidence-based practice? A systematic review and evaluation based on single-case research. *Exceptional Children*, *81*(2), 227-246.

Foorman, B. R., Breier, J. I., & Fletcher, J. M. (2003). Interventions aimed at improving reading success: An evidence-based approach. *Developmental neuropsychology*, *24*(2-3), 613-639.

Gajria, M., Jitendra, A. K., Sood, S., & Sacks, G. (2007). Improving comprehension of expository text in students with LD: A research synthesis. *Journal of learning disabilities*, *40*(3), 210-225.

Gardill, M. C., & Jitendra, A. K. (1999). Advanced story map instruction: Effects on the reading comprehension of students with learning disabilities. *The Journal of Special Education*, *33*(1), 2-17.

Gersten, R., Fuchs, L. S., Williams, J. P., & Baker, S. (2001). Teaching reading comprehension strategies to students with learning disabilities: A review of research. Review of Educational Research, 71, 279–320. doi:10.3102/00346543071002279

Gilmour, A. F., Fuchs, D., & Wehby, J. H. (2018). Are students with disabilities accessing the curriculum? A meta-analysis of the reading achievement gap between students with and without disabilities. Exceptional Children. Advanced online publication. doi:10.1177/0014402918795830

Gitomer, D., Bell, C., Qi, Y., McCaffrey, D., Hamre, B. K., & Pianta, R. C. (2014). The instructional challenge in improving teaching quality: Lessons from a classroom observation protocol. *Teachers College Record*, *116*(6), 1-32.

Hall, C. S. (2016). Inference instruction for struggling readers: A synthesis of intervention research. *Educational Psychology Review*, *28*(1), 1-22.

Hall, C., & Barnes, M. A. (2017). Inference instruction to support reading comprehension for elementary students with learning disabilities. *Intervention in School and Clinic*, *52*(5), 279-286.

Hall, E. (2014). A framework to support the validation of educator evaluation systems. *National Center for the Improvement of Educational Assessment*.

Hill, H., & Grossman P. (2013). Learning from teacher observations: Challenges and opportunities posed by new teacher evaluation systems. Harvard Educational Review, 83, 371–384.

Ho, A. D., & Kane, T. J. (2013). The Reliability of Classroom Observations by School Personnel. Research Paper. MET Project. *Bill & Melinda Gates Foundation*.

Hughes, C. A., Morris, J. R., Therrien, W. J., & Benson, S. K. (2017). Explicit instruction: Historical and contemporary contexts. *Learning Disabilities Research & Practice*, *32*(3), 140-148.

Jenkins, J. R., Peyton, J. A., Sanders, E. A., & Vadasy, P. F. (2004). Effects of reading decodable texts in supplemental first-grade tutoring. *Scientific Studies of Reading*, *8*(1), 53-85.

Jitendra, A. K., Edwards, L. L., Sacks, G., & Jacobson, L. A. (2004). What research says about vocabulary instruction for students with learning disabilities. *Exceptional children*, *70*(3), 299-322.

Johnson, E. S., Crawford, A., Moylan, L. A., & Zheng, Y. (2018). Using evidence-centered design to create a special educator observation system. *Educational Measurement: Issues and Practice*, *37*(2), 35-44.

Joseph, L. M., Alber-Morgan, S., Cullen, J., & Rouse, C. (2016). The effects of self-questioning on reading comprehension: A literature review. *Reading & Writing Quarterly*, *32*(2), 152-173.

Judge, S., & Bell, S. M. (2010). Reading achievement trajectories for students with learning disabilities during the elementary school years. *Reading & Writing Quarterly: Overcoming Learning Difficulties*, *27*(1-2), 153–78. doi:10.1080/10573569.2011.532722

Kaldenberg, E. R., Watt, S. J., & Therrien, W. J. (2015). Reading instruction in science for students with learning disabilities: A meta-analysis. *Learning Disability Quarterly*, *38*(3), 160-173.

Kang, E. Y., & Shin, M. (2019). The contributions of reading fluency and decoding to reading comprehension for struggling readers in fourth grade. *Reading & Writing Quarterly*, *35*(3), 179-192.

Kendeou, P., & Van Den Broek, P. (2007). The effects of prior knowledge and text structure on comprehension processes during reading of scientific texts. *Memory & cognition*, *35*(7), 1567-1577.

Kent, S. C., Wanzek, J., & Al Otaiba, S. (2012). Print reading in general education kindergarten classrooms: What does it look like for students at- risk for reading difficulties?. *Learning Disabilities Research & Practice*, *27*(2), 56-65.

Kim, W., Linan- Thompson, S., & Misquitta, R. (2012). Critical factors in reading comprehension instruction for students with learning disabilities: A research synthesis. *Learning Disabilities Research & Practice*, *27*(2), 66-78.

Kintsch, W. (2004). The construction-integration model of text comprehension and its implications for instruction. *Theoretical models and processes of reading*, *5*, 1270-1328.

Klingner, J. K., Urbach, J., Golos, D., Brownell, M., & Menon, S. (2010). Teaching reading in the 21st century: A glimpse at how special education teachers promote reading comprehension. *Learning Disability Quarterly*, *33*(2), 59-74.

Lane, H. B. (2014). Evidence-based reading instruction for grades K-5. *CEEDAR Document No. IC-12. Retrieved from http://ceedar. education. Ufl. edu/wp-content/uploads/2014/12/IC-12_FINAL_12-15-14. pdf.*

Lovett, M. W., Steinbach, K. A., & Frijters, J. C. (2000). Remediating the core deficits of developmental reading disability: A double-deficit perspective. *Journal of learning disabilities*, *33*(4), 334-358.

Mason, L. H. (2013). Teaching students who struggle with learning to think before, while, and after reading: Effects of self-regulated strategy development instruction. *Reading & Writing Quarterly*, *29*(2), 124-144.

Mason, L. H., & Hedin, L. R. (2011). Reading science text: Challenges for students with learning disabilities and considerations for teachers. *Learning Disabilities Research & Practice, 26*(4), 214-222.

Mathes, P. G., Denton, C. A., Fletcher, J. M., Anthony, J. L., Francis, D. J., & Schatschneider, C. (2005). The effects of theoretically different instruction and student characteristics on the skills of struggling readers. *Reading Research Quarterly*, *40*(2), 148-182.

McKeown, M. G., Beck, I. L., & Blake, R. G. (2009). Rethinking reading comprehension instruction: A comparison of instruction for strategies and content approaches. *Reading Research Quarterly*, *44*(3), 218-253.

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

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

Miller, B., McCardle, P., & Hernandez, R. (2010). Advances and remaining challenges in adult literacy research. *Journal of learning Disabilities*, *43*(2), 101-107.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). On the structure of educational assessments. Measurement: Interdisciplinary Research and Perspectives, 1, 3–67.

Moats, L. (2009). Knowledge foundations for teaching reading and spelling. *Reading and Writing*, *22*(4), 379-399.

Moats, L. C., & Foorman, B. R. (2003). Measuring teachers' content knowledge of language and reading. *Annals of Dyslexia*, *53*(1), 23-45.

Moats, L. C., & Hall, S. (2010). Language essentials for teachers of reading and spelling (LETRS®) Module 7—Teaching phonics, word study, and the alphabetic principle.

Moody, Sally Watson, et al. "Reading instruction in the resource room: Set up for failure." *Exceptional children* 66.3 (2000): 305-316.

National Center for Educational Statistics (2020, May). *The condition of education.* https://nces.ed.gov/programs/coe/indicator_cgg.asp

National Center for Education Statistics (2019). *Reading 2019: National Assessment of Educational Progress: An overview of NAEP*. Washington, D.C.: National Center for Education Statistics, Institute of Education Sciences, U.S. Dept. of Education.

National Reading Panel (US), National Institute of Child Health, Human Development (US), National Reading Excellence Initiative, National Institute for Literacy (US), United States. Public Health Service, & United States Department of Health. (2000). *Report of the National Reading Panel: Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction: Reports of the subgroups*. National Institute of Child Health and Human Development, National Institutes of Health.

Oakhill, J., & Cain, K. (2007). Issues of causality in children's reading comprehension. *Reading comprehension strategies: Theories, interventions, and technologies*, 47-71.

Oakhill, J. V., Cain, K., & Bryant, P. E. (2003). The dissociation of word reading and text comprehension: Evidence from component skills. *Language and cognitive processes*, *18*(4), 443-468.

Perfetti, C., & Stafura, J. (2014). Word knowledge in a theory of reading comprehension. *Scientific studies of Reading*, *18*(1), 22-37.

Scammacca, N. K., Roberts, G. J., Cho, E., Williams, K. J., Roberts, G., Vaughn, S. R., & Carroll, M. (2016). A century of progress: Reading interventions for students in grades 4–12, 1914–2014. *Review of Educational Research*, *86*(3), 756-800.

Schles, R. A., & Robertson, R. E. (2019). The role of performance feedback and implementation of evidence-based practices for preservice special education teachers and student outcomes: A review of the literature. *Teacher Education and Special Education*, *42*(1), 36-48.

Schulte, A. C., Stevens, J. J., Elliott, S. N., Tindal, G., & Nese, J. F. (2016). Achievement gaps for students with disabilities: Stable, widening, or narrowing on a state-wide reading comprehension test?. *Journal of Educational Psychology*, *108*(7), 925.

Snyder, P. A., Hemmeter, M. L., & Fox, L. (2015). Supporting implementation of evidence-based practices through practice-based coaching. *Topics in Early Childhood Special Education*, *35*(3), 133-143.

Solis, M., Ciullo, S., Vaughn, S., Pyle, N., Hassaram, B., & Leroux, A. (2012). Reading comprehension interventions for middle school students with learning disabilities: A synthesis of 30 years of research. *Journal of learning disabilities*, *45*(4), 327-340.

Solomon, B. G., Klein, S. A., & Politylo, B. C. (2012). The effect of performance feedback on teachers' treatment integrity: A meta-analysis of the single-case literature. *School Psychology Review*, *41*(2), 160-175.

Steacy, L. M., Elleman, A. M., Lovett, M. W., & Compton, D. L. (2016). Exploring differential effects across two decoding treatments on item-level transfer in children with significant word reading difficulties: A new approach for testing intervention elements. *Scientific Studies of Reading*, *20*(4), 283-295.

Stetter, M. E., & Hughes, M. T. (2010). Using story grammar to assist students with learning disabilities and reading difficulties improve their comprehension. *Education and treatment of children*, 115-151.

Swanson, E. A. (2008). Observing reading instruction for students with learning disabilities: A synthesis. *Learning Disability Quarterly*, *31*(3), 115-133.

Swanson, E. A., & Vaughn, S. (2010). An observation study of reading instruction provided to elementary students with learning disabilities in the resource room. *Psychology in the Schools*, *47*(5), 481-492.

Taylor, E. S., & Tyler, J. H. (2012). Can teacher evaluation improve teaching. *Education Next*, *12*(4), 78-84.

Torgesen, J. K., Alexander, A. W., Wagner, R. K., Rashotte, C. A., Voeller, K. K., & Conway, T. (2001). Intensive remedial instruction for children with severe

reading disabilities: Immediate and long-term outcomes from two instructional approaches. *Journal of learning disabilities*, *34*(1), 33-58.

Vaughn, S., Klingner, J. K., & Bryant, D. P. (2001). Collaborative strategic reading as a means to enhance peer-mediated instruction for reading comprehension and content-area learning. *Remedial and Special Education*, *22*(2), 66-74.

Vaughn, S., & Wanzek, J. (2014). Intensive interventions in reading for students with reading disabilities: Meaningful impacts. *Learning Disabilities Research & Practice*, *29*(2), 46-53.

Wanzek, J., Swanson, E., Vaughn, S., Roberts, G., & Fall, A. M. (2016). English learner and non-English learner students with disabilities: Content acquisition and comprehension. *Exceptional Children*, *82*(4), 428-442.

Washburn, E. K., Joshi, R. M., & Cantrell, E. B. (2011). Are preservice teachers prepared to teach struggling readers?. *Annals of dyslexia*, *61*(1), 21-43.

Wanzek, J., & Vaughn, S. (2007). Research-based implications from extensive early reading interventions. *School Psychology Review*, *36*(4), 541-561.

Weiser, B. L. (2013). Ameliorating reading disabilities early: Examining an effective encoding and decoding prevention instruction model. *Learning Disability Quarterly*, *36*(3), 161-177.

Weiser, B., & Mathes, P. (2011). Using encoding instruction to improve the reading and spelling performances of elementary students at risk for literacy difficulties: A best-evidence synthesis. *Review of Educational Research*, *81*(2), 170-200.

Williams, J. P. (2005). Instruction in reading comprehension for primary-grade students: A focus on text structure. *The Journal of Special Education*, *39*(1), 6-18.

Yuill, N., & Oakhill, J. (1988). Effects of inference awareness training on poor reading comprehension. *Applied Cognitive Psychology*, *2*(1), 33-45.

CHAPTER TWO: USING EVIDENCE-CENTERED DESIGN TO CREATE A SPECIAL EDUCATOR OBSERVATION SYSTEM

This chapter is published by John Wiley and Sons in *Educational Measurement: Issues and Practice* and should be referenced accordingly.

Reference:

Johnson, E. S., Crawford, A. R., Moylan, L. A., & Zheng, Y. (2018) Using Evidence-Centered Design to Create a Special Educator Observation System. *Educational Measurement: Issues and Practice,* **https://doi.org/10.1111/emip.12182**

*This chapter includes modifications from the originally published version. Modifications include format changes to meet dissertation requirements and updated citation formatting.

Using Evidence-Centered Design to Create a Special Educator Observation System

Evelyn S. Johnson, Angela Crawford, Laura A. Moylan, and Yuzhu Zheng

Boise State University

January 2018

Author Note

Evelyn S. Johnson, Department of Early and Special Education, Boise State University; Angela Crawford, Project RESET, Boise State University; Laura A. Moylan, Project RESET, Boise State University; Yuzhu Zheng, Project RESET; Boise State University.

This research was supported the Institute of Education Sciences, award number R324A150152 to Boise State University. The opinions expressed are solely those of the authors. Correspondence regarding this manuscript should be addressed to: Dr. Evelyn S. Johnson, Boise State University, 1910 University Dr., MS 1725, Boise, Idaho 83725-1725. Email: evelynjohnson@boisestate.edu

Abstract

The Evidence-Centered Design (ECD) framework was used to create a special education teacher observation system, Recognizing Effective Special Education Teachers (RESET). Extensive reviews of research informed the domain analysis and modeling stages, and led to the conceptual framework in which effective special education teaching is operationalized as the ability to effectively implement evidence-based practices for students with disabilities. In the assessment implementation stage, four raters evaluated 40 videos and provided evidence to support the scores assigned to teacher performances. An inductive approach was used to analyze the data and to create empirically derived, item level performance descriptors. In the assessment delivery stage, four different raters evaluated the same videos using the fully developed rubric. Many-facet Rasch measurement (MFRM) analyses showed that the item, teacher, lesson and rater facets achieved high psychometric quality. This process can be applied to other content areas to develop teacher observation systems that provide accurate evaluations and feedback to improve instructional practice.

Keywords: special education teacher evaluation, observation systems, Many-facet Rasch measurement

Introduction

Teacher observation systems are increasingly seen as an important component of education reform because they offer the opportunity to evaluate teaching practice and to provide teachers with feedback on how to improve instruction. Emerging analyses of teacher observation systems suggest that, when teachers are objectively evaluated and supported to improve instruction, there is a positive impact on student growth (Biancarosa et al., 2012). However, in the effort to adopt observation systems on a broad scale, many states and districts are using evaluation tools that are very generic in nature, or that have been designed primarily for accountability and therefore do not provide teachers with extensive feedback on practice (Hill & Grossman, 2013). If teacher observation systems are to fulfill their promise of improving instruction, considerable work remains to ensure that they are developed and implemented in ways that address the shortcomings of current tools.

To be useful, a teacher observation system must facilitate accountability, support growth and development of professional practice, and provide accurate, reliable ratings and feedback about the specific instructional adjustments teachers need to make (Hill & Grossman, 2013). Many observation systems however, are poorly aligned with the evidence-based instructional practices (EBPs) within the relevant content area, limiting the quality of the feedback provided to teachers through this mechanism (Grossman et al., 2009). This is especially the case for special education teachers, who are routinely evaluated with observation instruments designed for the general education setting (Johnson & Semmelroth, 2014). Additionally, large scale studies of current observation systems have indicated a propensity for bias in scores, in which the majority of teachers

are discovered to be proficient or better (Kane & Staiger, 2012). Recent state level reports confirm that in practice, the tendency for bias in teacher observation systems is significant (Farley, 2017).

Effective teacher observation systems require deliberate construction and thorough psychometric evaluation. An assessment that seeks to measure something as complex as instructional practice must be designed around the inferences that are to be made, the observations that will be used to draw these inferences, and the chain of reasoning that connects them (Messick, 1994). Evidence-Centered Design (ECD) provides a conceptual design framework to create complex, coherent assessments based on the principles of evidentiary reasoning (Mislevy et al., 2003). In brief, ECD consists of five stages: 1) domain analysis, 2) domain modeling, 3) conceptual framework, 4) assessment implementation, and 5) assessment delivery. Designing assessment products through the ECD framework ensures that the way that evidence is gathered and interpreted is consistent with the underlying construct the assessment is intended to address (Mislevy et al., 2003).

ECD has been applied to several, significant large-scale student assessment systems (Plake et al., 2010), but has not been used extensively to develop teacher observation instruments. In this manuscript, we describe the development of a special education teacher (SET) observation instrument that has been developed through the ECD framework with the goal of providing SETs clear and actionable signals about ways to improve their teaching practice, minimizing bias in the resulting evaluations, and providing reliable results across raters.

Recognizing Effective Special Education Teachers (RESET)

RESET is a federally funded project to create observation rubrics aligned with EBPs for students with high incidence disabilities. The goal is to leverage the extensive research on EBPs for this population of students to inform the development of observation instruments that provide feedback to SETs to improve their practice and ultimately, to improve outcomes for students with disabilities (SWD). To create the RESET observation system, we followed the five-stage ECD framework (Mislevy et al., 2003). Below, we describe each stage as it applies to the development of RESET, followed by a reporting of the studies undertaken to inform the assessment implementation and delivery stages.

Domain Analysis

The domain analysis stage involves collecting substantive information about the domain being assessed (Mislevy & Haertel, 2006); in this case, effective special education teaching. We reviewed the research on teacher impact to determine the salient aspects of the teacher's role in affecting student outcomes to create a definition of special education teaching. Drawing on the research on instructional practice, we identified common elements of effective instruction such as: 1) maintaining rigorous expectations; 2) creating an effective, engaging learner environment; 3) making content area knowledge relevant, and 4) providing learning experiences using effective research-based strategies (Hattie, 2009). Next, we engaged in a meta-review of the research on effective special education instructional practice, organizing our search through these four elements. Several meta-analyses of EBPs provided useful starting points for conducting our review (see for example: Bellini et al.; 2007; Berkeley et al., 2010; Gersten et al.,

2009; Swanson & Sachse-Lee, 2000). The result of this review led to a definition of effective special education teaching as the ability to assess a student's learning needs and implement EBPs to support academic and social/emotional growth.

Domain Modeling

We then moved to the domain modeling stage, in which the information and relationships identified in domain analysis are translated into assessment design options. Based on our definition of effective special education teaching, we concluded it is best assessed through observations of a SETs instruction that are evaluated using rubrics detailing the essential elements of the EBPs we expect to see in the classroom. To create assessment design options within the domain modeling stage, both characteristic and variable features are used to specify how SETs will produce performance tasks (Mislevy & Haertel, 2006). The characteristic tasks common across SETs include video recording the SET directly working with students in an instructional setting. However, because teaching contexts and instructional settings are highly variable in special education, the variable features of RESET include establishing criteria for evaluating a range of EBPs depending on the specific context in which the SET is working. SETS are responsible for providing instruction across content areas, grade levels, and various arrangements such as pull-out models or co-teaching. SETs also work with students who require specially designed instruction that is individualized depending on student need. SETs must be well-versed in numerous EBPs and be cognizant of various disability types to plan and implement effective instruction (Odom et al., 2005). Therefore, an effective SET observation system must capture a broad range of EBPs, delivered in a variety of contexts and adapted to meet individual student needs.

Conceptual Assessment Framework

With the framework developed in the domain modeling stage, we moved to create the blueprint for RESET, or the conceptual assessment framework, which is divided into models that bridge the assessment argument with the operational activities of the assessment system (Mislevy et al., 2003). The models included within RESET include the 1) teacher model; 2) evidence model, 3) task model, and 4) presentation model (Mislevy et al., 2003). The *teacher* model in RESET consists of a single variable, a SETs proficiency in the implementation of EBPs. Through our review of literature undertaken in the domain analysis stage, we organized EBPs into three main areas: 1) instructional methods, 2) content organization and delivery, and 3) individualization. Within each category, we outlined the rubrics associated with the EBPs to create an overall blueprint for RESET. The list of rubrics organized by category is included in Table 1. Through RESET, we obtain evidence that provides an estimate of a SETs proficiency to effectively deliver instruction, to organize and support content area learning, and to individualize instruction based on the students' presenting needs.

SETs submit video recordings of their lessons which are then evaluated using the appropriate rubric from each subscale. This process comprises the evidence model. The scoring rules are based on the SETs level of implementation of EBPs, and evaluated as implemented, partially implemented, or not implemented. The task model for RESET is any lesson delivered by the SET to SWD. The presentation model for RESET relies on the use of video recorded lessons and electronic versions of the relevant RESET rubrics. Observations are self-evaluated by the SET and evaluated by raters who have been trained in the use of RESET.

Assessment Implementation

The operational model derived from the conceptual assessment framework leads to the assessment implementation stage (Mislevy et al., 2003), the stage at which assessment items are created. As described above, RESET consists of a set of rubrics, each rubric reflects the items and performance-level descriptors (PLDs) for a specific EBP. To create individual items for each rubric, we conducted extensive reviews of the research on the EBPs included within RESET, then synthesized the descriptions of these practices across studies to create a set of items that detailed each EBP. To illustrate the item development process in more detail, we will use the Explicit Instruction (EI) rubric as an example (see the appendix).

A number of studies and meta-analyses have identified EI as one of the most effective approaches to teaching students with disabilities (see, e.g., Archer & Hughes, 2010; Brophy & Good, 1986; Christenson et al., 1989; Gersten et al., 2000; Rosenshine & Stevens, 1986; Swanson, 1999). We first extracted the critical elements of EI from the literature, then reviewed and synthesized them into a coherent set. Then, drawing on this review, we drafted a set of items to describe proficient implementation of EI. We refined the descriptors for proficient implementation by reviewing video recorded lessons collected from SETs, and discussing the clarity and utility of each item as written. We sent the rubric to subject matter experts for review, synthesized their feedback and completed revisions to create a set of elements that described proficient implementation of EI.

Because the purpose of RESET is to both evaluate and provide feedback to SETs, we needed to create a set of scoring rules that define and describe varying levels of

implementation (e.g. implemented, partially implemented, not implemented). Initially, we considered using general descriptor levels; however, rating scales can be imprecise when general descriptors are used (Hill & Grossman, 2013). Additionally, a key focus of ECD is to identify observable evidence to create performance-level descriptors (PLDs) that result in a transparent evidentiary argument and consistent evaluations of performance (Ewing et al., 2010). PLDs communicate what various levels of performance should look like, and serve a critical role in setting cut scores that ultimately determine the categorization of a person's performance (Ewing et al., 2010). Ewing et al (2010) describe an iterative process for articulating PLDs in which performances are mapped onto the performance continuum, with items that best target the meaning of a specific performance category as well as clearly differentiating the adjacent performance levels. An analysis is then undertaken to provide a synthesis of the salient content and skills that characterize and differentiate the categories along the performance continuum, and this analysis will reveal where more evidence may be needed to inform the PLDs. In this initial work to develop RESET, we began the process of PLD development through a study designed to create analytically developed descriptors (Knoch, 2009), with the intent in future studies to engage in the iterative process described by Ewing et al. (2010) to further refine the rubrics.

Assessment Delivery

In the assessment delivery stage items are piloted, feedback is collected, and psychometric analyses are conducted, the results of which are integrated into the final design of the assessment tool. Our primary objective was to create an observation instrument that provides reliable results across raters that provides SETs with clear and

actionable signals about legitimate ways to improve their teaching practice. Because there are multiple variables that can impact a SETs score within RESET, we employed many-facet Rasch measurement (MFRM) analysis to conduct a substantive investigation of the teacher, lesson, rater, and item facets, as well as the teacher and item difficulty. MFRM is an extension of the Rasch model that conceptualizes the expected performance of individuals as a function of their ability and the item difficulty (Smith & Kulikowich, 2004). MFRM allows us to include additional assessment variables such as rater severity into the analysis. MFRM also allows us to identify particular elements within a facet that are problematic and to conduct a bias analysis that identifies specific combinations of facet elements – particular rater-teacher combinations, for example - that are consistently different from the overall identified pattern (Eckes, 2011).

Teacher observation systems are high stakes assessments. They are used both to inform the instruction that students receive as well as to make critical decisions about teachers. To meet these demands, observation systems require a deliberate approach to development and a rigorous evaluation of their psychometric properties. The ECD framework provides a useful heuristic for creating observation systems suited for these purposes. In this review, we have described the application of the first three stages of the ECD process for creating RESET, an observation system specifically designed for SETs. Using one of the rubrics within the RESET system, the EI rubric, we now detail two studies undertaken that informed the assessment implementation and assessment delivery stages of the ECD process.

Methods

In this section, we describe two studies. The first study describes the processes undertaken to create an initial set of PLDs and the second study details the procedures used to analyze the reliability of the EI rubric.

Study 1. Performance-Level Descriptor Study.

Participants

Special Education Teachers

A total of ten special education teachers from three states provided four video recorded lessons each for a total of 40 videos. All teachers were female, with an average experience level of 11.55 years (8.46 SD). Nine teachers taught at elementary and one at the middle school level. All teachers had their special education certification, five had undergraduate degrees, and five had graduate degrees.

Raters

A total of four raters participated in the descriptor development study. Two of the raters were instructional coaches, and two were veteran special education teachers who served as department chair and lead teacher within their schools. Raters had an average of 15 years of experience. All four raters were female.

Procedures

Video Collection

During the 2015-16 school year, SETs provided weekly video recorded lessons from a consistent instructional period. Videos were recorded and uploaded

using the Swivl® capture system and ranged in length from 20 to 35 minutes. Each teacher contributed a total of 20 videos over the school year. From this video bank, four videos from each teacher were selected for inclusion in the study. To be included in the data set, videos had to have adequate video and audio quality (of the 800 total videos, 42 were found to be not usable due to poor video quality or lack of sound), and had to depict an instructional lesson for which the use of the EI rubric was applicable. If a teacher had more than four videos that met these criteria, we randomly selected four. Videos were assigned an ID number and listed in unique, random order for each rater to control for order effects.

Rater Training

Rater training took place over two days. Raters were provided with an overview of the RESET project goals, and a description of how the EI rubric was developed. Project staff then explained each item of the EI rubric and clarified any questions the raters had about the items. Then, raters watched a video that had been scored by project staff and scored the video with the EI rubric, and then the scores were reviewed and discussed to include the rationale for the score that each item received. Raters then watched and scored two videos independently, and scores were reconciled with a master coded rubric for each video. Any disagreements in scores were reviewed and discussed. Raters were then assigned a randomly ordered list of videos. Raters were asked to score each item, to provide time stamped evidence that they used as a basis for the score, and to provide a

brief explanation of the rationale for their score. Raters were given a timeframe of four weeks to complete their ratings.

Data Analysis

Performance-Level Descriptor (PLD) development.

To create the PLDs for each item, we compiled the evidence and explanations provided by the raters after they scored the videos. We used a general inductive approach (Thomas, 2006) to condense their input into themes and categories that emerged as key terms identified as influencing scoring decisions. The coding process included several phases: initial reading, identifying segments of information, labeling segments of information, creating categories, selecting categories, and creating themes. First, the evidence and explanations were reviewed until the researchers were familiar with their content and gained an understanding of the text. Then, text segments that contained meaningful units were identified. The identified segments were labeled as codes by using words, phrases, or sentences directly used in the segments to capture their key elements as closely as possible. Codes which had the same or similar key elements were grouped together to generate categories. Then, categories were selected to develop descriptors relevant to the rating scale of 1) not implemented, 2) partially implemented and 3) implemented, or (N/A) not applicable.

Several strategies were used to address the trustworthiness of the item level descriptors including consistency checking, peer debriefing, and stakeholder checking. Consistency checking involved independent parallel coding by two researchers (Thomas, 2006). Two researchers analyzed the raters' evidence and

explanations, then compared their analysis until they reached consensus in codes, categories, and descriptors. Peer debriefings were conducted with the research team (Creswell & Miller, 2000). The RESET team reviewed the codes and categories while referring to the evidence and explanations of raters, and participated in consensus building of descriptors. Stakeholder checking was conducted by requesting teachers and raters to review the descriptors. The researchers also kept procedural and analytic memos about the meaning of the data (Esterberg, 2002). The end result of this extensive process was a full set of descriptors for each item, a revision of the item descriptors for 'implemented' and paring down the number of items from 27 to 25. The final rubric is in the appendix.

Study 2. Many-facet Rasch Measurement Analysis

Participants

Special education teachers

The same teacher participants from Study 1 participated in Study 2.

Raters

A total of four raters participated in the MFRM reliability analysis study. One rater was a post-doctoral researcher, one a school-psychologist and RTI coordinator in her school, one a special education faculty member, and the fourth a special education teacher completing graduate studies in special education. Raters had an average of 17 years' experience. Three raters were female and one was male.

Procedures

Video collection

The same video set from Study 1 was used during Study 2.

Rater training

Rater training was conducted as described in Study 1, with the exception that raters in this study were trained using the fully developed EI rubric with PLDs for each item.

MFRM Analyses

We analyzed the data collected by the raters using the fully developed EI rubric through MFRM analyses. The raw scores assigned to the EI rubric are ordinal, making valid comparisons between teachers or items difficult, as equal raw score differences between pairs of points do not imply equal amounts of the construct under investigation (Smith & Kulikowich, 2004). With Rasch models, the ability estimates of teachers are freed from the distributional properties of the items, and the particular raters used to rate the performance (Eckes, 2011). Additionally, the estimated difficulty of items and severity of raters are freed from the distributional properties of the other facets of the assessment (Smith & Kulikowich, 2004). The model used for this analysis is given by:

$$ln\left(\frac{P_{nijok}}{P_{nijo(k-1)}}\right) = B_n - D_i - C_j - T_o - F_k$$

where P_{nijok} is the probability of teacher *n*, when rated on item *i* by judge (rater) *j* on occasion (lesson) *o*, being awarded a rating of *k*. $P_{nijo(k-1)}$ is the probability of teacher *n*, when rated on item *i* by judge *j* in occasion *o*, being awarded a rating of *k*-1, B_n is the ability of teacher *n*, D_i is the difficulty of item *i*, C_j is the severity

of judge *j*, T_o is the stringency of occasion *o*, and F_k is the extra difficulty overcome in being observed at the rating *k* relative to the rating *k*-1 (Eckes, 2011).

The MFRM analysis was conducted using the computer program FACETS version 3.71 (Linacre, 2014). MFRM analysis produces infit and outfit statistics for each facet, two quality control statistics that indicate whether the measures have been confounded by construct-irrelevant factors (Eckes, 2011). Ranges in fit statistics from .5 to 1.5 are considered acceptable (Eckes, 2011; Englehard, 1992). In addition to measures of fit, FACETS also provides reliability and separation indices. The reliability index indicates the reproducibility of the measures if the test were to be administered to another randomly selected sample from the same population (Bond & Fox, 2007). Separation indicates the number of statistically distinguishable strata in the data. Finally, MFRM allows for bias analysis of the scores to examine the discrepancy between observed and expected scores according to the severity levels of the raters. In this study, the biased interactions between teachers and raters were examined. Significant differences between expected and observed scores ($p < .05$) indicate the presence of bias (Eckes, 2011; Linacre, 2014b).

Results

Data collected from the raters who used the fully developed EI rubric was analyzed with the FACETS (Linacre, 2014a) program. The results of the analysis are shown in Figure 1 and Tables 2-6. Figure 1 includes the variable map and rank order of each facet. Tables 2-5 report the fit statistics and reliability and separation indices for each of the facets. Bias analysis results are reported in Table 6. All analyses are based on a total of 3952 observations. Category statistics showed that of the 3952 assigned scores,

51% were a 3 (implemented), 33% were a 2 (partially implemented) and 16% were a 1 (not implemented). Only 1% of items received an N/A.

The far-left column of Figure 1, titled "Measr," is the logit measure for the elements within each facet of the design. The second column contains the item measures, with more difficult items having larger logit values. Items 3, 13 and 12 were the most difficult, and items 21, 5, and 19 the least. Examining the items on the EI rubric (see the appendix), the rank order of items is logical. For example, items 12 and 13 require teachers to task analyze and to deliver instruction in ways that support the individual needs of their students. This is a difficult skill that likely develops over time and with training. Items that were the least difficult included #5, alignment of instruction to the stated goal, which, if the teacher is using an evidence-based program to guide her instruction, will meet this criterion. Additionally, item 19 focuses on providing students with opportunities to respond. Low teacher-student ratios may make implementing this item significantly easier than it might be in larger classrooms.

The third column contains the teacher facet, with more proficient teachers having higher logit values. Teacher 9 is the most proficient teacher (proficiency = 1.64 logits, *SE* = .10), and teacher 10 is the least proficient (proficiency = -.17 logits, *SE* = .08). The fourth column contains the lesson facet. In our data collection design the rank ordering of the lesson facet is somewhat difficult to interpret, because we did not specify the content or focus of the lessons but instead had the teachers select which lessons to submit. The fifth column contains the rater facet, with more severe raters having higher logit values. Rater 4 was our most severe rater (severity = .49 logits, *SE* = .05), and Rater 1 our most lenient (severity = -.64 logits, *SE* = .06).

Tables 2-5 report the fit statistics and the reliability and separation indices for the item, teacher, rater, and lesson facets. For all facets, all fit statistics fell within .8 to 1.2, which are within acceptable levels (Eckes, 2011). In addition to the fit statistics, reliability and separation information indices are reported. For items, the reliability coefficient was .97, separation = 5.62; for teachers, the reliability coefficient was .98, separation = 7.39. These statistics demonstrate reliable differences in item difficulty and teacher proficiency. For lessons, the reliability coefficient was .93, separation = 3.72, showing a discrimination across lessons, but lessons 1 and 2 have almost the same logits, providing some indication that we may be able to obtain reliable ratings with just three lessons instead of four. The reliability coefficient for raters was .98, separation = 9.07, suggesting differences in rater severity. The bias analysis (Table 6) indicated that a total of 31.5% of the variance in the observations (n = 3952) was explained by the model. 5.54% was explained by teacher/rater interactions, with 3.55% explained by teacher/lesson interactions, leaving 59.42% of the variance remaining in residuals. Table 6 presents only the rater/teacher pairs that showed bias and reports observed and expected scores, bias size in logits, t value and its probability. Of 40 possible teacher/rater interactions, 23 are biased. Teacher 3 was the only teacher with no biased interactions. Rater 4 had the fewest number of interactions. There was almost an even number of negative bias (n = 11) as positive (n = 12) interactions, with no clear pattern attributable to a specific teacher, rater or teacher/rater pair. As a whole, despite the presence of biased pairs, the EI rubric does not appear to exhibit a great deal of bias and the overall MFRM results suggest the facets function effectively.

Discussion

ECD is a framework that can guide efforts to create assessment systems that measure the complex construct of teaching, the inferences to be made about a teacher's ability to implement instruction, the observations that will be used to draw these inferences, and the chain of reasoning that connects them (Messick, 1994). In this manuscript, we described how the ECD framework was applied to create RESET, a special education teacher observation system (Johnson et al., 2016). The process described can be applied to other content areas to develop observation instruments of the caliber needed to realize the goal of improving practice.

We used a rigorous process in the assessment implementation stage that included having expert raters provide the evidence and rationale they used to assign scores. Then we created detailed performance level descriptors for each item. In the assessment delivery stage, we tested these descriptors with another set of raters to evaluate how well the EI rubric functioned. Through MFRM analyses, we were able to assess the reliability of the rubric and review how the various facets of the observation tool function.

Overall, our analyses provide strong evidence that we have created a rubric that will provide consistent evaluations of a SETs ability to implement EI. The psychometric reliability of items and teacher ability measures is supported by high reliability and separation statistics. That is, the RESET EI rubric reliably divided the items and teachers into statistically different strata, indicating the sensitivity of the instrument (Wright & Stone, 1999).

Although the results of the studies reported in this manuscript are promising for the continued development of the RESET observation rubrics, there are several

limitations that warrant caution in interpreting the results. The most significant limitation is that the sample sizes of both special education teachers (n = 10) and raters (n = 8 total) are small, and also limited in their representativeness of the larger population of special education teachers and potential raters. The benefit of using video observations however, is that over time, we can develop a video bank that will include a larger pool of teachers. Continued studies with larger samples of teachers and raters will be needed to verify the results of the studies reported in this manuscript.

A second limitation in the study reported here includes the process used to develop PLDs. Although we collected a significant amount of evidence from raters during our first study to inform descriptor development, within the process of ECD, the identification of claims and evidence to create PLDs should be iterative, with the goal of creating a transparent evidentiary argument (Huff et al., 2010). Future studies that continue this cycle of generating evidence and applying the mapping process to ensure that score interpretations are well-matched with the evidence and resulting PLDs are needed to further refine the RESET observation rubrics (Ewing et al., 2010; Plake et al., 2010).

Finally, scores provided on observation systems are a function not only of the teachers' ability but also of the severity of the rater evaluating them. Our analyses indicate that raters differed in their severity, but that the fit statistics for raters were within acceptable levels, suggesting no evidence of halo effects or noisy scoring. One advantage of using MFRM to analyze rater behavior is that it can account for differences in rater severity by adjusting the observed score and computing a fair score for teachers. This is different than other approaches to examining rater behavior that expect raters to

function as scoring machines, achieving perfect agreement against a master set of scores (Eckes, 2011). Research on rater behavior however, suggests that achieving perfect agreement across human raters who judge complex performances is an elusive goal and that acknowledging that raters will differ in their severity but can be trained to be consistent in their own scoring may be a more attainable reality (Eckes, 2011). The training provided to our raters appears to have achieved this goal, but further studies examining whether these findings will hold when raters who will likely serve as evaluators but who have less experience in special education (e.g. principals) are needed.

Despite these limitations, the results of our current analyses are promising. To fully realize the benefit of the RESET observation system, continued research on a variety of assessment aspects is needed. For example, the processes described in this manuscript must be applied to the other rubrics within the RESET system. Given the focus of RESET on improving teacher performance, we will also need to examine the impact of feedback and self-evaluation. Finally, teacher performance on RESET will need to be connected to student measures. A significant amount of research is needed to fully inform the development of teacher observation systems, but the ECD process is a useful blueprint for this undertaking.

Conclusion

Teacher observation systems are high stakes assessments. They are expected to significantly impact teacher behavior in ways that will lead to improved instruction and greater student gains. To achieve this vision, teachers must be held accountable through evaluation systems expressly designed for this purpose.

The development of RESET has been guided by the ECD framework to respond to the need for better teacher observation tools. Through adherence to the five stage process, we have adequately modeled the domain of effective special education teaching, created a conceptual assessment framework based on the research, and devised assessment items that reflect EBPs, result in reliable evaluations of teacher implementation, and are at a grain size sufficient to provide actionable feedback. Next steps in the process include collecting validity evidence for RESET through studies that examine the impact of receiving feedback, and studies that correlate teacher performance to student growth. The processes undertaken to create RESET could be applied to create observation systems across other content areas to support the improvement of instructional practice.

References

Archer, A. L., & Hughes, C. A. (2011). *Explicit instruction: Effective and efficient teaching*. Guilford Press: New York.

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. (2010). Reading Comprehension Instruction for Students With Learning Disabilities, 1995—2006: A Meta-Analysis. *Remedial and Special Education*, *31*(6), 423-436.

Biancarosa, G., Bryk, A. S., & Dexter, E. R. (2010). Assessing the value-added effects of literacy collaborative professional development on student learning. *The Elementary School Journal, 111,* (1), 7-34.

Bond, T. G., & Fox, C. M. (2007). Fundamental measurement in the human sciences. *Chicago, IL: Institute for Objective Measurement*.

Brophy, J., & Good, T.L. (1986). Teacher behavior and student achievement. In M.C. Wittrock (Ed.), *Handbook of research on teaching* (3rd ed.). New York: Macmillan.

Christenson, S. L., Ysseldyke, J. E., & Thurlow, M. L. (1989). Critical instructional factors for students with mild handicaps: An integrative review. *Remedial and Special Education, 10*(5), 21-31.

Creswell, J. W., & Miller, D. L. (2000). Determining validity in qualitative inquiry. *Theory into practice*, *39*(3), 124-130.

Eckes, T. (2011). *Introduction to many-facet rasch measurement*. Frankfurt: Peter Lang.

Engelhard Jr, G. (1992). The measurement of writing ability with a many-faceted Rasch model. *Applied Measurement in Education*, *5*(3), 171-191.

Esterberg, K. G. (2002). *Qualitative methods in social research*. London: McGraw-Hill.

Ewing, M., Packman, S., Hamen, C., & Thurber, A. C. (2010). Representing targets of measurement within Evidence-Centered Design. *Applied Measurement in Education, 23 (4)* 325-341. doi:10.1080/08957347.2010.510959.

Farley, A. N. (2017). Review of For Good Measure? Teacher Evaluation Policy in the ESSA Era. Boulder: National Education Policy Center. http://nepc.colorado.edu/

Gersten, R., Schiller, E. P., & Vaughn, S. (2000). *Contemporary special education research.* Mahwah, NJ: Erlbaum.

Gersten, R., Chard, D. J., Jayanthi, M., Baker, S. K., Morphy, P., & Flojo, J. (2009). Mathematics instruction for students with learning disabilities: A meta-analysis of instructional components. *Review of Educational Research*, *79*(3), 1202-42.

Grossman, P., Compton, C., Igra, D., Ronfeldt, M., Shahan, E., & Williamson, P. (2009). Teaching practice: A cross-professional perspective. *Teachers College Record, 111*(9), 2055-2100.

Hattie, J. A. C. (2009). Visible learning: A synthesis of 800+ meta-analyses on achievement. *Abingdon: Routledge*.

Herlihy, C., Karger, E., Pollard, C., Hill, H. C., Kraft, M. A., Williams, M., & Howard, S. (2014). State and local efforts to investigate the validity and reliability of scores from teacher evaluation systems. *Teachers College Record*, *116*(1), 1-28.

Hill, H., & Grossman P. (2013). Learning from teacher observations: Challenges and opportunities posed by new teacher evaluation systems. *Harvard Educational Review, 83,*(2), 371-384.

Huff, K., Steinberg, L., & Matts, T. (2010). The promises and challenges of implementing evidence-centered design in large-scale assessment. *Applied Measurement in Education, 23 (4),* 310-324. doi: 10.1080/08957347.2010.510956

Johnson, E. S., Crawford, A., Moylan, L. A., & Zheng, Y. (2016). *Recognizing Effective Special Education Teachers: Technical Manual*. Boise: Boise State University.

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

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

Knoch, U. (2009). Diagnostic assessment of writing: A comparison of two rating scales. *Language Testing*, *26*(2), 275-304.

Linacre, J. M. (2014a). *Facets 3.71. 4* [Computer software].

Linacre, J. M. (2014b). *A user guide to Facets, Rasch-model computer programs*. Chicago, IL: Winsteps.com.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher, 23*(2), 13-23.

Mislevy, R. J., & Haertel, G. D. (2006). Implications of evidence-centered design for educational testing. *Educational Measurement: Issues and Practice*, *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*, 3-67.

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

Plake, B. S., Huff, K., & Reshetar, R. (2010). Evidence-centered assessment design as a foundation for achievement-level descriptor development and for standard setting. *Applied Measurement in Education*, *23*(4), 342-357.

Rosenshine, B., & Stevens, R. (1986). Teaching functions. In M.C. Wittrock (Ed), *Handbook of research on teaching, 3rd ed (pp. 376-391)*. New York: Macmillan.

Smith Jr, E. V., & Kulikowich, J. M. (2004). An application of generalizability theory and many-facet Rasch measurement using a complex problem-solving skills assessment. *Educational and Psychological Measurement*, *64*(4), 617-639.

Swanson, H. L., (1999). Instructional components that predict treatment outcomes for students with learning disabilities: Support for a combined strategy and direct instruction model. *Learning Disabilities Research & Practice, 14,* 129-140.

Swanson, H. L., 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.

Thomas, D. R. (2006). A general inductive approach for analyzing qualitative evaluation data. *American journal of evaluation*, *27*(2), 237-246.

Wright, B. D., & Stone, M. H. (1999). Measurement essentials. *Wilmington. Wide Range Inc*, 221.

Table 2.1 Organization and Structure of RESET

Subscale	Content Area	Rubrics
Instructional Methods	N/A	Explicit Instruction
		Cognitive Strategy Instruction
		Peer Mediated Learning
Content Organization and Delivery	Reading	Letter Sound Correspondence
		Multi-Syllabic Words and Advanced Decoding
		Vocabulary
		Reading for Meaning
		Comprehension Strategy Instruction
		Comprehensive Reading Lesson
	Math	Problem Solving
		Conceptual Understanding of: Number Sense & Place Value, Operations, Fractions, Algebra
		Procedural Understanding of: Number Sense & Place Value, Operations, Fractions, Algebra
		Automaticity
	Writing	Spelling
		Sentence Construction
		Self Regulated Strategy Development
		Conventions
Individualization		Executive Function/Self-Regulation
		Cognitive Processing Accommodations
		Assistive Technology
		Duration/Frequency/Intensity

Table 2.2 Item Measure Report from Many-Facet Rasch Measurement Analysis

Item Number	Difficulty (Logits)	Model SE	Infit MNSQ	Outfit MNSQ
19	-1.61	.20	.81	.85
5	-.99	.16	.81	.80
21	-.80	.15	.86	1.03
18	-.72	.15	.83	.90
23	-.69	.14	.91	.84
6	-.53	.14	1.12	1.16
17	-.53	.14	.89	.91
4	-.48	.14	.77	.82
22	-.44	.14	.86	.87
10	-.39	.13	1.04	1.02
14	-.20	.13	1.11	1.09
20	-.15	.13	1.11	1.04
16	-.01	.12	.98	1.00
1	.16	.12	1.23	1.26
15	.20	.13	.84	.82
24	.34	.12	.91	.97
7	.38	.12	.93	.95
9	.38	.12	.97	.95
2	.44	.12	1.32	1.34
8	.47	.12	.96	.95
11	.49	.12	.95	.93
25	.60	.12	.92	.92
12	.93	.12	.92	.89
13	1.30	.12	1.11	1.09
3	1.86	.13	1.38	1.52
Mean (count = 25)	.00	.13	.98	1.00
SD	.76	.02	.16	.17

Note. Root mean square error (model) = .13; adjusted *SD* = .75; separation = 5.62; reliability = .97; fixed chi-square = 714.4; df = 24; significance = .00.

Table 2.3 Teacher Measure Report from Many-Facet Rasch Measurement Analysis

Teacher Number	Ability (Logits)	Model SE	Infit MNSQ	Outfit MNSQ
10	-.17	.08	.87	.97
3	.26	.07	.86	.89
1	.27	.08	.95	.93
4	.50	.08	1.16	1.10
8	.78	.08	1.10	1.06
7	.79	.08	.90	.88
5	1.29	.09	1.03	1.16
6	1.42	.09	1.07	1.02
2	1.52	.09	.94	.83
9	1.64	.10	1.14	1.12
Mean (count = 10)	.83	.08	.1.00	1.00
SD	.62	.01	.11	.11

Note. Root mean square error (model) = .08; adjusted *SD* = .61; separation = 7.39; reliability = .98; fixed chi-square = 492.7; df = 9; significance = .00.

Table 2.4 Lesson Measure Report from Many-Facet Rasch Measurement Analysis

Lesson Number	Difficulty (Logits)	Model SE	Infit MNSQ	Outfit MNSQ
3	-.26	.05	.99	.97
4	-.04	.05	1.02	1.05
1	.15	.05	1.04	1.04
2	.16	.05	.93	.93
Mean (count = 4)	.00	.05	.99	1.00
SD	.20	.00	.05	.06

Note. Root mean square error (model) = .05; adjusted *SD* = .19; separation = 3.72; reliability = .93; fixed chi-square = 43.4; df = 3; significance = .00.

Table 2.5 Rater Measure Report from Many-Facet Rasch Measurement Analysis

Rater Number	Severity (Logits)	Model SE	Infit MNSQ	Outfit MNSQ
1	-.64	.06	.84	.96
2	-.03	.05	1.17	1.13
3	.17	.05	.92	.85
4	.49	.05	1.02	1.05
Mean (count = 4)	.00	.05	.99	1.00
SD	.48	.00	.14	.12

Note. Root mean square error (model) = .05; adjusted *SD* = .47; separation = 9.07; reliability = .98; fixed chi-square = 232.7; df = 3; significance = .00.

Table 2.6 Bias Analysis Results – Teacher x Rater Interaction

Teacher - Rater	Observed Score	Expected Score	Bias Size	*t*	*p*
1 – 3	158	205.23	-1.06	-6.65	.000
10 – 3	157	184.68	-.65	-4.02	.000
5 – 2	234	255.21	-.57	-3.66	.000
4 – 2	188	212.15	-.56	-3.73	.000
7 – 3	205	228.46	-.52	-3.54	.000
2 – 1	266	277.13	-.48	-2.50	.014
6 – 4	222	241.33	-.46	-3.08	.002
5 – 1	254	264.52	-.42	-2.23	.027
8 – 3	210	228.01	-.40	-2.73	.007
8 – 1	245	258.38	-.39	-2.38	.019
9 – 4	234	247.21	-.35	-2.22	.028
8 – 4	228	213.68	.32	2.11	.037
7 – 2	250	236.67	.35	2.08	.039
1 – 4	206	188.70	.37	2.54	.012
1 – 2	231	211.43	.46	2.89	.004
8 – 2	247	229.91	.48	2.71	.008
10 – 2	197	176.89	.48	3.07	.002
7 – 1	272	258.70	.49	2.37	.019
4 – 1	262	242.15	.69	3.33	.001
6 – 3	274	253.03	.76	3.53	.000
2 – 3	277	256.34	.80	3.55	.000
9 – 3	287	260.03	1.33	4.58	.000
5 – 3	286	248.45	1.60	5.69	.000

Note. Observed and expected scores are based on the total possible number of points (300) across the observed count of items (100 = 25 items x 4 lessons).

Measr	-Items	+Teacher		-Lesson	-Raters	Scale
2						(3)
	3					
		Teacher 9				
		Teacher 2				
		Teacher 6				
	13	Teacher 5				
1						-------
	12					
		Teacher 7	Teacher 8			
	25					
	11, 8	Teacher 4			4	
	2, 9, 7					
	24	Teacher 1	Teacher 3			
	15, 1			2	3	
				1		
0	16			4	2	2
	20					
	14	Teacher 10				
				3		
	10,22					
	4, 17, 6					
					1	
	23, 18					
	21					
-1	5					-------
	19					
-2						1

Figure 2.1. Variable map of the RESET facets items, teachers, lessons and raters.

CHAPTER THREE: DEVELOPING A COMPREHENSION INSTRUCTION OBSERVATION RUBRIC FOR SPECIAL EDUCATION TEACHERS

This chapter is published by Taylor and Francis in *Reading and Writing Quarterly: Overcoming Learning Difficulties* and should be referenced accordingly.

Reference:

Johnson, E. S., Moylan, L. A., Crawford, A., & Zheng, Y. (2019). Developing a Comprehension Instruction Observation Rubric for Special Education Teachers. *Reading & Writing Quarterly*, 1-19.

*This chapter includes modifications from the originally published version. Modifications include format changes to meet dissertation requirements and updated citation formatting.

Developing a Comprehension Instruction Observation Rubric

Evelyn S. Johnson, Laura A. Moylan, Angela Crawford and Yuzhu Zheng

Boise State University

June 2018

Author Note

Evelyn S. Johnson, Department of Early and Special Education, Boise State University; Laura A. Moylan, Project RESET, Boise State University; Angela Crawford, Project RESET, Boise State University; Yuzhu Zheng, Project RESET; Boise State University.

This research was supported the Institute of Education Sciences, award number R324A150152 to Boise State University. The opinions expressed are solely those of the authors. Correspondence regarding this manuscript should be addressed to: Dr. Evelyn S. Johnson, Boise State University, 1910 University Dr., MS 1725, Boise, Idaho 83725-1725. Email: evelynjohnson@boisestate.edu

Abstract

In this study, we developed a Reading for Meaning special education teacher observation rubric that details the elements of evidence-based comprehension instruction and tested its psychometric properties using many-faceted Rasch measurement (MFRM). Video observations of classroom instruction from 10 special education teachers across three states during the 2015-16 school year were collected. External raters (n=4) were trained to observe and evaluate instruction using the rubric, and assign scores of 'implemented', 'partially implemented' or 'not implemented' for each of the items. Analyses showed that the item, teacher, lesson and rater facets achieved high psychometric quality for the instrument. Teacher performance was consistent with what has been reported in the literature. Implications for research and practice are discussed.

Keywords: special education teacher evaluation, reading comprehension, Many-facet Rasch measurement

Introduction

A critical outcome of school is proficient reading comprehension (National Institute of Child Health & Human Development, 2000). However, students with high incidence disabilities (SWD) tend to have significant achievement gaps in comprehension when compared to their peers in general education, and these gaps persist over time (Judge & Bell, 2010; Schulte et al., 2016, Vaughn & Wanzek, 2014; Wei et al., 2011). One potential explanation for this gap is the lack of evidence-based comprehension instruction provided to SWD. Observational studies of classroom practices consistently conclude that the quality of reading instruction in both general and special education settings is inadequate to meet the intensive instructional needs to support comprehension growth for students with reading disabilities (Klingner et al., 2010; Swanson, 2008; Vaughn & Wanzek, 2014). Inadequate instruction has been defined by (a) the limited amount of time that students actually spend reading (Kent et al., 2012; Vaughn et al., 2002); (b) the limited opportunity for active response and an emphasis on passive learning (Wanzek et al., 2014); and (c) the low quality of comprehension instruction (Swanson & Vaughn, 2010).

One way to improve reading instruction is to create a teacher observation instrument aligned with the instructional practices found to improve comprehension for SWD. Emerging analyses of general teacher observation systems suggest that when teachers are objectively evaluated and supported to improve instruction, there is a positive impact on student growth (Biancarosa et al., 2010; Taylor & Tyler, 2012). To impact instructional practice, an evaluator must be able to use an observation instrument to provide accurate, reliable ratings and feedback about the specific instructional

adjustments teachers need to make (Hill & Grossman, 2013). Many observation systems however, are very generic, limiting the quality and consistency of the feedback evaluators provide to teachers (Blazar et al., 2017; Grossman et al., 2009). This is especially the case for special education teachers, who are routinely evaluated with observation instruments designed for the general education setting (Johnson & Semmelroth, 2014).

Recognizing Effective Special Education Teachers (RESET) Reading for Meaning Rubric

The RESET *Reading for Meaning* rubric was designed to address the need for a more specific instructional observation tool that supports teachers' ability to improve reading comprehension instruction for SWD. The process of rubric development began with a synthesis of the research on effective comprehension instruction. One challenge with developing the Reading for Meaning rubric is that in order to create items that are relevant across multiple contexts and grade levels, the salient characteristics of this instructional practice needed to be reflected in a way that is both program and setting agnostic. An additional challenge with comprehension instruction is that there are a variety of instructional practices described in the research, including a recent meta-analysis suggesting that *multi*-component instructional strategies are more effective than single strategy approaches (Scammacca et al., 2016). Therefore, rather than creating multiple rubrics each detailing a specific approach to teaching reading comprehension, the key elements of effective reading comprehension instruction were identified and synthesized to create the Reading for Meaning rubric. Support for instructional practices that integrated strategies across five main areas were found: 1) comprehension strategies, 2) knowledge of text structures and features, 3) vocabulary, 4) developing background

knowledge, and 5) making inferences. In the following section we briefly review each of these areas. The complete list of studies used to inform the rubric is available at https://education.boisestate.edu/reset.

Comprehension Strategy Instruction

We began the review with the comprehension research synthesized by the National Reading Panel (NRP; NICHD, 2000), which was driven by a cognitive conceptualization of reading; the theory that readers actively and purposefully integrate prior knowledge, knowledge of text, and the content of the text to construct meaning. Two primary recommendations for teaching comprehension strategies based on this theory were included in the NRP executive summary: 1) comprehension can be improved through the *explicit* teaching of comprehension skills and strategies; and 2) teachers should be trained to teach and flexibly apply *multiple* strategies as dictated by the nature of the text (NICHD, 2000). Examples of the comprehension strategies to be taught include summarization, the use of graphic organizers and other content enhancement tools designed to structure and organize information, questioning strategies and comprehension monitoring. Highly effective strategies for SWD include identification of main idea, summarization and self-monitoring (Solis et al., 2012). The purposeful use of content enhancement tools provides students with a framework that helps them attend to, organize and retrieve important information (Ciullo et al., 2016). Content enhancement tools aligned with the text structure scaffold the reader's use of important information and support understanding and memory (Gersten at al., 2001; Kim et al., 2012).

Metacognitive strategies such as rereading, looking back in the text to locate important information, and using the text as a resource to clarify understandings are

critical scaffolds to support understanding (Englert & Mariage 1991; Gardill & Jitendra 1999; Mason, 2013; Vaughn et al., 2001). Strategy instruction has been found to be most effective when it includes practice to transfer strategies across texts (Gersten et al., 2001). A significant body of research supports the use of these strategies for SWD (Berkeley et al., 2010; Ciullo et al., 2016; El Zein et al., 2014; Gajria et al., 2007; Kim et al., 2012).

Text Structures

Students with learning disabilities have little awareness of text structures whether for narrative or expository text, and this lack of awareness leads to difficulties using text structure to facilitate comprehension (Williams et al., 2014). Text previews allow the teacher to engage background knowledge, assess what students already know, establish a framework for learning new information, and familiarize students with the text structure (Honig et al., 2000). Explicit instruction on text structures (e.g. story maps for narrative text) has been found to significantly support SWD's ability to comprehend both narrative and expository text (Alves et al., 2015; Gajria et al., 2007; Kaldenberg et al., 2015; Mason & Hedin, 2011; Stetter & Hughes, 2010). Knowledge of text structures leads students to focus their attention, to ask relevant questions, and to recall more of the information (Williams, 2005).

Vocabulary

The importance of vocabulary knowledge in reading comprehension is well documented (e.g. Nagy, Anderson & Herman, 1987; NICHD, 2000; Perfetti & Stafura, 2014). Differences in the amount of independent reading, a lack of strategies to learn words from context, and a limited knowledge of words or lexical quality (Perfetti, 2007), are significant obstacles to vocabulary development for students with learning disabilities

(Jitendra et al., 2004). Vocabulary instruction, including direct instruction, cognitive strategy instruction and morphological processing, has been shown to increase both vocabulary knowledge and comprehension, especially for struggling readers (Bryant et al., 2003; Elleman et al., 2009; Elleman et al., 2017; Jitendra et al., 2004; O'Connor et al., 2017). For SWD, it is often the case that readers have limited knowledge relevant to the text, which requires the teacher to build vocabulary, text structure and content knowledge *prior* to reading (Compton et al., 2014). As is the case with comprehension, effective vocabulary instruction relies on the use of multiple strategies (NICHD, 2000).

Background Knowledge

Background knowledge has been demonstrated to be highly predictive of comprehension ability (Catts & Kamhi, 2017; Compton et al., 2014; Elleman & Compton, 2017; Kendeou & van den Broek, 2007; McKeown et al., 2009; Willingham, 2007). Both general and text specific knowledge (e.g. text structure, content and vocabulary) impact the reader's ability to make inferences and build a coherent mental representation that integrates text information and background knowledge (Cain, 2010; Compton et.al, 2014, Kintsch, 2004; Perfetti & Stafura, 2014). Students with high incidence disabilities typically have limited background knowledge for reading most texts, especially those in the content areas (Gersten et al., 2001). Therefore, more recent recommendations for comprehension instruction focus on content centered approaches in which texts are selected for their relevance and critical meanings, and used to support students' development of a corpus of knowledge (Catts & Kamhi, 2017). McKeown et al. (2009) demonstrated that students taught through a content-centered approach

outperformed students taught through a strategy-centered approach on measures of narrative recall and expository learning probes.

Inference making

The ability to make inferences is essential to reading comprehension (Cain & Oakhill, 2007; Elleman, 2017; Kintsch, 2005). Inference making is the process by which a reader integrates information within or across texts using background knowledge to support that which is not explicitly stated (Elleman, 2017). Poor comprehenders demonstrate difficulties with inference making (Barth et al., 2015; Cain et al., 2001), but studies of inference making interventions report moderate to large effects on general and inferential comprehension outcomes for both skilled and less-skilled readers (Elleman, 2017). Connections to relevant background knowledge and schema support the ability to make inferences (Cain et al., 2004; Hall, 2015). When students are taught to monitor their comprehension and use strategies to better understand text, inference making skills have been shown to improve (McNamara et al. 2006; Yuill & Oakhill, 1988).

Multi-Component Strategies

Across the comprehension instruction research, there is strong support for approaches that integrate multiple components (Boardman et al., 2016; Scammacca et al., 2016; Wanzek et al., 2016). Multicomponent interventions tend to employ strategies across stages of reading (e.g. before, during and after), and the combination of strategies throughout the reading process is thought to support students' achievement. Collaborative Strategic Reading (CSR; Klingner et al., 2012), represents a multicomponent reading comprehension instructional model, but there are many examples of effective, multi-component interventions across the comprehension intervention research (see O'Connor

et al., 2017; Scammacca et al., 2016). Comprehension intervention that includes a focus on content and the integration of effective questioning leads students to attend more carefully and to think more systematically about the text as it is being read (Berkeley et al., 2010). The key characteristics of effective questioning practices include that they a) encourage active, engaged, and reflective reading, b) are purposeful and well-designed, c) focus on the integration of information and active construction of meaning, and d) are clear (McKeown et al, 2009). Questions may be teacher directed, or the teacher may guide students can use self-questioning strategies (Joseph et al., 2016).

Reading for Meaning Rubric Components, Structure, and Rating

Following this review, we organized the rubric to capture the complexity of effective comprehension instruction into four components designed to follow the progression of a lesson. The components include: 1) Preparing to Read – Setting a Purpose for Reading, 2) Preparing to Read – Activating Background Knowledge and Schema, 3) Reading for Meaning and Monitoring Understanding, and 4) Teacher Questioning Practices. The *Reading for Meaning* rubric is located in Appendix A. The first and second components (items 1-6) focus on how the teacher establishes a clear purpose for reading and how the teacher engages and develops the knowledge the reader brings to the text (Snow, 2002). By establishing and maintaining a clear purpose, the reader is more likely to read intentionally and attend to critical information. The third component (items 7-15) is composed of items that align with the processes of identifying, attending to and integrating information during and after reading. The items in this component focus on providing appropriate guidance and support as students identify and attend to the main idea and important details (Jitendra et al., 2000), summarize key ideas

or critical passages (Kim et al., 2012; Solis et al., 2012) and make inferences or predictions (Cain et al., 2004; Hall, 2015). The fourth component (items 16-18) focuses on questioning practices that promote understanding and focus the reading.

Across the four components there are a total of 18 items. Each item is scored on a 3-point scale, where a 3 is proficient implementation, a 2 is partial implementation, and a 1 is not implemented. The RESET Reading for Meaning rubric is designed for use with video recorded lessons that are observed and evaluated by raters who are knowledgeable of comprehension instruction and who are trained to use the rubric (training procedures are described in the Methods section). The RESET Reading for Meaning rubric is intended to be used in two main ways, 1) to provide teachers with an objective evaluation of their ability to implement this evidence-based practice and 2) to provide feedback to teachers on specific elements of the practice. Teaching reading comprehension to SWD is critical to help close the reading achievement gap, but it is also complex. Teachers must have strong knowledge of both the content of the text and of effective strategies to facilitate comprehension. They must be able to support the use of the most effective strategy across content types, and effectively teach and model strategy use for the purpose of building understanding. However, observation studies of reading instruction indicate that in general, SWD are exposed to instruction that is inadequate for supporting strong comprehension development (Klingner et al., 2010; Swanson, 2008; Vaughn & Wanzek, 2014). The *Reading for Meaning* rubric was designed to capture the complexity of effective comprehension instruction so that teachers could receive an evaluation of their ability to implement this evidence-based instructional practice.

The Reading for Meaning rubric is a high-inference observation instrument, designed to capture a complex instructional practice and to be used by observers with high levels of expertise. As a result, it can be difficult to obtain consistent interpretation and application of the scoring criteria to observations of multiple teachers' lessons across multiple raters. In fact, the *instructional* dimensions of observation protocols are the most challenging for raters to score reliably (Bell et al. 2015, Bill and Melinda Gates Foundation, 2011; Gitomer et al, 2014). Across multiple large-scale studies of teacher observation, raters account for between 25 to 70% of the variance in scores assigned to the same lesson (Casabianca et al., 2015). Methods to improve rater reliability and consistency such as increased training and calibration requirements have been investigated, but issues persist even as raters gain experience and with ongoing calibration efforts (Casabianca et al., 2015). Research on rater behavior suggests that achieving perfect agreement across raters who judge complex performances is an elusive goal and that acknowledging that raters will differ in their severity but can be trained to be consistent in their own scoring may be a more attainable reality (Eckes, 2011; Linacre, 1994).

Many-faceted Rasch measurement (MFRM) is an approach to data analysis that recognizes and models two aspects of rater behavior: 1) severity, and 2) stochastic differences, and can investigate bias interactions among raters and other facets of the observation, such as rater/teacher interactions or rater/item interactions (Linacre, 1994). In MFRM analyses, rater behavior is captured through a "severity" parameter, and that parameter characterizes the rater in the same way that an ability parameter characterizes the teacher being evaluated, and a difficulty parameter characterizes an item of the rubric

(Linacre, 1994). MFRM also reports on the amount of error that raters display. All raters are expected to demonstrate some degree of error, but too much error threatens the validity of the measurement process (Linacre, 1994). By examining rater severity, error, and bias, MFRM analyses can provide important insights that can be used to improve rater training efforts, leading to more consistent evaluations and feedback over time (Wigglesworth, 1993). In this study, we employed MRFM analyses to examine the data and provide more information about these analyses in the methods section.

Purpose of the Current Study

Teacher observations are high stakes assessments because they are used to make critical decisions about teachers' employment status (Adnot et al., 2016), and more importantly, because they should be used to improve the quality of reading instruction that SWD receive. Given these goals, observation instruments require a deliberate approach to development and a rigorous psychometric evaluation of all facets that can impact a teacher's observed scores (e.g. items, lessons, teachers, raters). The purpose of this study therefore, was to examine the psychometric quality through MFRM analyses of the *Reading for Meaning* rubric for use as an evaluative observation instrument of a teacher's ability to effectively each reading comprehension to SWD.

Methods

Participants

Special education teachers

A total of ten special education teachers from three states (Idaho, Wisconsin, Florida) each provided three video recorded lessons for a total of 30 videos. Participating teachers were part of a larger data collection effort for the RESET rubric development process that includes 46 teachers across grade levels 2 – 8 from 3 states. Teachers were recruited by contacting state and district special education directors, who then distributed consent forms throughout their district. Inclusionary criteria included having special education teaching certification and providing regular instruction to a group or individual SWD. All participating teachers were white females and taught at the elementary school level, with an average experience level of 13.07 years (9.03 SD). Three teachers had undergraduate degrees, and seven had graduate degrees.

Raters

A total of four raters from three states (Idaho, Washington, Georgia) participated in this study. Raters were recruited through a purposive sampling technique, focused on selecting raters with deep knowledge of comprehension instruction and teacher observation. One rater held a doctoral degree in special education and literacy and works as a clinical supervisor for pre-service special education teachers at a university in the Mountain West, with 10 years total experience in the field. One rater was a special education teacher with a master's degree, 13 years of experience, and was Nationally Board Certified as an Exceptional Needs Specialist. One rater held a doctoral degree in special education and literacy and works as the district RTI coordinator in a large, urban

district in the Southeast. One rater held a doctoral degree in literacy, and works as an independent consultant with more than 35 years of experience as a special education teacher, district and state level administrator. All raters were white females.

Procedures

Video collection

During the 2015-16 school year, teachers provided weekly video recorded lessons from a consistent instructional period. Videos were recorded and uploaded using the Swivl® capture system and ranged in length from 20-60 minutes. Each teacher contributed 20 videos over the school year. From this video bank, three videos (one from the beginning, middle and end of school year) from each teacher were randomly selected by research project staff for inclusion in the study. To be included in the data set, videos had to have adequate video and audio quality, and had to depict a lesson for which the use of the *Reading for Meaning* rubric was applicable. Videos were assigned an ID number and listed in random order for each rater to control for order effects.

Rater training

Rater training consisted of four, four-hour training sessions conducted by RESET project staff. Raters were provided with an overview of the RESET project goals, and a description of how the *Reading for Meaning* rubric was developed. Project staff then explained each item of the *Reading for Meaning* rubric and clarified any questions the raters had about the items. Raters were also provided with a training manual that included a more in-depth explanation of each of the items, along with examples of observations that would be considered ‘Implemented’, ‘Partially Implemented’ or ‘Not Implemented’, scored as a 3, 2, or 1 respectively. Then, raters watched and scored a video that had been

scored by project staff. The scores were reviewed and discussed. Raters then watched and scored two videos independently, and scores were reconciled with a master coded rubric for each video. Any disagreements in scores were reviewed and discussed. Raters were then assigned a randomly ordered list of videos to control for order effects. Instead of having each rater observe every video, we created a rating scheme that allowed for the connection of ratings across all rater pairs and across teachers (Eckes, 2011). Twenty-four of the 30 videos were scored by three raters, and six of the 30 videos scored by four raters. Raters scored each item for each video, to provide time stamped evidence of what they observed and used as a basis for the score, and provided a brief explanation of the rationale for their score. Raters were given a timeframe of four weeks to complete their ratings.

Data Analysis

Data were analyzed through many-faceted Rasch measurement (MFRM) analyses. The raw scores assigned to the rubric are ordinal, making valid comparisons between teachers or items difficult, as equal raw score differences between pairs of points do not imply equal amounts of the construct under investigation (Smith & Kulikowich, 2004). With Rasch models, the ability estimates of teachers are freed from the distributional properties of the items, and the particular raters used to rate the performance (Eckes, 2011).

The model used for the MFRM analysis in this study is given by:

$$ln\left(\frac{P_{nijok}}{P_{nijo(k-1)}}\right) = B_n - D_i - C_j - T_o - F_k$$

where P_{nijok} is the probability of teacher *n*, when rated on item *i* by judge (rater) *j* on occasion (lesson) *o*, being awarded a rating of *k*. $P_{nijo(k-1)}$ is the probability of teacher *n*, when rated on item *i* by judge *j* in occasion *o*, being awarded a rating of *k*-1, B_n is the ability of teacher *n*, D_i is the difficulty of item *i*, C_j is the severity of judge *j*, T_o is the stringency of occasion *o*, and F_k is the difficulty overcome in being observed at the rating *k* relative to the rating *k*-1 (Eckes, 2011).

The MFRM analysis was conducted using the computer program FACETS version 3.71 (Linacre, 2014). MFRM analysis produces infit and outfit statistics for each facet, two quality control statistics that indicate whether the measures have been confounded by construct-irrelevant factors (Eckes, 2011). Ranges in fit statistics from .5 to 1.5 are considered acceptable (Eckes, 2011; Englehard, 1992). In addition to measures of fit, FACETS also provides reliability and separation indices. The reliability index indicates the reproducibility of the measures if the test were to be administered to another randomly selected sample from the same population (Bond & Fox, 2007). Separation indicates the number of statistically distinguishable strata in the data. Finally, MFRM allows for bias analysis of the scores to examine the discrepancy between observed and expected scores according to the severity levels of the raters. In this study, the biased interactions between teachers and raters, and between items and raters were examined. Significant differences between expected and observed scores ($p < .05$) indicate the presence of bias (Linacre, 2014).

Results

The results of the analysis are shown in Figure 1 and Tables 1 through 6. All analyses are based on a total of 1728 assigned scores. Category statistics showed that of

the 1728 assigned scores, 28% were a 3 (implemented), 31% were a 2 (partially implemented) and 41% were a 1 (not implemented).

Figure 1 includes the variable map and rank order of each facet. The far left column of Figure 1, titled "Measr," is the logit measure for the elements within each facet of the design. The second column contains the item measures, with "more difficult" items having larger logit values. Items on which teachers tended to receive low scores are considered to be more difficult than those items on which teachers tended to receive higher scores. Items 5, 9 and 8 were the most difficult, and items 15 and 16 were less difficult. Item 17 was the least difficult with a logit value of -2. Examining the items on the rubric (see Appendix), the rank order of items is logical. For example, item 5 examines the teacher's use of text preview strategies. Throughout the recorded lessons, very few teachers employed this strategy as a part of the lesson, with 87.5% of possible responses for this item scored as not implemented. Item 9 is related to a teacher's encouragement of students making predictions and confirming them during and after reading. In most videos, this item was also not observed (81% scored a 1). The implemented descriptor for Item 8 reads, *The teacher focuses attention on relevant text features and/or structures to organize thinking and support comprehension.* The majority of responses for this item were scored as not implemented (67%), and most occasions when it was observed it was scored as partially implemented (24%), with comments suggesting that teachers pointed out text features, but not in a way that supported comprehension. Only 9% of items were scored as implemented.

In reviewing the less difficult items, Item 15 focuses on a teacher's cueing and correction of decoding errors. 58% of the possible responses were scored as a 3 or

implemented, and for those that were scored as partially implemented, it was generally noted that the teacher did not have the student reread the word, or that they did not encourage the use of strategies to decode unknown words. Item 16 examines the teacher's general questioning practices, and whether they promote understanding of the text. 48% of possible responses were scored as implemented, and items that were scored as partially implemented tended to comment on the pacing or whether the questions were too teacher directed. 76% of the possible responses on item 17 were scored as implemented, and 22% were scored as partially implemented. When the item was scored as partially implemented, the comments included by raters indicated that teachers were inflexible in their ability to reframe questions when students were not able to provide a response.

The third column contains the teacher facet, with more proficient teachers having higher logit values. Teacher 1 is the most proficient teacher (proficiency = .38 logits, *SE* = .11), and teacher 10 is the least proficient (proficiency = -1.08 logits, *SE* = .12). The fourth column contains the lesson facet. In our data collection design the rank ordering of the lesson facet is somewhat difficult to interpret, because we did not specify the content or focus of the lessons but instead had the teachers select which lessons to submit. Consistent with research on teacher observation, our results show that there are differences in teacher performance across lessons, which is why it is important to observe a teacher multiple times throughout the school year (Mantzicopoulos et al., 2018; Patrick & Mantzicopoulous, 2016). The fifth column contains the rater facet, with more severe raters having higher logit values. Rater 2 was our most severe rater (severity = .50 logits, *SE* = .07), with Raters 1, 3 and 4 relatively consistent with one another in severity (severity = -.12, -.16, -.21 respectively logits, *SE* = .07).

Tables 1-4 report the fit statistics and reliability and separation indices for each of the facets. For all facets, all fit statistics fell within .6 to 1.4, which are within acceptable levels (Eckes, 2011). In addition to the fit statistics, reliability and separation information indices are reported. For items, the reliability coefficient was .97, separation = 5.43; for teachers, the reliability coefficient was .91, separation = 3.27. These statistics demonstrate reliable differences in item difficulty and teacher proficiency. For lessons, the reliability coefficient was .88, separation = 2.68, showing a discrimination across lessons. The reliability coefficient for raters was .94, separation = 3.96, suggesting differences in rater severity. The bias analysis indicated that a total of 31.13% of the variance in the observations (n = 1728) was explained by the model. 2.3% was explained by teacher/rater interactions, and 5.7% by item/rater interactions, leaving 60.87% of the variance remaining in residuals.

Table 5 presents only the teacher/rater and item/rater pairs that showed bias and reports observed and expected scores, bias size in logits, standard error, t value and its probability. Of 40 possible teacher/rater interactions, only 3 are biased, and 2 of those interactions involve rater 3. Examining the item/rater interactions, rater 2 is involved in 3 of the 6 significant interactions, scoring item 17 more severely than expected, and items 10 and 11 more leniently than expected. As a whole, the results of this analysis do not appear to exhibit a great deal of bias and the overall MFRM results suggest the facets function effectively. Table 6 includes the rank order of teachers as a measure of their average observed score across all items and lessons, and compares this to the Fair Average score, a score that accounts for rater severity. With the exception of Teachers 7

and 5, the rank order of teacher performance is consistent across observed and fair average scores.

Discussion

The results of the MFRM analyses suggest that we have developed a rubric that will provide reliable evaluations of a teacher's ability to implement reading comprehension instruction consistent with the effective instructional practices described in the research. The high separation and reliability statistics support that the *Reading for Meaning* rubric reliably divided the items and teachers into statistically different strata, indicating the sensitivity of the instrument (Wright & Stone, 1999). The bias analysis indicates limited bias, with 2.3% of the variance accounted for by teacher by rater bias interactions, and 5.7% by item by rater interactions.

The goal of developing the *Reading for Meaning* rubric is to improve teachers' reading comprehension instruction. Whereas observation instruments used in studies of teacher practice have focused on either categorizing elements of instruction (Swanson & Vaughn, 2010), or examining the amount of time spent on various components of instruction (Kent et al., 2012; Vaughn et al., 2002), the RESET *Reading for Meaning* rubric is designed to capture the salient elements of effective comprehension instruction at a grain size that allows for specific, consistent feedback to teachers. The results of this study suggest that this rubric can be used to establish baseline performances of teachers' ability to implement evidence-based comprehension instruction. Next steps in rubric development include examining its impact as a formative assessment used to guide improvements in teacher practice. Following a baseline evaluation, teachers can set goals for improvement, and receive feedback with the rubric throughout the school year.

Although we have not yet tested the *Reading for Meaning* rubric for that purpose, our initial studies with other RESET rubrics suggest that routine observations coupled with feedback can lead to improvements in teacher practice (Authors et al., under review).

A longer-term goal for the development of the RESET observation rubrics is to connect teacher performance to student growth, and to examine the relative contribution of the elements of each instructional practice reflected at the item level. In the case of the *Reading for Meaning* rubric, this would allow teacher preparation and professional development efforts to focus on those elements of comprehension instruction that have the most impact on the reading achievement of SWD, or to create a scope and sequence for teacher training based on those elements of comprehension instruction that are found to have the greatest impact on student performance.

Although the main goal of this study was to investigate the psychometric properties of the observation instrument and not to provide an evaluation of the participating teachers' ability to implement comprehension instruction, the results of the raters' relatively low evaluations of this sample of teachers are consistent with the performance reported in other observation studies. Unfortunately, as evidenced by the distribution of scores across teachers: Implemented, 28%; Partially Implemented, 31%; and Not Implemented, 41%, as well as the distribution of teacher performance depicted on the variable map, our sample of teachers and their recorded lessons did not include examples of high quality comprehension instruction.

When breaking down performance at the item level, the variable map (Figure 1) indicates that the rubric includes a range of items that discriminate across different levels of teacher ability. The 'easier' items, or those on which more teachers were likely to

receive a score of implemented or partially implemented, were focused on decoding and questioning practices (items 15, 16, and 17). This finding is consistent with observation studies of reading instruction that indicate the majority of time is spent on decoding, and that comprehension instruction has historically focused on asking students questions about what they have read (Swanson & Vaughn, 2010). The more difficult items as identified on the variable map included those that focus on strategies such as the use of text preview strategies (item 5), making and confirming predictions (item 9), focusing on relevant text structures (item 8), identifying the main idea and details (item 10), summarizing (item 11) and making inferences (item 12). While effective questioning practices have been shown to be an important strategy for improving comprehension, when questioning routines are not coupled with other strategies the impact on student achievement is likely limited.

An important consideration for the development of observation systems is that the scores provided are a function not only of the teachers' ability but also of the severity of the raters evaluating them. A teacher's performance should not vary considerably when evaluated across raters. Examining the adjustments made using the Fair Average instead of the Observed score show that no changes to a teacher's categorical evaluation or rank ordering occurred. Our analyses indicate that raters differed in their severity, with Rater 2 being the most severe, but the fit statistics were within acceptable levels, with a limited number of bias interactions, suggesting no evidence of halo effects or noisy scoring. Our levels of exact agreement across raters (54.3%) are consistent with those reported across other studies (Cash et al., 2012; Kane & Staiger, 2012).

Although the results are promising, there are limitations in this study that warrant caution. The most significant limitation is that the sample sizes of both special education teachers (n = 10) and raters (n = 4) are small, and somewhat limited in their representativeness of the larger population of special education teachers and potential raters (e.g. all participants were White females). Exploratory work using Rasch analysis can be performed with small samples, though recommendations for stable estimates are typically 30 per parameter (Wright & Stone, 1979). One benefit of using video observations however, is that over time, we can develop a video bank that will include a larger and more diverse pool of teachers. Continued studies with larger samples of teachers and raters can be conducted to verify the results of the studies reported in this manuscript. Additionally, although our larger pool of RESET teacher participants includes teachers across the grade levels, to test the Reading for Meaning rubric, only elementary level teachers could be included, as there were no videos at the secondary level that captured comprehension instruction. Despite these limitations, the results of our analysis are promising. If we can evaluate a teacher's ability to implement evidence-based comprehension instruction, it follows that the rubric can be used to provide feedback and individualized coaching to help improve practice.

For decades, the reading achievement of SWD has remained significantly behind that of their general education peers. Over the same time frame, a significant body of research investigating best practices to improve the comprehension abilities of SWD has been published. One potential explanation for the continued poor achievement of SWD is that research-based practices are either not implemented within the school setting, or they are not implemented with sufficient fidelity to realize the positive effects reported in the

literature. A number of observational studies of instruction support this idea (e.g. Boardman et al., 2005; Klingner et al., 2010; McLeskey & Billingsley, 2008; Vaughn et al., 2002). Klingner et al., (2010) commented in one of their studies that "most special education teachers seemed unsure of how to promote their students' reading comprehension" (p. 59). This is consistent with what we have observed while developing the RESET observation system. Although most teachers are doing their best to serve SWD well, there is a significant disconnect between the practices in the classroom with what is described in the research-base. If we are to improve reading outcomes for SWD, we must create observation systems that align targets for high quality comprehension instruction with observations of teachers who deliver these practices.

References

Adnot, M., Dee, T., Katz, V., & Wyckoff, J. (2017). Teacher turnover, teacher quality, and student achievement in DCPS. Educational Evaluation and Policy Analysis, 39(1), 54-76.

Alves, K. D., Kennedy, M. J., Brown, T. S., & Solis, M. (2015). Story grammar instruction with third and fifth grade students with learning disabilities and other struggling readers. *Learning Disabilities: A Contemporary Journal*, *13*(1), 73-93.

Authors, (under review). Improving special education teacher's implementation of evidence-based practices through feedback on an observation instrument. *Exceptional Children*

Barth, A. E., Barnes, M., Francis, D., Vaughn, S., & York, M. (2015). Inferential processing among adequate and struggling adolescent comprehenders and relations to reading comprehension. *Reading and writing*, *28*(5), 587-609.

Bell, C A., Yi Q, Croft, A J., Leusner, D., McCaffrey, D. F., Gitomer, D. H. & Pianta, R.C. (2015). Improving Observational Score Quality. Designing Teacher Evaluation Systems: New Guidance from the Measures of Effective Teaching Project, 50- 97. San Francisco: Jossey-Bass.

Berkeley, S., Scruggs, T. E., & Mastropieri, M. A. (2010). Reading comprehension instruction for students with learning disabilities, 1995—2006: A Meta-analysis. *Remedial and Special Education*, *31*(6), 423-436.

Biancarosa, G., Bryk, A. S., & Dexter, E. R. (2010). Assessing the value-added effects of literacy collaborative professional development on student learning. *The elementary school journal*, *111*(1), 7-34.

Bill and Melinda Gates Foundation. (2011). Learning about teaching: Initial findings from the Measures of Effective Teaching project. Bellevue, WA: Author. Retrieved from www.gatesfoundation.org/collegeready-education/Documents/preliminaryfindings-research-paper.pdf

Blazar, D., Braslow, D., Charalambous, C. Y., & Hill, H. C. (2017). Attending to general and mathematics-specific dimensions of teaching: Exploring factors across two observation instruments. *Educational Assessment*, *22*(2), 71-94.

Boardman, A. G., Vaughn, S., Buckley, P., Reutebuch, C., Roberts, G., & Klingner, J. (2016). Collaborative Strategic Reading for students with learning disabilities in upper elementary classrooms. *Exceptional Children*, *82*(4), 409-427.

Bryant, D. P., Goodwin, M., Bryant, B. R., & Higgins, K. (2003). Vocabulary instruction for students with learning disabilities: A review of the research. *Learning Disability Quarterly*, *26*(2), 117-128.

Cain, K. (2010). *Reading development and difficulties* (Vol. 8). John Wiley & Sons.

Cain, K., Oakhill, J., & Bryant, P. (2004). Children's reading comprehension ability: Concurrent prediction by working memory, verbal ability, and component skills. *Journal of educational psychology*, *96*(1), 31.

Cain, K., Oakhill, J. V., Barnes, M. A., & Bryant, P. E. (2001). Comprehension skill, inference-making ability, and their relation to knowledge. *Memory & cognition*, *29*(6), 850-859.

Casabianca, J. M., Lockwood, J. R., & McCaffrey, D. F. (2015). Trends in classroom observation scores. *Educational and Psychological Measurement, 75*(2), 311-337.

Cash, A. H., Hamre, B. K., Pianta, R. C., & Myers, S. S. (2012). Rater calibration when observational assessment occurs at large scale: Degree of calibration and characteristics of raters associated with calibration. *Early Childhood Research Quarterly, 27*, 529-542.

Catts, H. W., & Kamhi, A. G. (2017). Prologue: Reading comprehension is not a single ability. *Language, Speech, and Hearing Services in Schools*, *48*(2), 73-76.

Ciullo, S., Lo, Y. L. S., Wanzek, J., & Reed, D. K. (2016). A synthesis of research on informational text reading interventions for elementary students with learning disabilities. *Journal of learning disabilities*, *49*(3), 257-271.

Compton, D. L., Miller, A. C., Elleman, A. M., & Steacy, L. M. (2014). Have we forsaken reading theory in the name of "quick fix" interventions for children with reading disability?. *Scientific Studies of Reading*, *18*(1), 55-73.

Eckes, T. (2011). *Introduction to many-facet rasch measurement*. Frankfurt: Peter Lang.

Elleman, A. M. (2017). Examining the impact of inference instruction on the literal and inferential comprehension of skilled and less skilled readers: A meta-analytic review. *Journal of Educational Psychology*, *109*(6), 761.

Elleman, A. M., & Compton, D. L. (2017). Beyond comprehension strategy instruction: What's next?. *Language, Speech, and Hearing Services in Schools*, *48*(2), 84-91.

Elleman, A. M., Lindo, E. J., Morphy, P., & Compton, D. L. (2009). The impact of vocabulary instruction on passage-level comprehension of school-age children: A meta-analysis. *Journal of Research on Educational Effectiveness*, *2*(1), 1-44.

Elleman, A. M., Steacy, L. M., Olinghouse, N. G., & Compton, D. L. (2017). Examining Child and Word Characteristics in Vocabulary Learning of Struggling Readers. *Scientific Studies of Reading*, *21*(2), 133-145.

El Zein, F., Solis, M., Vaughn, S., & McCulley, L. (2014). Reading comprehension interventions for students with autism spectrum disorders: A synthesis of research. *Journal of Autism and Developmental Disorders*, *44*(6), 1303-1322.

Engelhard Jr, G. (1992). The measurement of writing ability with a many-faceted Rasch model. *Applied Measurement in Education*, *5*(3), 171-191.

Gajria, M., Jitendra, A. K., Sood, S., & Sacks, G. (2007). Improving comprehension of expository text in students with LD: A research synthesis. *Journal of learning disabilities*, *40*(3), 210-225.

Gersten, R., Fuchs, L. S., Williams, J. P., & Baker, S. (2001). Teaching reading comprehension strategies to students with learning disabilities: A review of research. *Review of educational research*, *71*(2), 279-320.

Grossman, P., Compton, C., Igra, D., Ronfeldt, M., Shahan, E., & Williamson, P. (2009). Teaching practice: A cross-professional perspective. *Teachers College Record*, *111*(9), 2055-2100.

Hall, C. S. (2016). Inference instruction for struggling readers: A synthesis of intervention research. *Educational Psychology Review*, *28*(1), 1-22.

Hill, H., & Grossman, P. (2013). Learning from teacher observations: Challenges posed by new teacher evaluation systems. *Harvard educational review*, *83*(2), 371-384.

Honig, B., Diamond, L., & Gutlohn, L. (2000). *Teaching Reading: Sourcebook for Kindergarten through Eighth Grade*. Arena Press, 20 Commercial Boulevard, Novato, CA 94949-6191

Jitendra, A. K., Edwards, L. L., Sacks, G., & Jacobson, L. A. (2004). What research says about vocabulary instruction for students with disabilities. *Exceptional Children*, *70*(3), 299-322.

Jitendra, A. K., Kay Hoppes, M., & Xin, Y. P. (2000). Enhancing main idea comprehension for students with learning problems: The role of a summarization strategy and self-monitoring instruction. *The Journal of Special Education*, *34*(3), 127-139.

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

Johnson, E. S., Crawford, A. R., Moylan, L. A., & Zheng, Y. (in press), Using evidence-centered design to create a special educator observation system, *Educational Measurement: Issues and Practice*

Joseph, L. M., Alber-Morgan, S., Cullen, J., & Rouse, C. (2016). The effects of self-questioning on comprehension: A literature review. *Reading & Writing Quarterly*, *32*(2), 152-173.

Judge, S., & Bell, S. M. (2010). Reading achievement trajectories for students with learning disabilities during the elementary school years. Reading & Writing Quarterly: Overcoming Learning Difficulties, 27, 153–178.

Kaldenberg, E. R., Watt, S. J., & Therrien, W. J. (2015). Reading instruction in science for students with disabilities: A meta-analysis. *Learning Disability Quarterly*, *38*(3), 160-173.

Kane, T. J., & Staiger, D. O. (2012). Gathering feedback for teaching: Combining high-quality observations with student surveys and achievement gains. Research Paper. MET Project, *Bill & Melinda Gates Foundation*.

Kendeou, P., & van den Broek, P. (2007). The effects of prior knowledge and text structure on comprehension during reading of scientific texts. *Memory & cognition*, *35*(7), 1567-1577.

Kent, S. C., Wanzek, J., & Al Otaiba, S. (2012). Print reading in general education kindergarten classrooms: What does it look like for students at-risk for reading difficulties?. *Learning Disabilities Research & Practice*, *27*(2), 56-65.

Kim, W., Linan-Thompson, S., & Misquitta, R. (2012). Critical factors in reading comprehension instruction for students with learning disabilities: A research synthesis. *Learning Disabilities Research & Practice*, *27*(2), 66-78.

Kintsch, W. (2004). The construction-integration model of text comprehension and its implications for instruction. *Theoretical models and processes of reading*, *5*, 1270-1328.

Kintsch, W. (2005). An overview of top-down and bottom-up effects in comprehension: The CI perspective. *Discourse processes*, *39*(2-3), 125-128.

Klingner, J. K., Urbach, J., Golos, D., Brownell, M., & Menon, S. (2010). Teaching reading in the 21st century: A glimpse at how special education teachers promote reading comprehension. *Learning Disability Quarterly*, *33*(2), 59-74.

Linacre, J. M. (2014). *Facets 3.71. 4* [Computer software].

Linacre, J. M. (1994). *Many-facet Rasch measurement.* University of Chicago Press: Chicago, IL

McNamara, D. S., O'Reilly, T. P., Best, R. M., & Ozuru, Y. (2006). Improving adolescent students' reading comprehension with iSTART. *Journal of Educational Computing Research*, *34*(2), 147-171.

Mason, L. H. (2013). Teaching students who struggle with learning to think before, while, and after reading: Effects of self-regulated strategy development instruction. *Reading & Writing Quarterly*, *29*(2), 124-144.

Mason, L. H., & Hedin, L. R. (2011). Reading science text: Challenges for students with learning disabilities and considerations for teachers. *Learning Disabilities Research & Practice*, *26*(4), 214-222.

McKeown, M. G., Beck, I. L., & Blake, R. G. (2009). Rethinking reading comprehension instruction: A comparison of instruction for strategies and content approaches. *Reading Research Quarterly*, *44*(3), 218-253.

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

Nagy, W. E., Anderson, R. C., & Herman, P. A. (1987). Learning word meanings from context during normal reading. *American educational research journal*, *24*(2), 237-270.

National Reading Panel (US), National Institute of Child Health, & Human Development (US). (2000). *Report of the national reading panel: Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction: Reports of the subgroups*. National Institute of Child Health and Human Development, National Institutes of Health.

Oakhill, J., & Cain, K. (2007). Issues of causality in children's reading comprehension. *Reading comprehension strategies: Theories, interventions, and technologies*, 47-72.

O'Connor, R. E., Sanchez, V., Beach, K. D., & Bocian, K. M. (2017). Special Education Teachers Integrating Reading with Eighth Grade US History Content. *Learning Disabilities Research & Practice*, *32*(2), 99-111.

Perfetti, C. (2007). Reading ability: Lexical quality to comprehension. *Scientific studies of reading*, *11*(4), 357-383.

Perfetti, C., & Stafura, J. (2014). Word knowledge in a theory of reading comprehension. *Scientific Studies of Reading*, *18*(1), 22-37.

Scammacca, N. K., Roberts, G. J., Cho, E., Williams, K. J., Roberts, G., Vaughn, S. R., & Carroll, M. (2016). A century of progress: Reading interventions for students in grades 4–12, 1914–2014. *Review of educational research*, *86*(3), 756-800.

Schulte, A. C., Stevens, J. J., Elliott, S. N., Tindal, G., & Nese, J. F. (2016). Achievement gaps for students with disabilities: Stable, widening, or narrowing on a state-wide reading comprehension test?. *Journal of Educational Psychology*, *108*(7), 925.

Smith Jr, E. V., & Kulikowich, J. M. (2004). An application of generalizability theory and many-facet Rasch measurement using a complex problem-solving skills assessment. *Educational and Psychological Measurement*, *64*(4), 617-639.

Snow, C. (2002). *Reading for understanding: Toward an R&D program in reading comprehension*. Rand Corporation.

Solis, M., Ciullo, S., Vaughn, S., Pyle, N., Hassaram, B., & Leroux, A. (2012). Reading comprehension interventions for middle school students with learning disabilities: A synthesis of 30 years of research. *Journal of learning disabilities*, *45*(4), 327-340.

Stetter, M. E., & Hughes, M. T. (2010). Using story grammar to assist students with learning disabilities and reading difficulties improve their comprehension. *Education and Treatment of Children*, *33*(1), 115-151.

Swanson, E. A. (2008). Observing reading instruction for students with learning disabilities: A synthesis. *Learning Disability Quarterly*, *31*(3), 115-133.

Swanson, E. A., & Vaughn, S. (2010). An observation study of reading instruction provided to elementary students with learning disabilities in the resource room. *Psychology in the Schools*, *47*(5), 481-492.

Taylor, E. S., & Tyler, J. H. (2012). Can teacher evaluation improve teaching. *Education Next*, *12*(4), 78-84.

Vaughn, S., Klingner, J. K., & Bryant, D. P. (2001). Collaborative strategic reading as a means to enhance peer-mediated instruction for reading comprehension and content-area learning. *Remedial and Special Education*, *22*(2), 66-74.

Vaughn, S., Levy, S., Coleman, M., & Bos, C. S. (2002). Reading instruction for students with LD and EBD: A synthesis of observation studies. *The Journal of Special Education*, *36*(1), 2-13.

Vaughn, S., & Wanzek, J. (2014). Intensive interventions in reading for students with reading disabilities: Meaningful impacts. *Learning Disabilities Research & Practice*, *29*(2), 46-53.

Wanzek, J., Swanson, E., Vaughn, S., Roberts, G., & Fall, A. M. (2016). English learner and non-English learner students with disabilities: Content acquisition and comprehension. *Exceptional Children*, *82*(4), 428-442.

Wei, X., Blackorby, J., & Schiller, E. (2011). Growth in reading achievement of students with disabilities, ages 7 to 17. *Exceptional Children*, *78*(1), 89-106.

Wigglesworth, G. (1993). Exploring bias analysis as a tool for improving rater consistency in assessing oral interaction. *Language Testing, 10,* 305-335.

Williams, J. P. (2005). Instruction in reading comprehension for primary-grade students: A focus on text structure. *The Journal of Special Education*, *39*(1), 6-18.

Williams, J. P., Pollini, S., Nubla-Kung, A. M., Snyder, A. E., Garcia, A., Ordynans, J. G., & Atkins, J. G. (2014). An intervention to improve comprehension of cause/effect through expository text structure instruction. *Journal of Educational Psychology*, *106*(1), 1.

Willingham, D. T. (2007). Critical thinking. *American Educator, 31*(3), 8-19.

Wright, B. D., & Stone, M. H. (1979). *Best test design*. University of Chicago Press: Chicago, IL

Yuill, N., & Oakhill, J. (1988). Effects of inference awareness training on poor reading comprehension. *Applied Cognitive Psychology*, *2*(1), 33-45.

Measr	-Items	+Teacher	-Lesson	-Raters	Scale
2					(3)
	5				
	9				
1					-------
	8				
	10			2	
		Teacher 1			
	11,12,14,4				
		Teacher 2	1		
0	13, 6	Teacher 8			2
	7	Teacher 4	3	1	
	3		2	3, 4	
	1	Teachers 3, 5, 7, 9			
		Teacher 6			
	18, 2				
-1	15, 16				-------
		Teacher 10			
-2	17				1

Figure 3.1. Variable map of the Reading for Meaning rubric facets items, teachers, lessons and raters

Table 3.1 Item Measure Report from Many-Facet Rasch Measurement Analysis

Item Number	Difficulty (Logits)	Model SE	Infit MNSQ	Outfit MNSQ
17	-1.98	.21	.86	.86
16	-1.04	.15	.61	.61
15	-1.00	.15	1.20	1.18
18	-.63	.15	1.05	1.04
2	-.61	.14	.99	.98
1	-.36	.14	.75	.77
3	-.18	.14	.90	.89
7	-.12	.14	1.32	1.29
6	.01	.14	.77	.75
13	.01	.14	1.15	1.17
14	.29	.14	.88	.86
11	.31	.14	1.11	1.10
12	.31	.14	.81	.80
4	.35	.14	1.11	1.04
10	.46	.15	1.13	1.11
8	.89	.17	1.08	1.03
9	1.39	.20	1.32	1.22
5	1.89	.25	1.33	1.24
Mean (count = 18)	.00	.16	1.02	1.00
SD	.88	.03	.21	.19

Note. Root mean square error (model) = .16; adjusted *SD* = .86; separation = 5.43; reliability = .97; fixed chi-square = 393.2; df = 17; significance = .00.

Table 3.2 Teacher Measure Report from Many-Facet Rasch Measurement Analysis

Teacher Number	Ability (Logits)	Model SE	Infit MNSQ	Outfit MNSQ
10	-1.08	.12	.94	.83
6	-.50	.11	1.00	1.10
7	-.40	.12	1.08	1.02
5	-.37	.11	.81	.92
3	-.35	.11	1.17	1.13
9	-.35	.11	1.11	1.04
4	-.11	.11	.86	.80
8	-.03	.11	1.03	.99
2	.20	.11	1.11	1.08
1	.38	.11	.95	1.09
Mean (count = 10)	-.26	.11	1.01	1.00
SD	.38	.00	.11	.11

Note. Root mean square error (model) = .11; adjusted *SD* = .37; separation = 3.27; reliability = .91; fixed chi-square = 111.9; df = 9; significance = .00.

Table 3.3 Lesson Measure Report from Many-Facet Rasch Measurement Analysis

Lesson Number	Difficulty (Logits)	Model SE	Infit MNSQ	Outfit MNSQ
2	-.17	.06	.98	.99
3	-.07	.06	1.04	1.03
1	.24	.06	.98	.98
Mean (count = 3)	.00	.06	1.00	1.00
SD	.18	.00	.03	.02

Note. Root mean square error (model) = .06; adjusted *SD* = .16; separation = 2.68; reliability = .88; fixed chi-square = 23.9; df = 2; significance = .00.

Table 3.4 Rater Measure Report from Many-Facet Rasch Measurement Analysis

Rater Number	Severity (Logits)	Model SE	Infit MNSQ	Outfit MNSQ
4	-.21	.07	1.10	1.04
3	-.16	.07	1.00	1.06
1	-.12	.07	1.08	1.02
2	.50	.07	.82	.87
Mean (count = 4)	.00	.07	1.00	1.00
SD	.29	.00	.11	.07

Note. Root mean square error (model) = .07; adjusted *SD* = .28; separation = 3.96; reliability = .94; fixed chi-square = 65; df = 3; significance = .00.

Table 3.5 Bias Analysis Results

Teacher - Rater	Observed Score	Expected Score	Bias Size	Model S.E.	*t*	*p*
1 – 3	72	80.66	-.50	.24	-2.10	.043
6 – 3	111	99.21	.45	.19	2.30	.025
2 – 2	81	68.84	.69	.24	2.83	.007
Item - Rater	**Observed Score**	**Expected Score**	**Bias Size**	**Model S.E.**	***t***	***p***
11-1	30	42.53	-1.31	.42	-3.15	.004
18-4	40	52.64	-.99	.29	-3.46	.002
17-2	58	65.57	-.70	.28	-2.51	.019
10-2	43	35.48	.66	.28	2.40	.024
11-2	45	36.93	.66	.27	2.45	.022
16-3	67	59.25	1.02	.45	2.26	.033
15-1	68	59.32	1.25	.50	2.48	.021

Note. Observed and expected scores are based on the total possible number of points across the observed count of items.

Table 3.6 Teacher Measurement Report

Teacher	Observed Score	Fair Average Score	Measure	S.E.
10	1.53	1.43	-1.08	.12
6	1.79	1.71	-.50	.11
7	1.82	1.76	-.40	.12
5	1.81	1.78	-.37	.11
3	1.84	1.79	-.35	.11
9	1.83	1.79	-.35	.11
4	1.93	1.93	-.11	.11
8	1.99	1.98	-.03	.11
2	2.10	2.12	.20	.11
1	2.12	2.21	.38	.11

CHAPTER FOUR: DEVELOPING A COMPREHENSIVE DECODING INSTRUCTION OBSERVATION PROTOCOL FOR SPECIAL EDUCATION TEACHERS

This chapter is an unpublished manuscript prepared for submission to a peer-reviewed journal and should be referenced appropriately.

Reference:

Moylan, L., Johnson, E.S., & Zheng, Y. (unpublished manuscript). Developing a comprehensive decoding instruction observation protocol for special education teachers.

Developing a Comprehensive Decoding Instruction Observation Protocol for Special Education Teachers

Laura A. Moylan, Evelyn S. Johnson, and Yuzhu Zheng

Boise State University

Author Note

Laura A. Moylan, MEd., Project RESET, Boise State University; Evelyn S. Johnson, Ed.D., Department of Early and Special Education, Boise State University; Yuzhu Zheng, Ph.D., Project RESET; Boise State University.

This research was supported the Institute of Education Sciences, award number R324A150152 to Boise State University. The opinions expressed are solely those of the authors. Correspondence regarding this manuscript should be addressed to: Dr. Evelyn S. Johnson, Boise State University, 1910 University Dr., MS 1725, Boise, Idaho 83725-1725. Email: evelynjohnson@boisestate.edu

Abstract

This study describes the development of a special education teacher observation protocol detailing the elements of effective decoding instruction. The psychometric properties of the protocol were investigated through many-facet Rasch measurement (MFRM). Video observations of classroom decoding instruction from 20 special education teachers across three states were collected. Twelve external raters were trained to observe and evaluate instruction using the protocol and assigned scores of "implemented", "partially implemented", or "not implemented" for each of the items. Analyses showed that the item, teacher, lesson, and rater facets achieved high levels of reliability. Teacher performance was consistent with what is reported in the literature. Implications for practice are discussed.

Keywords: evidence-based decoding instruction, observation systems, many-facet Rasch measurement

Introduction

Reading is a complex process requiring the reader to integrate, coordinate, and execute multiple skills and processes in order to extract meaning from text (Cain et al., 2004; Cain, 2009; Perfetti, 2007). While the ability to accurately and efficiently read words does not ensure comprehension will occur, word reading proficiency is a necessary component of this complex process, as evidenced by its role as a key predictor of reading comprehension ability (Cain et al., 2004; Castles et al., 2018; Denton & Al Otaiba, 2011; Ehri et al., 2001; Kang & Shin, 2019). In a comprehensive synthesis of the science of reading, phonics instruction was emphasized as the foundation by which students acquire mastery of the alphabetic code, fluent word recognition, and skilled comprehension (Castles et al., 2018). Empirical evidence consistently supports the need for students with or at risk for reading disabilities (hereafter abbreviated as SWD) to receive intensive, explicit, and systematic instruction in word reading skills and strategies emphasizing phonological (sound) and orthographic (written) connections (Blachman et al., 2004; Denton et al., 2013; Ehri et al., 2001; Lovett et al., 2000; Torgesen et al., 2001).

Despite the depth of the literature base describing the evidence-based practices (EBP) that promote strong word reading skills, observation studies of teachers' practice routinely indicate a lack of consistent and effective implementation of these EBPs, particularly in classrooms focused on providing instruction to SWD (Moody et al., 2000; Swanson, 2008; Vaughn & Wanzek, 2014). Observational studies of classroom practices have consistently concluded the quality, intensity and content of reading instruction is inadequate to meet the intensive instructional needs for students SWD (Kent et al., 2017; Klingner et al., 2010; Swanson, 2008; Vaughn & Wanzek, 2014). Additionally, concerns

have been raised about the lack and depth of content knowledge among teachers providing reading instruction, inhibiting their ability to explain concepts effectively, select appropriate examples, be diagnostic, and provide targeted feedback to students (Moats & Foorman, 2003; Moats, 2009; Washburn et al., 2011). This gap between research and practice may provide some explanation for on-going reading achievement concerns, with significant numbers of SWD performing below proficiency on state and national measures of reading, and persistent gaps in performance between SWD and their general-education peers (Judge & Bell, 2010; NCES, 2019; Schulte et al., 2016).

If we are to improve reading outcomes for SWD, it is essential to ensure teachers have the knowledge of EBPs, the skills to sustain fidelity to implementation, and the ongoing support to consistently provide high quality instruction to SWD (McLeskey & Billingsley, 2008). To inform these efforts, it is also critical to establish baseline levels of teacher instructional performance and to define teacher development as observable, measurable progress toward an ambitious, well-articulated standard for practice. Providing both in-service and preservice teachers with feedback on their implementation of instructional practices has been shown to have positive effects on teacher performance and the effective implementation of EBPs (Fallon et al., 2015; Schles & Robertson, 2019; Solomon et al., 2012). Coaching models that include observations paired with specific performance-based feedback have the potential to produce observable and measurable changes in the accuracy of EBP implementation (Fallon et al., 2015; Kretlow & Bartholomew, 2010). Teachers significantly improved their knowledge and implementation of systematic phonics instruction and student outcomes following a year-long mentoring program including on-going and consistent modeling and feedback

aligned to specific practices (Ehri & Flugman, 2018). Observation protocols that identify and define the components of a specific practice provide opportunities for focused feedback on both content and delivery and have the potential to promote and incentivize effective implementation (Hill & Grossman, 2013), equipping educators with common language and a framework to guide the systematic and continuous implementation of EBP.

Comprehensive Decoding Lesson Observation Protocol

The Recognizing Effective Special Education Teachers (RESET) observation system is a federally funded project to create teacher observation protocols aligned to EBP for SWD (Johnson et al., 2018). The goal of RESET is to create and validate an observation system comprised of observation protocols that leverage the extensive research on instructional EBPs for SWD (Johnson et al., 2020). The RESET system was developed using the principles of evidence-centered design to create observation protocols that effectively capture the complexities of EBPs (Johnson et al., 2018; Mislevy et al., 2003). The observation protocol of interest for the current study is the Comprehensive Decoding Lesson Protocol (hereafter abbreviated as CDLP). The CDLP was designed to evaluate and support the implementation of systematic and explicit phonics instruction and practice, providing teachers with content specific targets aligned with the essential features of a comprehensive decoding intervention for SWD.

The first step in developing the CDLP was to identify the components of decoding instruction as described in the research. The RESET research team conducted a systematic review of the literature and identified the critical components of a comprehensive decoding lesson as: (a) systematic instruction, (b) explicit instruction in

phoneme-grapheme correspondence and word reading skills and strategies, (c) encoding, (d) the integration of word meaning, reading and processing decodable text, and (e) consistent monitoring and feedback throughout the lesson. Each of these components is briefly described.

Systematic Instruction

In order for SWD to make significant progress toward word reading proficiency, instruction must be highly explicit, efficient and intensive, providing students with extended opportunities to practice (Blachman et al., 2004; Denton & Al Otaiba, 2011). Systematic phonics instruction is characterized by a planned set of elements or concepts that are taught and practiced sequentially, then build logically upon one another providing students with the prerequisite skills needed to learn new concepts and advance their ability to decode and read words in isolation and in context (Brady, 2011; Ehri et al., 2001). Concepts are presented as part of a coherent system, and instruction includes regular step-by-step procedures or routines such as the "I do", "We do", "You do" procedures found in explicit instruction or systematic cues for routines such as "Blend it" or "What's the word?" (Archer & Hughes, 2010). Well established and implemented routines and procedures lead to a more fluid, efficient, and focused lesson where students know what is expected and have clear opportunities to respond (Archer & Hughes, 2010).

Phoneme-Grapheme Correspondence

Understanding of the alphabetic principle, the awareness that letters and letter patterns represent the sounds in language, and the understanding that these relationships are systematic and predictable is central to learning to read, and the foundation of effective reading instruction and intervention for SWD (Blachman et al., 2004; Ehri et al.,

2001; Foorman et al., 2003; Steacy et al., 2016; Torgesen et al., 2001). To become fluent word readers, students must be able to distinguish the distinct phonemes in words, such as *bat*, /b/ /a/ /t/, understand that the /b/ in bat is the same as the /b/ in b-a-g, and connect those phonemes to the corresponding graphemes by linking the sound /b/ with the grapheme *b*. The acquisition of these fundamental skills requires systematically designed explicit instruction and practice with increased levels of intensity for SWD (Blachman et al., 2004; Denton et al., 2013; Torgesen et al., 2001).

Word Reading

Accurate and fluent word reading skills are integral to the process of comprehending text (Castles et al., 2018; Gough & Tunmer, 1986; Perfetti & Stafura, 2014). Two approaches to support accurate word reading for SWD are synthetic phonics instruction (mapping phonemes to graphemes and blending to decode words) and analytic phonics instruction (recognizing larger word parts and patterns such as onset, rimes, syllables) or a combination of both methods (Denton & Al Otaiba, 2011; Ehri et al., 2001; Lovett et al., 2000; Torgesen et al., 2001; Wanzek & Vaughn, 2007). Multiple exposures and frequent opportunities to read words enhance students' ability to retain them in memory, and to learn orthographic patterns that facilitate orthographic mapping (Ehri, 2014).

Reading Decodable Text and Developing Word Knowledge

The ultimate goal of reading is comprehension, and instruction must be designed to not only facilitate word level reading skills, but to engage with texts, to develop background knowledge and to increase vocabulary. Reading performance improves when students are provided with explicit and systematic instruction in decoding paired with the

opportunity to successfully apply skills in text reading (Blachman et al., 2004; Denton et al., 2010; Jenkins et al., 2004; Mathes et al., 2005). Despite the documented effect sizes when interventions include both phonics instruction and daily opportunities to read and respond to text at the appropriate level of difficulty, observation studies indicate that SWD spend limited amounts of time engaged in reading, as low as 1-4% of classroom instructional time (Vaughn & Wanzek, 2014).

Forming a coherent mental representation of a text requires automaticity with word level reading skills, but word reading skills must also be situated within the larger framework of the reading process. While the primary focus of the CDLP is on instructional practices that target a student's word level reading abilities, a comprehensive approach to reading instruction includes a focus on word meaning and comprehending what is read. Comprehension is scaffolded by engaging background knowledge prior to reading and providing students the opportunity for discussion appropriate to the text. Vocabulary knowledge has been identified as impacting both word identification and comprehension (Perfetti & Stafura, 2014; Tunmer & Chapman, 2012), suggesting that instruction on word meaning should be included as part of decoding instruction.

Encoding

Students demonstrate greater levels of improvement in reading and spelling when they engage in explicit decoding instruction paired with encoding instruction focused on phoneme-grapheme mapping (Denton et al., 2013; Weiser, 2013). To be most effective in reinforcing phoneme-grapheme relationships, the encoding portion of a decoding lesson must make explicit connections between phonemes and graphemes and may include

exercises such as writing dictated words and manipulating tools such as letter tiles to form words paired with immediate corrective or reinforcing feedback (Weiser & Mathes, 2011).

Monitoring and Feedback

Feedback as a general construct has the potential to powerfully influence learning; the impact is dependent upon the type of feedback provided and how it is given (Hattie & Timperley, 2007). Providing students with timely, corrective and/or affirmative feedback is a critical component of effective instruction (Denton & Al Otaiba, 2011) and a key component of explicit, systematic instruction (Archer & Hughes, 2010; Hughes et al., 2017).

Comprehensive Decoding Lesson Protocol Structure and Scoring

Once the review of literature was complete, the CDLP was drafted to capture the critical components of a comprehensive decoding lesson as outlined above (see Appendix A for a copy of the CDLP). A total of 18 items on the CDLP are organized by the seven components: 1) Systematic Instruction, 2) Phoneme-Grapheme Correspondence, 3) Word Reading, 4) Encoding, 5) Word Meaning, 6) Reading Decodable Text, and 7) Monitoring and Feedback Throughout the Lesson. Items aligned to these components were developed through an iterative process involving the translation of practices from the literature, drafting an item, testing items with video, eliciting subject matter expert input, and revision. Once we developed the set of items that described proficient implementation, studies were conducted to inform the performance level descriptors for each item across a three-point scale, where a 3 is “proficient implementation”, a 2 is “partial implementation”, and a 1 is “not implemented” (see Johnson et al., 2018 for a full

description of this process). The RESET CDLP is designed for use with video recorded lessons that are observed and evaluated by raters with expertise in reading instruction for SWD and who receive training to accurately and consistently apply the scoring criteria.

Purpose of the Current Study

Observation protocols require a deliberate approach to development and a rigorous evaluation of the various facets that can impact a teacher's observed scores. If the RESET CDLP is to serve the purpose of improving teachers' ability to effectively implement EBPs, it must result in reliable evaluations of teachers' ability to effectively implement decoding instruction, and must provide teachers with specific, and actionable feedback on how to improve. Reliable observations of teacher practice can serve as a baseline performance that can inform professional development efforts. Therefore, the purpose of the current study was: 1) to examine the psychometric quality of the CDLP through MFRM analysis and 2) to analyze teachers' performance on the implementation of effective decoding instruction.

Methods

Participants

Teachers

Twenty special education teachers from three states (Idaho, Florida, Wisconsin) each provided three video recorded lessons for a total of 60 videos. Teachers were recruited by contacting district special education directors, who then distributed consent forms to eligible candidates. All participating teachers identified as white females teaching SWD at the elementary school level, and had an average experience level of 13.5 years. Seven teachers held a master's degree in special education or literacy, eleven

teachers held a bachelor's degree in special education and two teachers held a bachelor's degree in elementary education.

Raters

Twelve raters, one male and eleven females, from six states (Idaho, Maryland, Illinois, Oregon, Pennsylvania, Utah) participated in this study. Raters were recruited through a purposive sampling technique, focused on selecting individuals with strong knowledge of special education instruction and reading intervention. Three raters held a doctoral degree and worked as a special education teacher, a school administrator and a special education faculty member at a university in the eastern part of the U.S. Three raters held master's degrees and were doctoral students. Five raters held master's degrees and were working as special education teachers or reading specialists. One rater was a retired teacher who held a master's degree in education.

Procedures

Video Collection

Video observations of classroom instruction were collected over a three-year period (2015-2018). Participating teachers provided video recorded lessons from a consistent instructional period each week. Videos were recorded and uploaded using the Swivl ® capture system and ranged in length from 30 – 45 minutes of instructional time. Videos were organized into three time periods from across the school year (e.g. September – December, January – March, April – June). From this video bank, one video from each of these three time periods from each teacher was selected for inclusion in this study, for a total of 3 videos per teacher, 60 videos for the study. Videos were assigned an

ID number and listed in random order for each rater according to the rating scheme described below, to control for order effects.

Rater Training

Rater training took place over four days. Each day consisted of a four-hour training session conducted by the RESET project staff, followed by additional video viewing and scoring assignments to be completed independently prior to the following day's training. Raters were provided with an overview of the RESET project goals, and a description of how the CDLP was developed. Project staff then explained each item in the CDLP using a video model to demonstrate 'implemented' and clarified any questions raters had about the items. Raters were also provided with a detailed training manual that included in-depth explanations of each item, definitions of reading terms and exemplars of performance levels. Over the course of training, raters independently watched and scored three videos and documented evidence observed in the video aligned with their scoring decisions. Scores and evidence were reviewed and discussed the following day during training sessions with project staff and reconciled with a master coded rubric for each video. Any disagreements were reviewed and discussed.

Rater Scoring Design

Raters were assigned a randomly ordered list of videos to control for order effects. Instead of having each rater observe every video, we created a rating scheme that allowed for the connection of ratings across all rater pairs and across teachers (Eckes, 2011). Each rater scored 22 of the 60 videos. Two videos were scored by all twelve raters, 19 videos were scored by five raters, 28 videos were scored by four raters and eleven videos were scored by three raters. Three raters scored each teacher at least one time. Raters were

asked to score each item on the protocol for each video, to provide time-stamped evidence of what they observed and used as a basis for the score, and to provide a brief explanation of the rationale for their score. Raters were given six weeks to complete their ratings and enter the data into an electronic version of the CDLP.

Data Analysis

The instructional dimensions of observation protocols have been reported to be the most challenging for raters to score reliably (Gitomer et al., 2014; Ho & Kane, 2013). Rater behavior research suggests achieving perfect agreement across raters who judge complex performances is an elusive goal, and that acknowledging raters will differ in severity but can be trained to be consistent in their own scoring may be a more attainable reality (Eckes, 2011). Therefore, data were analyzed using MFRM analyses. MFRM analyses allow for the investigation of teacher, lesson, rater, and item facets (Eckes, 2011). Using the Rasch model, the ability estimates of teachers are freed from the distributional properties of the items, lessons, and the particular raters used to rate the performance (Eckes, 2011). Additionally, the estimated difficulty of items and severity of raters are freed from the distributional properties of the other facets of the assessment (Smith & Kulikowich, 2004).

The model used for the MFRM analyses in this study is given by:

$$ln\left(\frac{P_{nijok}}{P_{nijo(k-1)}}\right) = B_n - D_i - C_j - T_o - F_k$$

where P_{nijok} is the probability of teacher *n*, when rated on item *i* by judge (rater) *j* on occasion (lesson) *o*, being awarded a rating of *k*. $P_{nijo(k-1)}$is the probability of teacher *n*, when rated on item *i* by judge (rater) *j* on occasion (lesson) *o*, being awarded a rating of *k*-1, B_n is the ability of teacher *n*, D_i is the difficulty of item *i*, C_j is the severity of

judge *j*, T_o is the stringency of the occasion *o*, and F_k is the difficulty overcome in being observed at the rating *k* relative to the rating *k*-1 (Eckes, 2011).

The MFRM analysis was conducted using the computer program FACETS version 3.71 (Linacre, 2014a). MFRM analysis produces infit and outfit statistics for each facet, two quality control statistics that indicate whether the measures have been confounded by construct-irrelevant factors (Eckes, 2011). Ranges in fit statistics from +/- .5 to 1.5 are considered acceptable (Eckes, 2011; Linacre, 2014b). FACETS also provides reliability and separation indices. The reliability index indicates the reproducibility of the measures if the test were to be administered to another randomly selected sample from the same population (Bond & Fox, 2007). Separation indicates the number of statistically distinguishable strata in the data. Finally, MFRM analyses produce a "fair average score" that accounts for rater severity. In addition to the MFRM analyses, the score distributions by item and protocol component were analyzed to examine which aspects of evidence-based reading instruction were most frequently implemented and which aspects of instruction were not implemented by the teachers in this sample.

Results

The results of the MFRM analyses are shown in Figure 1 and Tables 1 through 5. All analyses were based on a total of 4,824 assigned scores. Exact rater agreement was 52.4%. Figure 1 includes the variable map and rank order of the four facets (a) item, (b) teacher, (c) lessons, and (d) raters on a common scale. The scale along the left of Figure, titled "Measr," represents the logit scale, ranging from -2 to +2, which is estimated from the pattern of the data. Placing the facets on a common scale allows for comparisons within and among the facets (Smith & Kulikowich, 2004). A higher location on the

vertical rulers indicates less frequent implementation of items, higher proficiency for teachers, more severity for raters, and more difficulty for lessons. The column heading for items ranks the items from least implemented by teachers to most implemented by teachers and is commonly referred to in MFRM as a measure of "difficulty". Item 15, *The teacher effectively engages background knowledge and/or activates schema relevant to the text prior to reading* was implemented the least often across teachers and lessons. Items 4, *The teacher makes explicit connections between sounds and letters or letter groups* and Item 5, *The teacher clearly and accurately models articulation* were the items most often implemented. The column labeled teacher ranks the teachers in order of proficient implementation with T1 scoring the most items as proficient and T7 the least proficient. The rater column ranks raters by their level of severity, with most severe raters at the top of the scale.

Item Facet and Fit Statistics

Table 1 reports the item difficulty, fit statistics, separation and reliability indices for the item facet. As reported in Figure 1, and in more detail here, the item "difficulty" ranges from 1.81 logits (*SE*=.12) for Item 15, to -.74 logits (*SE*=.10) for Item 5. The fit statistics range from .74 (Item 6) to 1.63 (Item 15) and outfit statistics from .70 (Item 17) to 1.52 (Item 15) placing Item 15 slightly higher than the upper bound of the acceptable range of .50 to 1.50 (Eckes, 2011). Item fit statistics indicate whether raters have scored items in a consistent manner. The fit statistics for Item 15 *The teacher effectively engages background knowledge and/or activates schema relevant to the text prior to reading* indicate potential misfit. Fit statistics are sensitive to extreme values (Linacre, 2014b) and in the present analysis the higher fit statistics for Item 15 are likely the result of teachers

who performed well on the other items of the protocol receiving a low score because they did not implement this item. For example, teachers (T1, T11, T20) were the most proficient at implementing the items across the CDLP. When examining the 7% of instances where these three teachers were scored as not implementing an item, half of the 'not implemented' scores were assigned to Item 15. The item reliability of separation of .98 demonstrates that item difficulties are separated along the continuum of implementation. This separation was statistically significant with a chi-square of 592.8 and 17 degrees of freedom *(p<.001)*. These statistics demonstrate reliable differences in item difficulty.

In examining the items on the CDLP and their rank order on the variable map, as well as the overall percentages of scores received for each item (see Table 6), the rank order appears to be logical. For example, teachers were most frequently observed as proficient (50% of possible responses) on Item 5 *The teacher clearly and accurately models articulation,* with only 12% of possible responses scoring as not implemented. A majority of teachers also implemented or partially implemented Item 4 *The teacher makes explicit connections between sounds and letters or letter groups,* with only 10% of possible responses scoring as not implemented. Both of these items would be expected to have high levels of implementation in a lesson specifically targeting decoding skills, especially when teachers use scripted, evidence-based programs. The items that were least often implemented were those related to text reading, scaffolding, and comparing and contrasting learned patterns (Items 10, 13, 15, and 16).

Teacher Facet and Fit Statistics

The teacher column of Figure 1 lists teachers from most proficient (Teacher 1) at the top to least proficient (Teacher 7) at the bottom. Table 2 reports the teachers overall fair average score on the CDLP protocol, along with the fit statistics and the reliability and separation indices for the teacher facet. The teachers' proficiency with implementing the items on the CDLP ranges from 1.56 logits (*SE*=.10) for Teacher 1 who is the most proficient to -.77 logits (*SE*= .10) for Teacher 7, who is the least proficient. The fair average score, which accounts for rater severity, ranges from 2.65 for Teacher 1 to 1.63 for Teacher 7. The fit statistics measure the extent to which a teacher's pattern of responses matches that predicted by the model, and can be used to identify teachers who have been evaluated in a consistent manner. Table 2 shows that all fit statistics are within acceptable ranges (-0.5 to 1.5), indicating that the evaluation with the rubric has been consistently applied to determine teachers' ability to implement a comprehensive decoding lesson. The reliability of separation is .98, with a statistically significant chi square of 836.9 and 19 degrees of freedom *(p<.001)*. This indicates that teachers differ in their ability to proficiently implement decoding instruction as measured by the CDLP.

Rater Facet and Fit Statistics

The rater column on Figure 1 ranks the raters from the most severe (Rater 5) at the top to the most lenient rater (Rater 11) at the bottom. Table 3 shows that the raters' severity ranges from -.84 logits (*SE*=.08) to .50 logits (*SE*=.08). The fit statistics help determine whether raters are consistent with their own ratings on the protocol and can be used to identify severe or lenient ratings that are unexpected given the rater's overall scoring pattern, or used to identify biases for a particular item or teacher. Fit statistics fell

within the acceptable range. The reliability coefficient of .96, on a chi-square of 257.9 and 11 degrees of freedom *(p < .001)* along with the spread from -.84 to .50 logits suggests that raters differ in their overall ratings and severity level. The bias analysis indicated a total of 29.14% of the variance in the observations (n = 4,824) was explained by the model. 11.49% was explained by teacher/rater interactions, and 4.6% by item/rater interactions.

Table 4 presents the rank order of teachers as a measure of their average observed score across all items and lessons, and compares this to their fair average score to examine whether teacher rankings might vary as a result of having had a different set of raters. With the exception of Teachers 7 and 12, whose observed average scores differed from their fair average scores by .01 of a point, the rank order of teacher performance is consistent across observed and fair average scores, suggesting that rater severity did not have a significant impact on the ratings assigned to teachers.

Lesson Facet and Fit Statistics

As shown in Figure 1, the Lesson facet shows little variability in its range across the logit scale. The lesson facet is somewhat difficult to interpret as we did not specify the content or focus of the lessons in advance, but selected video labeled as reading with decoding instruction. Additionally, participating teachers were requested to only include video with the same group of students, as observation research has suggested that teacher performance may vary depending on class composition, and our goal in the current study was to first examine teacher performance with a consistent instructional group. Table 5 shows that each of the three lessons were of approximately the same difficulty with a range of -.03 to .03 logits. Fit statistics are all within the acceptable range of -0.5 to 1.5.

The reliability of separation of .00 was not statistically significant, suggesting lesson "difficulty" did not significantly differ.

Distribution of Scores across CDLP Components and Items

As discussed, the variable map (Figure 1) provides a rank order of the items of the CDLP, allowing for an initial understanding of which elements of effective decoding instruction were the least often provided to the SWD in this sample. Table 6 presents the items by component in the order in which they appear on the CDLP, and includes the number and percentage of assigned scores for each item, and across each component. Category statistics showed that across all items of the CDLP and the 4,824 total assigned scores, 33% were assigned a score of 3 (implemented), 41% were assigned a score of 2 (partially implemented), and 26% were assigned a score of 1 (not implemented).

When examining performance across the seven components of the CDLP, teachers in this sample had the highest level of proficient implementation (45% of assigned scores) on the *Phoneme-Grapheme Correspondence* component, which is comprised of Items 4-6. The *Systematic Instruction* component (Items 1-3), had the next highest percentage of "proficient implementation" (40%). The most problematic components for this sample of teachers and observations were *Word Meaning* (Item 12) and *Reading Connected Text* (Items 13-16), with only 22% and 25% of assigned scores of "proficient implementation", respectively. As depicted in Table 6 and indicated in Figure 1, Item 5 had the highest number of scores of "implemented" assigned (50%), and Item 15 had the highest number of scores of "not implemented" (75%).

Discussion

The purpose of this study was to test the psychometric properties of the RESET CDLP and to examine the distribution of scores across items and components that comprise effective, evidence-based practice. The CDLP was developed to allow for the reliable and accurate observation of a teacher's ability to effectively implement decoding instruction as described in the research on EBPs for SWDs. The results of the MFRM analyses suggest the CDLP will provide reliable evaluations of a teacher's ability to implement decoding instruction for SWD and will support delivery of specific and actionable feedback recommended for effective evaluation instruments (Hill & Grossman, 2013). The sensitivity of the CDLP is supported by high separation and reliability statistics dividing the items and teachers into statistically different strata (Linacre, 2014b), indicating the CDLP can reliably differentiate between both item difficulty and teacher implementation proficiency.

As an observation instrument aligned to EBP with high levels of reliability across its multiple facets, the CDLP can be part of an effective observation system of support to not only systematically improve practice, but to also promote the sustained use of practices identified in the research. For example, following a baseline evaluation using the CDLP, teachers can set goals for improvement and receive feedback specifically aligned to decoding instruction practices over time. Further, the CDLP provides common language and a framework to guide teachers' self-reflection, professional development planning and implementation.

Consistent with existing research, the scoring distribution across the CDLP items and components suggest that the teachers in this sample are not implementing reading

intervention with a level of adherence to EBPs needed to improve outcomes for SWD (Vaughn & Wanzek, 2014). The overall distribution of assigned scores from this sample of video observations suggests a need for instruments like the CDLP to inform and improve teacher practice. The score distribution across components indicate that practices to develop word meaning and reading connected text are not implemented at a level to support students' development in these areas.

Integrating word meaning and reading connected text have been identified as important and effective instructional practices for SWD (Jenkins et al., 2004; Tunmer & Chapman, 2012). Our findings are consistent with classroom observation studies reporting that students spend limited amounts of time engaging with print (Vaughn & Wanzek, 2014). It may be teachers are either not provided with sufficient time for intervention sessions, are not appropriately pacing instruction to ensure adequate time for this important practice, or are unaware of the importance of this practice to promoting stronger reading outcomes. The underlying causes and appropriate solutions can only be identified once consistent, reliable observation data are collected. In this way, observations conducted with the CDLP highlight instructional areas of concern, allowing for a set of related goals and an articulated plan of support to be put into place. Ongoing observation data provide routine progress monitoring and equip teachers with the specific, actionable feedback they need to improve practice.

In addition to the specific feedback that can be provided through an observation conducted with the CDLP, it is critical that observations of teacher practice not be subject to differences in rater severity. Our analyses indicate raters differed in their severity, with Raters 2 and 5 being the most severe in their ratings and Rater 11 the least severe. Fit

statistics for the rater facet were within acceptable range, indicating that raters were consistent in their own scoring, and the fair average scores and observed scores did not result in significant changes to overall teacher scores or to rank order. When multiple raters watch multiple teachers and multiple lessons, statistical adjustments to account for rater differences are possible. However, in practice, it is likely that only one rater will watch a teacher, and for this reason it is important to consider the implications of the low level of perfect agreement across raters.

Our findings are consistent with the broader research that highlights the difficulty of scoring the instructional aspects of observation instruments with perfect agreement (Casabianca et al., 2015), particularly when observation protocols are specific rather than generic in focus (Gitomer et al., 2014). Our findings, taken in context with those reported in the research, highlight the variability in observations and feedback provided to teachers that is likely to occur from one observer to the next. Some researchers argue exact rater agreement is unlikely even with extensive training (Casabianca et al., 2015; Eckes, 2011). This must be taken into consideration as teachers are observed and evaluated; different evaluators may have different perspectives on quality and degrees of implementation, or not share the same background knowledge about the practice being observed. Therefore, continued research on how to feasibly support raters in accurately and consistently applying the scoring criteria is needed.

Limitations

Although the results of this study are promising, there are limitations that warrant caution when generalizing results. The most significant limitation is the small sample sizes of both special education teachers (n = 20) and raters (n = 12), and the limited

representation of the samples to the larger population of special education teachers and potential raters. One benefit of using video observations as part of the larger RESET project, we can develop a video bank that will include a larger and more diverse pool of teachers and lessons. Continued studies with larger samples of teachers and raters can be conducted to verify the results reported in this manuscript. Despite these limitations, the results of our current analysis are promising.

Conclusion

Over the past several decades a significant body of research detailing and validating best practices for improving decoding, word, and text reading abilities has emerged, yet these practices are not consistently implemented in the classroom (Kent et al., 2017; Klingner et al., 2010). At the same time, reading achievement of SWD has continued to lag. Teacher observation systems offer a potential solution to bridging the research to practice gap. However, realizing the promise of observation systems will require an approach that aligns observation protocols, support systems and teacher learning opportunities (Hill & Grossman, 2013). Systems that integrate improving teacher knowledge with ongoing opportunities to practice and receive feedback on their application of EBPs have been consistently shown to be more effective in changing practice and improving student outcomes (Snyder et al., 2015). The RESET CDLP represents a first step towards developing such a system. Continued research examining the impact of using the CDLP to establish baseline levels of performance, and to provide teachers with ongoing feedback and support is needed if we are to improve teachers' ability to implement evidence-based reading instruction and to improve reading achievement for SWD.

References

Archer, A. L., & Hughes, C. A. (2010). *Explicit instruction: Effective and efficient teaching*. Guilford Press.

Blachman, B. A., Schatschneider, C., Fletcher, J. M., Francis, D. J., Clonan, S. M., Shaywitz, B. A., & Shaywitz, S. E. (2004). Effects of intensive reading remediation for second and third graders and a 1-year follow-up. *Journal of Educational Psychology*, *96*(3), 444.

Brady, S. A. (2011). *Efficacy of phonics teaching for reading outcomes: Indications from post-NRP research.* In S. A. Brady, D. Braze, & C. A. Fowler (Eds.), *New directions in communication disorders research. Explaining individual differences in reading: Theory and evidence* (p. 69–96). Psychology Press.

Cain, K., Oakhill, J., & Bryant, P. (2004). Children's reading comprehension ability: Concurrent prediction by working memory, verbal ability, and component skills. *Journal of Educational Psychology*, *96*(1), 31-46.

Cain, K. (2009). Making sense of text: Skills that support text comprehension and its development. *Perspectives on Language and Literacy*, *35*(2), 11-14.

Casabianca, J. M., Lockwood, J. R., & McCaffrey, D. F. (2015). Trends in classroom observation scores. *Educational and Psychological Measurement*, *75*(2), 311-337.

Castles, A., Rastle, K., & Nation, K. (2018). Ending the reading wars: Reading acquisition from novice to expert. *Psychological Science in the Public Interest*, *19*(1), 5-51.

Denton, C. A., Nimon, K., Mathes, P. G., Swanson, E. A., Kethley, C., Kurz, T. B., & Shih, M. (2010). Effectiveness of a supplemental early reading intervention scaled up in multiple schools. *Exceptional Children*, *76*(4), 394-416.

Denton, C. A., & Al Otaiba, S. (2011). Teaching word identification to students with reading difficulties and disabilities. *Focus on Exceptional Children*, 254245149.

Denton, C. A., Tolar, T. D., Fletcher, J. M., Barth, A. E., Vaughn, S., & Francis, D. J. (2013). Effects of tier 3 intervention for students with persistent reading

difficulties and characteristics of inadequate responders. *Journal of Educational Psychology*, *105*(3), 633-648.

Ehri, L. C., & Flugman, B. (2018). Mentoring teachers in systematic phonics instruction: Effectiveness of an intensive year-long program for kindergarten through 3rd grade teachers and their students. *Reading and Writing*, *31*(2), 425-456.

Ehri, L. C., Nunes, S. R., Stahl, S. A., & Willows, D. M. (2001). Systematic phonics instruction helps students learn to read: Evidence from the National Reading Panel's meta-analysis. *Review of Educational Research*, *71*(3), 393-447.

Eckes, T. (2011). *Introduction to many-facet rasch measurement*. Frankfurt: Peter Lang.

Fallon, L. M., Collier-Meek, M. A., Maggin, D. M., Sanetti, L. M., & Johnson, A. H. (2015). Is performance feedback for educators an evidence-based practice? A systematic review and evaluation based on single-case research. *Exceptional Children*, *81*(2), 227-246.

Foorman, B. R., Breier, J. I., & Fletcher, J. M. (2003). Interventions aimed at improving reading success: An evidence-based approach. *Developmental Neuropsychology*, *24*(2-3), 613-639.

Gitomer, D., Bell, C., Qi, Y., McCaffrey, D., Hamre, B. K., & Pianta, R. C. (2014). The instructional challenge in improving teaching quality: Lessons from a classroom observation protocol. *Teachers College Record*, *116*(6), 1-32.

Gough, P. B., & Tunmer, W. E. (1986). Decoding, reading, and reading disability. *Remedial and Special Education*, *7*(1), 6-10.

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, *77*(1), 81-112.

Hill, H., & Grossman, P. (2013). Learning from teacher observations: Challenges and opportunities posed by new teacher evaluation systems. *Harvard Educational Review*, *83*(2), 371-384.

Ho, A. D., & Kane, T. J. (2013). The Reliability of Classroom Observations by School Personnel. Research Paper. MET Project. *Bill & Melinda Gates Foundation*.

Hughes, C. A., Morris, J. R., Therrien, W. J., & Benson, S. K. (2017). Explicit instruction: Historical and contemporary contexts. *Learning Disabilities Research & Practice*, *32*(3), 140-148.

Johnson, E. S., Crawford, A. R., Moylan, L. A., & Zheng, Y. (2020). Validity of a special education teacher observation system. *Educational Assessment. 25(*1) 31-46.

Johnson, E. S., Crawford, A., Moylan, L. A., & Zheng, Y. (2018). Using evidence-centered design to create a special educator observation system. *Educational Measurement: Issues and Practice*, *37*(2), 35-44.

Jenkins, J. R., Peyton, J. A., Sanders, E. A., & Vadasy, P. F. (2004). Effects of reading decodable texts in supplemental first-grade tutoring. *Scientific Studies of Reading*, *8*(1), 53-85.

Judge, S., & Bell, S. M. (2010). Reading achievement trajectories for students with learning disabilities during the elementary school years. *Reading & Writing Quarterly: Overcoming Learning Difficulties*, *27*(1-2), 153–78. doi:10.1080/10573569.2011.532722

Kang, E. Y., & Shin, M. (2019). The contributions of reading fluency and decoding to reading comprehension for struggling readers in fourth grade. *Reading & Writing Quarterly*, *35*(3), 179-192.

Kent, S. C., Wanzek, J., & Al Otaiba, S. (2017). Reading instruction for fourth-grade struggling readers and the relation to student outcomes. *Reading & Writing Quarterly*, *33*(5), 395-411.

Klingner, J. K., Urbach, J., Golos, D., Brownell, M., & Menon, S. (2010). Teaching reading in the 21st century: A glimpse at how special education teachers promote reading comprehension. *Learning Disability Quarterly*, *33*(2), 59-74.

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.

Linacre, J. M. (2014a). *Facets 3.71. 4* [Computer software].

Linacre, J. M. (2014b). *A user guide to Facets, Rasch-model computer programs*. Chicago, IL: Winsteps.com.

Lovett, M. W., Steinbach, K. A., & Frijters, J. C. (2000). Remediating the core deficits of developmental reading disability: A double-deficit perspective. *Journal of Learning Disabilities*, *33*(4), 334-358.

Mathes, P. G., Denton, C. A., Fletcher, J. M., Anthony, J. L., Francis, D. J., & Schatschneider, C. (2005). The effects of theoretically different instruction and student characteristics on the skills of struggling readers. *Reading Research Quarterly*, *40*(2), 148-182.

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

Mislevy, R. J., Almond, R. G., & Lukas, J. F. (2003). A brief introduction to evidence-centered design. *ETS Research Report Series*, *2003*(1), i-29.

Moats, L. C., & Foorman, B. R. (2003). Measuring teachers' content knowledge of language and reading. *Annals of Dyslexia*, *53*(1), 23-45.

Moats, L. (2009). Knowledge foundations for teaching reading and spelling. *Reading and Writing*, *22*(4), 379-399.

Moody, S. W., Vaughn, S., Hughes, M. T., & Fischer, M. (2000). Reading instruction in the resource room: Set up for failure. *Exceptional children*, *66*(3), 305-316.

National Center on Education Statistics (NCES, 2019). *Reading 2019: National Assessment of Educational Progress at Grades 4 and 8.* NCES. Washington, DC: US Dept of Ed.

Perfetti, C. (2007). Reading ability: Lexical quality to comprehension. *Scientific Studies of Reading*, *11*(4), 357-383.

Perfetti, C., & Stafura, J. (2014). Word knowledge in a theory of reading comprehension. *Scientific Studies of Reading*, *18*(1), 22-37.

Schles, R. A., & Robertson, R. E. (2019). The role of performance feedback and implementation of evidence-based practices for preservice special education teachers and student outcomes: A review of the literature. *Teacher Education and Special Education*, *42*(1), 36-48.

Schulte, A. C., Stevens, J. J., Elliott, S. N., Tindal, G., & Nese, J. F. (2016). Achievement gaps for students with disabilities: Stable, widening, or narrowing on a state-wide reading comprehension test? *Journal of Educational Psychology*, *108*(7), 925-940.

Smith Jr, E. V., & Kulikowich, J. M. (2004). An application of generalizability theory and many-facet Rasch measurement using a complex problem-solving skills assessment. *Educational and Psychological Measurement*, *64*(4), 617-639.

Snyder, P. A., Hemmeter, M. L., & Fox, L. (2015). Supporting implementation of evidence-based practices through practice-based coaching. *Topics in Early Childhood Special Education*, *35*(3), 133-143.

Solomon, B. G., Klein, S. A., & Politylo, B. C. (2012). The effect of performance feedback on teachers' treatment integrity: A meta-analysis of the single-case literature. *School Psychology Review*, *41*(2), 160-175.

Steacy, L. M., Elleman, A. M., Lovett, M. W., & Compton, D. L. (2016). Exploring differential effects across two decoding treatments on item-level transfer in children with significant word reading difficulties: A new approach for testing intervention elements. *Scientific Studies of Reading*, *20*(4), 283-295.

Swanson, E. A. (2008). Observing reading instruction for students with learning disabilities: A synthesis. *Learning Disability Quarterly*, *31*(3), 115-133.

Torgesen, J. K., Alexander, A. W., Wagner, R. K., Rashotte, C. A., Voeller, K. K., & Conway, T. (2001). Intensive remedial instruction for children with severe reading disabilities: Immediate and long-term outcomes from two instructional approaches. *Journal of Learning Disabilities*, *34*(1), 33-58.

Tunmer, W. E., & Chapman, J. W. (2012). The simple view of reading redux: Vocabulary knowledge and the independent components hypothesis. *Journal of learning disabilities*, *45*(5), 453-466.

Vaughn, S., & Wanzek, J. (2014). Intensive interventions in reading for students with reading disabilities: Meaningful impacts. *Learning Disabilities Research & Practice*, *29*(2), 46-53.

Wanzek, J., & Vaughn, S. (2007). based implications from extensive early reading interventions. *School Psychology Review*, *36*(4), 541-561.

Washburn, E. K., Joshi, R. M., & Cantrell, E. B. (2011). Are preservice teachers prepared to teach struggling readers?. *Annals of Dyslexia*, *61*(1), 21-43.

Weiser, B., & Mathes, P. (2011). Using encoding instruction to improve the reading and spelling performances of elementary students at risk for literacy difficulties: A best-evidence synthesis. *Review of Educational Research*, *81*(2), 170-200.

Weiser, B. L. (2013). Ameliorating reading disabilities early: Examining an effective encoding and decoding prevention instruction model. *Learning Disability Quarterly*, *36*(3), 161-177.

Table 4.1 Item Measure Report from Many-Facet Rasch Measurement Analysis

Item Number	Difficulty (Logits)	Model SE	Infit MNSQ	Outfit MNSQ
15	1.81	.12	1.63	1.52
10	.62	.10	.95	.94
13	.56	.09	.98	.97
16	.55	.09	1.37	1.38
12	.40	.09	1.01	1.01
11	.19	.09	1.28	1.27
9	.05	.09	.92	.93
17	-.03	.09	.70	.70
7	-.10	.09	.99	1.01
14	-.22	.09	.99	1.02
2	-.29	.09	1.05	1.03
18	-.30	.09	.79	.78
3	-.33	.09	.80	.78
8	-.43	.10	1.22	1.28
6	-.46	.10	.74	.75
1	-.56	.10	.93	.90
4	-.71	.10	.81	.79
5	-.74	.10	1.06	1.08
Mean	.00	.10	1.01	1.01
SD	.62	.01	.24	.23

Note. Root mean square error (model) = .10; adjusted *SD* = .61; separation = 6.38; reliability = .98; fixed chi-square = 592.8; df = 17; significance = .00.

Table 4.2 **Teacher Measure Report from Many-Facet Rasch Measurement Analysis**

Teacher Number	Fair Average	Ability (Logits)	Model SE	Infit MNSQ	Outfit MNSQ
1	2.65	1.56	.10	1.12	1.11
11	2.64	1.52	.13	1.24	1.17
20	2.48	1.06	.10	.95	.94
5	2.21	.43	.10	.87	.87
2	2.21	.43	.10	1.24	1.24
10	2.16	.32	.10	1.17	1.23
19	2.16	.31	.10	.86	.86
6	2.15	.30	.10	1.07	1.09
17	2.09	.19	.10	.84	.83
15	2.05	.11	.10	.73	.72
4	1.99	-.02	.10	1.10	1.08
16	1.93	-.14	.10	1.18	1.27
14	1.88	-.23	.11	.84	.84
13	1.84	-.33	.10	.87	.86
9	1.80	-.42	.10	.76	.82
3	1.76	-.50	.10	1.22	1.17
18	1.69	-.64	.11	1.10	1.04
8	1.69	-.64	.10	1.00	.99
12	1.66	-.70	.11	1.05	1.13
7	1.63	-.77	.10	.87	.90
Mean	2.03	.09	.10	1.00	1.01
SD	.31	.68	.01	.16	.17

Note. Root mean square error (model) = .10; adjusted *SD* = .68; separation = 6.63; reliability = .98; fixed chi-square =836.9; df = 19; significance = .00.

Table 4.3 Rater Measure Report from Many-Facet Rasch Measurement Analysis

Rater Number	Severity	Model SE	Infit MNSQ	Outfit MNSQ
5	.50	.08	.78	.82
2	.47	.08	.98	.97
1	.32	.08	.75	.76
6	.19	.08	1.10	1.10
12	.17	.08	.97	.99
7	.12	.08	1.05	1.09
8	-.02	.08	.90	.88
9	-.11	.08	1.31	1.28
3	-.12	.08	1.01	1.05
10	-.17	.08	1.29	1.34
4	-.50	.08	.91	.95
11	-.84	.08	.92	.88
Mean	.00	.08	1.00	1.01
SD	.39	.00	.17	.18

Note. Root mean square error (model) = .08; adjusted *SD* = .38; separation = 4.87; reliability = .96; fixed chi-square = 257.9; df = 11; significance = .00.

Table 4.4 Teacher Measurement Report

Teacher	Observed Score	Fair Average Score	Measure	S.E.
7	1.7	1.63	-.77	.10
12	1.6	1.66	-.70	.11
8	1.7	1.69	-.64	.10
18	1.7	1.67	-.64	.11
3	1.8	1.76	-.50	.10
9	1.8	1.80	-.42	.10
13	1.9	1.84	-.33	.11
14	1.9	1.88	-.23	.10
16	2.0	1.93	-.14	.10
4	2.0	1.99	-.02	.10
15	2.0	2.05	.11	.10
17	2.1	2.09	.19	.10
6	2.2	2.15	.30	.10
19	2.2	2.16	.31	.10
10	2.2	2.16	.32	.10
2	2.1	2.21	.43	.10
5	2.1	2.21	.43	.10
20	2.4	2.48	1.06	.09
11	2.6	2.64	1.52	.13
1	2.6	2.65	1.56	.10

Table 4.5 Lesson Measure Report from Many-Facet Rasch Measurement Analysis

Lesson Number	Difficulty	Model SE	Infit MNSQ	Outfit MNSQ
3	.03	.04	1.00	1.02
2	.00	.04	.99	1.00
1	-.03	.04	1.01	1.00
Mean (count =3)	.00	.04	1.00	1.01
SD	.03	.00	.01	.01

Note. Root mean square error (model) = .04; adjusted *SD* = .00; separation = 0.00; reliability = .00; fixed chi-square = 1.3; df = 2; significance = .51.

Table 4.6 Score Distribution Across Components and Items of the Comprehensive Decoding Lesson Protocol

Component		Item	Number of Assigned Scores			Percentage of Assigned Scores		
			3	**2**	**1**	**3**	**2**	**1**
Systematic Instruction			323	348	133	40	43	17
	1.	Skills are taught systematically within the lesson in a logical, clearly defined, graduated sequence.	120	110	38	45	41	14
	2.	The teacher provides a focused review of word reading skills.	111	97	60	41	36	22
	3.	The teacher uses effective step by step procedures or routines with appropriate pacing.	92	141	35	34	53	13
Phoneme-Grapheme Correspondence			366	358	88	45	44	11
	4.	The teacher makes explicit connections between sounds and letters or letter groups.	125	116	27	47	43	10
	5.	The teacher clearly and accurately models articulation.	133	103	32	50	38	12
	6.	The teacher engages all students in the pronunciation of the target sound or sounds with a sufficient emphasis on accurate articulation.	100	139	29	37	52	11
Word Reading			335	455	282	31	43	26
	7.	Blending strategies focused on accurate orthographic (written) and phonological (sound) connections are used clearly and consistently throughout the lesson.	84	131	53	31	49	20
	8.	When a word is segmented, the teacher consistently ensures the word is also read as a whole word at the normal rate.	119	93	56	44	35	21

	9.	The teacher provides students with adequate practice designed to reinforce orthographic (written) and phonological (sound) connections aligned to the target skill.	76	128	64	28	48	24
	10.	The teacher guides students to compare and contrast learned patterns.	56	103	109	21	38	41
Encoding			78	108	82	29	40	31
	11.	The teacher explicitly reinforces precise letter-sound correspondence through encoding exercises aligned to the target skill(s).	78	108	82	29	40	31
Word Meaning			59	118	91	22	44	34
	12.	The teacher effectively integrates word meaning into the lesson.	59	118	91	22	44	34
Reading Connected Text			268	313	491	25	29	46
	13.	The teacher scaffolds the transfer of new word reading skills to text reading as needed for students to experience success.	63	94	111	24	35	41
	14.	The teacher provides sufficient opportunities for all students to engage in reading decodable text.	98	112	58	37	42	22
	15.	The teacher effectively engages background knowledge and/or activates schema relevant to the text prior to reading.	32	36	200	12	13	75
	16.	The teacher effectively scaffolds meaning and understanding through questioning and/or discussion appropriate to the text.	75	71	122	28	26	46
Monitoring and Feedback			162	286	88	30	53	17
	17.	Throughout the lesson the teacher provides affirmative and corrective feedback consistently focused on	69	151	48	26	56	18

		reinforcing the application of word reading skills and strategies.					
	18.	When errors are detected, the teacher consistently elicits the correct response from the student throughout the lesson.	93	135	40	35	50	15
Total			**1583**	**1986**	**1255**	**33**	**41**	**26**

Measr	-Items	+Teacher	-Rater	-Lesson	Scale
2					(3)
	Item 15				
		T1			
		T11			
		T20			---
1					
	Item 10 Item 13 Item 16				
			2 5		
	Item 12	T2 T5			
		T10 T19 T6	1		
	Item 11	T17	12 6		
	Item 9	T15	7		
0	Item 17	T4	8	1 2 3	2
	Item 7	T16	3 9		
	Item 14	T14	10		
	Item 18 Item 2 Item 3	T13			
	Item 8	T9			
	Item 6	T3	4		
	Item 1	T18 T8			
	Item 4 Item 5	T12			
		T7	11		
-1					(1)

Figure 4.1 Variable map of the CD rubric facets items, teachers, raters, and lessons.

CHAPTER FIVE

Summary

Rigorously developed teacher observation systems aligned to the content specific instructional practices found to be effective for SWD have the potential for improving teacher practice and ultimately, outcomes for students. Such systems can provide a framework for developing a shared understanding about effective practices and provide teachers with accurate, actionable, specific feedback designed to support growth. Observations of classroom instruction consistently show a need for improvements in instructional content and delivery. The purpose of this body of work was to develop special education observation protocols detailing the elements of evidence-based practices for decoding and comprehension instruction, using deliberate approaches to development, and rigorous evaluation of the multiple facets which impact a teacher's observed scores.

Chapter Two provided an introduction to the development of the larger RESET observation system using Evidence-Centered Design (ECD) to create a reliable and sound observation system. In this chapter, the five stages of the ECD framework were explained in the context of the RESET observation system. The RESET Explicit Instruction protocol is used to illustrate the assessment implementation and assessment delivery stages of the ECD process. Two studies are described. The first study describes the inductive approach used in the development of performance level descriptors. The second study describes the analysis of the fully developed Explicit Instruction protocol using

MFRM, with results indicating the development of a psychometrically sound instrument. This process is applied in later studies to other content areas to develop observation instruments designed with the rigor and structure needed to achieve the goal of improving practice.

Chapter Three is a study describing the development of the Reading for Meaning RESET observation protocol. This protocol details the evidence-based practices for comprehension instruction extracted from the research and tested the psychometric properties of the protocol using MFRM. In this paper the elements of effective comprehension instruction for SWD are discussed. The procedures for testing the reliability of the protocol include video recorded comprehension lessons which are rated by a set of content area experts trained in the Reading for Meaning protocol. Data were analyzed using MFRM, with results indicating the development of a protocol that will provide reliable evaluations of a teacher's ability to implement the components of reading comprehension instruction as they are presented in the Reading for Meaning protocol.

Chapter Four describes the development of the RESET Comprehensive Decoding Lesson Protocol (CDLP). The purposes of this study were to test the psychometric properties of the CDLP and to analyze the implementation of practices by examining the distribution of scores across the items indicating effective decoding instruction. In this paper the components of effective decoding instruction for SWD are explained. The study procedures for testing the protocol follow a similar pattern with video recorded lessons identified as decoding instruction rated by a set of content area experts trained in the use of the CDLP. Data analyzed using MFRM indicate the CDLP will provide reliable evaluations of a teacher's ability to implement the elements of decoding

instruction presented in the CDLP and provide a framework for specific and actionable feedback for teachers to improve or sustain their practices. Consistent with what has been reported in observational studies, the scoring distribution across the CDLP items suggest teachers in this sample are not consistently implementing decoding intervention to the degree necessary for SWD to be successful.

In conclusion, this collection of work represents important steps toward developing a special education teacher observation system with the potential for improving practice and ultimately outcomes for students. Findings from each article demonstrate we are able to develop reliable instruments for the purpose of providing accurate teacher evaluation and feedback and supporting on-going professional development, through a well-defined framework and the use of rigorous processes. While not without limitations, this is a promising development moving toward improving instructional practice for SWD.

APPENDIX

Rubrics

Explicit Instruction Rubric

RESET Explicit Instruction Rubric - 2017-18

Components	Item	3 - Implemented	2 - Partially Implemented	1 - Not Implemented
Identifying and Communicating Goals	1	The goals of the lesson **are clearly** communicated to the students.	The goals of the lesson are **not clearly** communicated to the students.	The goals of the lesson are **not** communicated to the students.
	2	The stated goal(s) is/are **specific.**	The stated goal(s) is/are **broad or vague.**	There is **no stated goal**.
	3	The teacher **clearly** explains the relevance of the stated goal to the students.	The teacher **tries to explain** the relevance of the stated goal to the students, but the **explanation is unclear or lacks detail.**	The teacher **does not explain** the relevance of the stated goal to the students.
Alignment	4	Instruction is **completely aligned** to the stated or implied goal.	Instruction is **partially or loosely aligned** to the stated or implied goal.	Instruction is **not aligned** to the stated or implied goal.
	5	**All** of the examples or materials selected **are aligned** to the stated or implied goal.	**Some** of the examples or materials **are aligned** to the stated or implied goal; **OR** examples and materials are **somewhat aligned** to the stated or implied goal.	Examples or materials selected **are not aligned** to the stated or implied goal.
	6	Examples or materials selected **are aligned** to the instructional level of **most or all** of the students.	Examples or materials selected **are aligned** to the instructional level of **some** of the students.	Examples or materials selected **are not aligned** to the instructional level of **most students**.
Teaching Procedures	7	The teacher **effectively** reviews prior skills and/or engages background knowledge **before beginning** instruction.	The teacher reviews prior skills and/or engages background knowledge **before beginning** instruction, **but not effectively.**	The teacher **does not** review prior skills and/or engage background knowledge **before beginning** instruction.
	8	The teacher **provides clear** demonstrations of proficient performance.	The teacher **does not provide clear** demonstrations of proficient performance.	The teacher **does not provide any** demonstrations of proficient performance.

	9	The teacher **provides an adequate number** of demonstrations given the nature and complexity of the skill or task.	The teacher **does not provide an adequate number** of demonstrations given the nature and complexity of the skill or task.	The teacher **does not provide** demonstrations.
	10	The teacher uses language that is **clear, precise, and accurate** throughout the lesson.	The teacher uses language that is **not always clear, precise, and accurate**.	The teacher uses language that is **confusing, unclear, imprecise, or inaccurate** throughout the lesson.
	11	Scaffolding is provided **when it is needed** to facilitate learning.	**Some** scaffolding is provided, but **more is needed** to facilitate learning.	Scaffolding **is needed, but minimal or no** scaffolding is provided to facilitate learning.
	12	Complex skills or strategies **are broken down** into logical instructional units to address cognitive overload, processing demands, or working memory.	Complex skills or strategies **are not effectively broken down** to address cognitive overload, processing demands, or working memory.	Complex skills and strategies **are not broken down as needed** into logical instructional units to address cognitive overload, processing demands, or working memory.
	13	The teacher **systematically withdraws** support as the students move toward independent use of the skills.	The teacher withdraws support, but **it is not withdrawn systematically.**	The teacher **does not withdraw** support; **OR** the teacher provides very limited support and then **abruptly withdraws** it.
Guided Practice	14	Guided practice is **focused** on the application of skills or strategies related to the stated or implied goal.	Guided practice is **somewhat focused** on the application of skills or strategies related to the stated or implied goal.	Guided practice is **not focused** on the application of skills or strategies related to the stated or implied goal.
	15	The teacher **consistently prompts** students to apply skills or strategies throughout guided practice.	The teacher prompts students to apply skills or strategies, but **not consistently OR not effectively** throughout guided practice.	The teacher **does not prompt** students to apply skills or strategies throughout guided practice.
Pacing	16	The teacher maintains an **appropriate** pace **throughout the lesson**.	The teacher maintains an **appropriate** pace during **some of the lesson**.	The teacher maintains an **inappropriate** pace **throughout the lesson**.

	17	The teacher **allows adequate** time for students to think or respond **throughout** the lesson.	The teacher **sometimes allows adequate** time for students to think or respond but **inconsistently throughout** the lesson.	The teacher **never allows** adequate time to students to think or respond.
	18	The teacher **maintains focus** on the stated or implied goal **throughout** the lesson.	The teacher **inconsistently focuses** on the stated or implied goal.	The teacher **does not focus** on the stated or implied goal.
Engagement	19	The teacher **provides frequent** opportunities for students to engage or respond during the lesson.	The teacher **provides limited** opportunities for students to engage or respond during the lesson.	The teacher **does not provide** opportunities for students to engage or respond during the lesson.
	20	There are **structured and predictable** instructional routines throughout the lesson.	Instructional routines are **not consistently** applied throughout the lesson.	There is **no** instructional routine.
	21	The teacher **monitors** students to ensure they remain engaged.	The teacher **monitors inconsistently** throughout the lesson; **OR** the teacher **does not consistently** monitor **all students** to ensure they remain engaged.	The teacher **does not** monitor students to ensure they remain engaged.
Monitoring and Feedback	22	The teacher **consistently** checks for understanding **throughout the lesson**.	The teacher only checks **some students** for understanding; **OR** the teacher does **not consistently** check for understanding throughout the lesson.	The teacher does **no or very minimal** checking for understanding.
	23	The teacher provides timely feedback **throughout the lesson**.	The teacher **occasionally** provides timely feedback.	The teacher **does not** provide feedback; **OR** it is **not timely**.
	24	Feedback is specific and informative **throughout** the lesson.	Feedback is **not consistently** specific and informative throughout the lesson.	There is **no** feedback; **OR** it is **not at all** specific and informative.
	25	The teacher **makes** adjustments to instruction **as needed** based on the student responses.	The teacher **makes some** adjustments to instruction as needed based on the student responses, **but more adjustments are needed.**	The teacher **does not make** adjustments to instruction **as needed** based on the student responses.

Reading for Meaning Protocol

RESET Comprehension - Reading for Meaning

Components	Item	3 Implemented	2 Partially Implemented	1 Not Implemented
Preparing to Read **Purpose for Reading**	**1**	The teacher communicates a **content specific purpose** for reading the text.	The teacher communicates a purpose for reading the text, but the purpose is **broad, vague, or not specific to the content of the text.**	The teacher **does not communicate** a purpose for reading the text.
	2	The purpose for reading **is sustained** throughout the lesson.	The purpose for reading is **inconsistently sustained** throughout the lesson.	The purpose for reading is **not sustained** throughout the lesson.
Preparing to Read **Background and Schema**	**3**	The teacher **effectively engages** background knowledge and/or activates schema relevant to the text prior to reading.	The teacher **attempts to engage** background knowledge and/or activate schema **but does not** maintain the focus on relevant information.	The teacher **does not engage** background knowledge and/or activate schema relevant to the text prior to reading.
	4	The teacher **effectively** pre-teaches or reviews **key** concepts.	The teacher pre-teaches or reviews **key** concepts **but not effectively.**	The teacher **does not** pre-teach or review **key** concepts.
	5	The teacher **purposefully uses** text preview strategies that are **focused** on text structure **and aligned** with the purpose for reading.	The teacher **uses** text preview strategies that are **somewhat focused** on text structure **and aligned** with the purpose for reading.	The teacher **does not use** text preview strategies; **OR** text preview is **not at all focused** on text structure and purpose for reading.
	6	The teacher reviews or teaches **key** vocabulary **prior to reading** using words that are clear, precise, and accurate.	The teacher reviews or teaches **some key** vocabulary **as they are encountered AND/OR** uses words that are not always clear, precise, and accurate.	The teacher **does not** review or teach **key** vocabulary.

Reading for Meaning and Monitoring Understanding	**7**	The teacher **actively engages** students in the use of content enhancement tools that are **aligned** to facilitate comprehension (e.g., advanced and graphic organizers, visual displays, mnemonic instruction).	The teacher **provides** content enhancement tools that are **aligned** to facilitate comprehension **but does not actively engage** students in their use.	The teacher **does not provide** content enhancement tools at all; **OR** the teacher **provides** content enhancement tools that are **not aligned** to facilitate comprehension **AND/OR** refers to content enhancement tools **but does not implement** them.
	8	The teacher **focuses** attention on relevant text features and/or structures **to organize thinking and support comprehension.**	The teacher **points out some** text features and/or structures **but does not deliberately use** them to organize thinking and support comprehension.	The teacher **does not use** text features and/or structures.
	9	The teacher **guides** students to make predictions about the text **AND** to confirm, disconfirm, and/or extend them.	The teacher **asks** students to make predictions **AND** gives the opportunity to confirm, disconfirm, and/or extend them **but without adequate guidance** (e.g., lacks connection to relevant information or background knowledge).	The teacher **does not ask** students to make predictions; **OR** the teacher **does not provide the opportunity** to confirm, disconfirm, or extend predictions that are made.
	10	The teacher **supports** the students in identifying the main idea and supporting details.	The teacher provides **some support** for identifying main idea and supporting details **but more is needed** (e.g., lacks clear process).	The teacher **does not support** the identification of main idea and supporting details.
	11	The teacher **guides** students to summarize key ideas and/or critical passages to support understanding.	The teacher provides **some guidance** for summarizing, **but more is needed** (e.g. focus, structure, more opportunity).	The teacher **does not guide** students to summarize key ideas and/or critical passages to support understanding.
	12	The teacher **supports** making inferences by helping students identify and connect relevant information, fill gaps, and/or connect to prior knowledge.	The teacher **supports** making inferences **but more support is needed (**e.g. identify and connect relevant information, fill gaps, and/or connect to prior knowledge).	The teacher **does not support** making inferences.

	13	The teacher **guides** students to support their responses with information from the text.	The teacher **guides** students to support their responses with information from the text, **but more guidance is needed.**	The teacher **does not guide** students to support their responses with information from the text.
	14	The teacher **consistently guides** students to reread **as needed** to support comprehension.	The teacher **misses some opportunities** for students to reread as needed to support comprehension **AND/OR** does **not always provide sufficient guidance.**	The teacher **does not guide** students to reread **as needed** to support comprehension.
	15	The teacher **consistently** cues or provides correction of decoding or word level errors **as needed AND** has the student reread the word correctly.	The teacher **inconsistently** cues or provides correction of decoding or word level errors **AND/ORinconsistently** has the student reread the word correctly.	The teacher **does not** cue or provide correction of decoding or word level errors**OR does not** have the student reread the word correctly; **OR**the teacher has selected a text that is **not at the instructional level** of most students and decoding errors inhibit comprehension.
Questioning and Discussion Practices	**16**	The teacher's questioning practices **effectively** promote understanding, guide, and focus the reading.	The teacher's questioning practices **somewhat** promote understanding, guide, and focus the reading.	The teacher's questioning practices **do not** promote understanding, guide, and focus the reading; **OR** the teacher **does not ask** questions.
	17	The teacher asks questions using wording that is **consistently understandable** for the students (e.g. clear, not too long, avoid multiple questions within a question).	The teacher asks questions using wording that is **not always understandable** for the students.	The teacher asks questions using wording that is **confusing** for the students (e.g., unclear, too long, multiple questions within a question); **OR** the teacher does not ask questions.
	18	The teacher **consistently and accurately** uses academic language (e.g., predict, compare, contrast, infer).	The teacher **uses** academic language but **not consistently AND/OR not always accurately.**	The teacher **does not use** academic language **OR** uses it **inaccurately.**

Comprehensive Decoding Instruction Protocol

RESET Comprehensive Decoding Rubric

Components	Item	3 Implemented	2 Partially Implemented	1 Not Implemented
Systematic Instruction	1	Skills are **taught systematically** within the lesson in a logical, clearly defined, graduated sequence.	Skills are **taught somewhat systematically** within the lesson in a logical, clearly defined, graduated sequence.	Skills are **not taught systematically** within the lesson in a logical, clearly defined, graduated sequence; **instruction is incidental.**
	2	The teacher provides a **focused** review of word reading skills.	The teacher provides a review, but the review is **limited or lacking in focus**.	The teacher **does not** provide a review.
	3	The teacher uses **effective** step-by-step procedures or routines with **appropriate** pacing.	The teacher uses step-by-step procedures or routines that are **somewhat** effective **AND/OR not always** paced appropriately.	The teacher **does not** use effective step-by-step procedures or routines throughout instruction, **OR** pacing **negatively impacts** learning.
Phoneme-Grapheme Correspondence	4	The teacher makes **explicit** connections between sounds and letters or letter groups.	The teacher makes connections between sounds and letters or letter groups but **not always explicitly.**	The teacher **does not** make explicit connections between sounds and letters or letter groups, **OR** connections are inaccurate.
	5	The teacher **clearly and accurately** models articulation.	The teacher models articulation but **not always clearly**.	The teacher **does not** model articulation **OR** models **inaccurately.**
	6	The teacher engages **all** students in the pronunciation of the target sound or sounds with a **sufficient emphasis** on accurate articulation.	The teacher **engages some, but not all,** students in the pronunciation of the target sound or sounds **OR** does **not sufficiently** emphasize accurate articulation**.**	The teacher **does not** engage students in the pronunciation of the target sound or sounds with an emphasis on accurate articulation OR allows for **inaccurate articulation.**
Word Reading	7	Blending strategies focused on accurate orthographic (written) and phonological (sound) connections are used **clearly and consistently** throughout the lesson.	Blending strategies focused on accurate orthographic (written) and phonological (sound) connections are used but **not always clearly and/or consistently** throughout the lesson.	Blending strategies focused on accurate orthographic (written) and phonological (sound) connections **are not used** throughout the lesson.

	8	When a word is segmented, the teacher **consistently** ensures the word is also read as a whole word at the normal rate.	When a word is segmented, the teacher **inconsistently** ensures the word is also read as a whole word at the normal rate.	When a word is segmented, the teacher **does not** ensure the word is also read as a whole word at the normal rate **OR** words are not segmented.
	9	The teacher provides students with **adequate** practice designed to reinforce orthographic (written) and phonological (sound) connections aligned to the target skill.	The teacher provides students with **somewhat adequate** practice designed to reinforce orthographic (written) and phonological (sound) connections aligned to the target skill.	The teacher provides students with **inadequate** practice designed to reinforce orthographic (written) and phonological (sound) connections aligned to the target skill.
	10	The teacher **guides** students to compare and contrast learned patterns.	The teacher provides students with the opportunity to compare and contrast learned patterns but **without appropriate guidance**.	The teacher **does not** provide students with the opportunity to compare and contrast learned patterns.
Encoding	11	The teacher **explicitly reinforces** precise letter-sound correspondence through encoding exercises **aligned** to the target skill(s). • Writing (letters, words or sentences) AND/OR • Using manipulatives to build words (tiles, cards)	The teacher engages students in encoding exercises that are **not aligned** to the target skills, **OR** the teacher **does not explicitly reinforce** precise letter-sound correspondence.	The teacher **does not** engage students in encoding exercises.
Word Meaning	12	The teacher **effectively** integrates word meaning into the lesson.	The teacher integrates word meaning into the lesson, but important **opportunities are missed.**	The teacher does **not effectively** integrate word meaning into the lesson.
Reading Decodable Text	13	The teacher scaffolds the transfer of new word reading skills to text reading **as needed** for students to experience success.	The teacher **provides some** scaffolding for the transfer of new word reading skills to text reading, but **more is needed**.	The teacher **does not scaffold** the transfer of new word reading skills to text reading.
	14	The teacher provides **sufficient** opportunities for **all** students to engage in reading decodable text.	The teacher provides **limited** opportunities for students to engage in reading decodable text, AND/OR **not all students** are engaged.	The teacher **does not** provide opportunities for students to engage in reading decodable text, **OR** the text is not decodable for most of the students.

	15	The teacher **effectively** engages background knowledge and/or activates schema relevant to the text prior to reading.	The teacher a**ttempts** to engage background knowledge and/or activate schema relevant to the text prior to reading but **not effectively**.	The teacher **does not** engage background knowledge and/or activate schema relevant to the text prior to reading.
	16	The teacher **effectively** scaffolds meaning and understanding through questioning and/or discussion appropriate to the text.	The teacher **somewhat** scaffolds meaning and understanding through questioning and/or discussion appropriate to the text.	The teacher **does not** scaffold meaning and understanding through questioning and/or discussion appropriate to the text.
Monitoring and Feedback Throughout the Lesson	17	**Throughout the lesson** the teacher provides affirmative and corrective feedback **consistently** focused on reinforcing the application of word reading skills and strategies.	**Throughout the lesson** the teacher provides **some** affirmative and/or corrective feedback reinforcing the application of word reading skills and strategies but **more is needed**.	**Throughout the lesson** the teacher **does not** provide feedback OR feedback is **not focused** on reinforcing the application of word reading skills and/or strategies.
	18	When errors are detected, the teacher **consistently** elicits the correct response from the student **throughout the lesson.** **OR** **No errors** are made by the student(s) throughout the lesson.	When errors are detected, the teacher **inconsistently** elicits the correct response from the student **throughout the lesson.**	When errors are detected, the teacher **does not** elicit the correct response from the student **throughout the lesson.**