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

Citation: 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.

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

Developing a Comprehension Instruction Observation Rubric for Special Education Teachers

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, 2011; Schulte et al., 2016, Vaughn & Wanzek, 2014; Wei, Blackorby & Schiller, 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, Urbach, Golos, Brownell & Menon, 2010; Swanson, 2008; Vaughn & Wanzek, 2014). Inadequate instruction has been defined by (a) the limited amount of time that students actually spend reading (Kent, Wanizek, & Al Otaiba, 2012; Vaughn et al., 2002); (b) the limited opportunity for active response and an emphasis on passive learning (Wanzek, Roberts, & Al Otaiba, 2013); 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, Bryk, & Dexter, 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, Braslow, Charalambous, & Hill, 2017; Grossman, Compton, Igra, Ronfeldt, Shahan & Williamson, 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, Ciullo, Vaughn, Pyle, Hassaram & Leroux, 2012). The purposeful use of content enhancement tools provides students with a framework that helps them attend to, organize and retrieve important information (Ciullo, Lo, Wanzek, Reed, 2016). Content enhancement tools aligned with the text structure scaffold the reader's use of important information and support understanding and memory (Gersten, Fuchs, Williams, Baker, 2001; Kim, Linan-Thompson & Misquitta, 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, Klingner, & Gryant 2001). Strategy instruction has been found to be most effective when it includes practice to transfer strategies across texts (Gersten, Fuchs, Williams, Baker 2001). A significant body of research supports the use of these strategies for SWD (Berkeley, Scruggs &

Mastropieri, 2010; Ciullo, Lo, Wanzek & Reed, 2016; El Zein, Solis, Vaughn, McCulley, 2014; Gajria, Jitendra, Sood & Sacks, 2007; Kim, Linan-Thompson & Misquitta, 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, Pollini, 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, Diamond, & Gutlohn, 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 (Alvis, Kennedy, Brown & Solis, 2015; Gajria, Jitendra, Sood & Sacks, 2007; Kaldenberg, Watt, & Therrien, 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, Edwards, Sacks & Jacobson, 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, Goodwin, Bryant & Higgins, 2003; Elleman, Lindo, Morphy & Compton, 2009; Elleman, Steacy, Olinghouse & Compton, 2017; Jitendra et al., 2004; O'Connor et al., 2017). It is often the case that SWD 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). 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, Miller, Elleman & Steacy, 2013; Elleman & Compton, 2017; Kendeou & van den Broek, 2007; McKeown, Beck and Blake, 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, 2013, 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; Garnham & Oakhill, 2014; 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. 2016; 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, Swanson, Vaughn, Roberts & Fall, 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, Scruggs, & Mastropieri, 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, Alber-Morgan, Cullen, & Rouse, 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, Hoppes & Yin, 2000), summarize key ideas or critical passages (Kim, Linan-Thompson, & Misquitta 2012; Solis, Ciullo, Vaughn, Hassaram, & Leroux 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.

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, Lockwood & MCCaffrey, 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).

Purpose of the Current Study

Teacher observations are high stakes assessments because they are used to make critical decisions about teachers' employment status (Adnot, Dee, Katz & Wyckoff, 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.

Methods

Participants

Special education teachers. Ten special education teachers from 3 states (Idaho, Wisconsin, Florida) each provided 3 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. 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. 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 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 3 raters, and six of the 30 videos scored by 4 raters. Raters scored each item for each video, providee 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. MFRM allows for bias analysis of the scores to examine the discrepancy between observed and expected scores according to the raters' severity levels. 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 A), 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*. Most responses for this item were scored as not implemented (67%). 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, Patrick, Strati & Wesson, 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 = 5.59; for teachers, the reliability coefficient was .92, separation = 3.47. These statistics demonstrate reliable differences in item difficulty and teacher proficiency. For lessons, the reliability coefficient was .92, separation = 3.36, showing a discrimination across lessons. The reliability coefficient for raters was .95, separation = 4.61, 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. 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 x rater bias interactions, and 5.7% by item x 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 each of the items to student growth. 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 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. 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, 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, Arguelles, Vaughn, Hughes & Klingner, 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

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.

Anderson, R. C., & Freebody, P. (1981). Vocabulary knowledge. In 1. T. Guthrie (Ed.), Comprehension and teaching: Research reviews (pp. 77-117). Newark. *DE: International Reading Association*.

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

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.

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.

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

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

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.

Chiesi, H. L., Spilich, G. J., & Voss, J. F. (1979). Acquisition of domain-related information in relation to high and low domain knowledge. *Journal of verbal learning and verbal behavior*, *18*(3), 257-273.

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.

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.

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

Mantzicopoulos, P., Patrick, H., Strati, A., & Watson, J. S. (2018). Predicting kindergarteners' achievement and motivation from observational measures of teaching effectiveness. *Journal of Experimental Education, 86(2),* 214-232.

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.

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.

Patrick, H., & Mantzicopoulos, P. (2016). Is effective teaching stable? *Journal of Experimental Education, 84*(1), 23-47.

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.

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., Roberts, G., & Al Otaiba, S. (2014). Academic responding during instruction and reading outcomes for kindergarten students at-risk for reading difficulties. *Reading and writing*, *27*(1), 55-78.

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