



# The influence of language knowledge and test components on reading comprehension scores

Alyson A. Collins<sup>1</sup> · Esther R. Lindström<sup>2</sup> · Micheal Sandbank<sup>3</sup>

Received: 5 February 2020 / Accepted: 10 December 2020 / Published online: 13 January 2021  
© The International Dyslexia Association 2021

## Abstract

This study investigated the dependability of reading comprehension scores across different text genres and response formats for readers with varied language knowledge. Participants included 78 fourth-graders in an urban elementary school. A randomized and counterbalanced  $3 \times 2$  study design investigated three response formats (open-ended, multiple-choice, retell) and two text genres (narrative, expository) from the Qualitative Reading Inventory (QRI-5) reading comprehension test. Standardized language knowledge measures from the Woodcock Johnson III Tests of Achievement (Academic Knowledge, Oral Comprehension, Picture Vocabulary) defined three reader profiles: (a)  $< 90$  as emerging, (b)  $90\text{--}100$  as basic, and (c)  $> 100$  as proficient. Generalizability studies partitioned variance in scores for reader, text genre, and response format for all three groups. Response format accounted for 42.8 to 62.4% of variance in reading comprehension scores across groups, whereas text genre accounted for very little variance (1.2–4.1%). Single scores were well below a 0.80 dependability threshold (absolute phi coefficients = 0.06–0.14). Decision studies projecting dependability achieved with additional scores varied by response format for each language knowledge group, with very low projected dependability on open-ended and multiple-choice scores for readers with basic language knowledge. Multiple-choice scores had similarly low projected dependability levels for readers with emerging language knowledge. Findings evidence interactions between reader language knowledge and response format in reading comprehension assessment practices. Implications underscore the limitations of using a single score to classify readers with and without proficiency in foundational skills.

**Keywords** Assessment · Language knowledge · Reading comprehension · Reliability

Identifying specific reading comprehension deficits (S-RCD) is challenging due to a number of contributing factors influencing this diagnosis. To improve understanding of these factors, prior research has examined reader characteristics resulting in different profiles of S-RCD

---

✉ Alyson A. Collins  
alysonacollins@txstate.edu

(e.g., Compton, Fuchs, Fuchs, Lambert, & Hamlett, 2012; Keenan et al., 2014). Such investigations aim to identify strengths and weaknesses among students with S-RCD and ultimately to enhance identification accuracy and intervention. Empirical simulations of these methods indicate strong specificity in distinguishing between students with and without learning disabilities in certain academic areas (Stuebing, Fletcher, Branum-Martin, & Francis, 2012), although some conflicting evidence suggests inconsistent diagnosis remains when applying different methods (Keenan et al., 2014) and using contrasting tests (Colenbrander, Nickels, & Kohlen, 2017; Keenan & Meenan, 2014). Still, research on this topic has further translated to common practice, with school psychologists reporting use of profile analysis to identify patterns of deficits contributing to specific learning disabilities, such as S-RCD, and identifying this method as a preferred diagnostic tool (Pfeiffer, Reddy, Kletzel, Schmelzer, & Boyer, 2000). Therefore, research examining S-RCD profiles offers promise toward improving dependability of identification methods.

Research on profiles of students with S-RCD identifies language knowledge difficulties as a predominant characteristic that may exacerbate difficulties for up to one-third of students with S-RCD (Nation, 2019) and gain importance with age (Francis, Fletcher, Catts, & Tomblin, 2005; Vellutino, Tunmer, Jaccard, & Chen, 2007). Specifically, early language development difficulties may impede reading comprehension performance later in elementary school, even when early S-RCD is not evident (Catts, Compton, Tomblin, & Bridges, 2012). However, the role of language knowledge in S-RCD identification and diagnosis is less clear.

First, language knowledge encompasses multiple skills associated with reading comprehension, including general academic knowledge, listening comprehension, and vocabulary. A reader's accumulation of prior content-relevant knowledge (i.e., academic knowledge) contributes to their reading comprehension performance (Kintsch & Rawson, 2007; Perfetti & Stafura, 2014), and students with less content knowledge are more likely to struggle with expository texts (Best, Floyd, & McNamara, 2008). Likewise, elementary students' vocabulary knowledge (including breadth and word meanings) influences their performance on reading comprehension measures (Tannenbaum, Torgesen, & Wagner, 2006). Therefore, weak vocabulary knowledge may further contribute to challenges with reading comprehension for students with S-RCD (Spencer, Quinn, & Wagner, 2014). In addition, the ability to construct meaning from oral language (i.e., listening comprehension) is directly related to reading comprehension (see Hogan, Adlof, & Alonzo, 2014) and gains prominence in upper elementary children with S-RCD (Catts, Hogan, & Adlof, 2005). In the present study, we examined these three variables (general academic knowledge, listening comprehension, vocabulary) as a composite score of language knowledge to distinguish their contribution to reading comprehension scores, apart from code and fluency-based reading abilities.

A second complicating factor in understanding how language knowledge may inform identification of S-RCD stems from prior studies reporting discrepancies in S-RCD diagnoses on reading comprehension measures. Previous studies report the overlap of S-RCD diagnoses of the same students ranged from as low as 25% (Cutting & Scarborough, 2006; see also Keenan & Meenan, 2014) to as high as 85% (Colenbrander et al., 2017), depending on the tests administered. Language knowledge partially explained these discrepancies, accounting for a large proportion of score variance in certain tests, but not others (Cutting & Scarborough, 2006; Francis, Fletcher, Catts, & Tomblin, 2005; Keenan, Betjemann, & Olson, 2008; Nation & Snowling, 1997; Spear-Swerling, 2004). Researchers attributed these differential contributions of language knowledge across reading comprehension measures to the lack of equivalence among test components that pertain to the texts and assessment methods. Specifically,

assessments that involved longer, multi-paragraph texts and used multiple-choice or open-ended response formats (including retell) to assess students relied more heavily on a reader's language knowledge over other foundational reading skills (see Keenan, 2014). Study methods did not allow examination of effects of response format, isolated from these other contributing variables.

Interactions with test components further complicate the contribution of language knowledge to reading comprehension performance. Research indicates that reader abilities and text types (i.e., narrative vs. expository text genres) may interact to produce differential reading comprehension scores for students in upper elementary and middle school grades (Best et al., 2008; Eason, Goldberg, Young, Geist, & Cutting, 2012; Miller et al., 2014). Recent studies have investigated contributions of question types and response formats to achievement, specifically examining how these test components interact with reader abilities such as language knowledge (Collins et al., 2020; Kulesz, Francis, Barnes, & Fletcher, 2016; Reardon, Kalogrides, Fahle, Podolsky, & Zárate, 2018; Spencer et al., 2019; Walker, 2017). Yet, the extent to which reader abilities such as language knowledge predict student scores may change across distributions of performance levels (Hua & Keenan, 2017), further substantiating the potential for reader and test component interactions in reading comprehension assessment. Overall, more research is needed to examine characteristics of readers, texts, and activities together to better understand how reading comprehension differs under various conditions and circumstances.

## Purpose of the present study

The present study investigated the contribution of language knowledge, test components, and interactions among these variables to variance in reading comprehension scores. By estimating the influence of reader and test components, we aimed to further understand how S-RCD profiles differ across tests as a result of the potential reader and test interactions. In a previous study, Collins, Compton, Lindström & Gilbert(2020) conceptualized variance in reading comprehension scores using the RAND framework (RAND Reading Study Group, 2002). This model suggests a reader's abilities in foundational skills (e.g., decoding) and cognitive competencies (e.g., working memory) contribute to reading comprehension, with strengths and weaknesses in these skills predictive of assessment performance (e.g., Keenan et al., 2008; Spencer, Gilmour, Miller, Emerson, Saha & Cutting 2019). Furthermore, the RAND framework suggests text complexities (e.g., reading level, genre) and activities involved in reading comprehension assessment (e.g., response format, reading mode) alter demands placed on the reader, which interact with reader abilities to produce differential assessment performance (Eason et al., 2012; Francis, Kulesz, & Benoit, 2018; Hua & Keenan, 2017; Kulesz et al., 2016; Miller, Davis, Gilbert, Cho, Toste, Street & Cutting 2014; Spencer et al., 2019; Wixson, 2017).

In the previous study, Collins et al., (2020) applied this framework, in conjunction with item response and crossed random effects methodologies to identify statistically significant predictors of reading comprehension scores on the Qualitative Reading Inventory (QRI-5; Leslie & Caldwell, 2011). Analyses examined contributions of reader abilities, text genre, and response format to variance in reading comprehension scores for fourth-grade students of varying skill levels. Results identified response format as one test component contributing to differential effects on student outcomes. Text genre (narrative and expository) also predicted

differences in retell scores, with students performing less well on expository texts regardless of specific reader abilities. Most relevant to the present study, language knowledge—represented as composite of general academic knowledge, listening comprehension, and vocabulary—predicted differences in reading comprehension scores across all three response formats (open-ended, multiple-choice, and retell). Furthermore, an interaction between language knowledge and response format suggested that readers with low language knowledge may score less well on open-ended tests. These findings converged with prior investigations reporting language knowledge as a predominant predictor of QRI comprehension performance (Keenan et al., 2008; Keenan & Meenan, 2014), but differential contributions of this ability on low to high scores (Hua & Keenan, 2017).

In the present study, we extended this investigation by focusing more narrowly on contributions of response format and text genre in reading comprehension scores and these test component interactions with readers' language knowledge. As such, we shifted our theoretical lens. Whereas the RAND (2002) framework identifies variance in reading comprehension scores as evidence of an interaction between the reader, the text, and the activity, classical test theory divides these same sources of variance into two categories: (a) true sources of score variance related to individual differences among persons (i.e., characteristics and abilities of the reader) and (b) sources of measurement error (i.e., components of the test or measurement process). Categorizing sources of variance in this way permits the quantification of a score's dependability—the extent to which it can be relied upon for decision-making because it reflects a given reader's true score rather than measurement error—as the ratio of true score variance to observed score variance. Applying this classical theory, language knowledge is a reader characteristic that contributes to true differences in test scores, supported by empirical evidence validating its importance in reading comprehension (e.g., Nation, 2019; Vellutino et al., 2007). As such, response format and text genre exemplify test components that contribute to measurement error in reading comprehension scores, thereby inhibiting identification of readers' true scores. In this regard, interactions between sources of true and error variance in test scores may undermine the reliability and validity of current reading comprehension tests, and compromise the degree to which reading comprehension scores may be relied upon to identify readers with S-RCD.

## **Application of generalizability theory to reading comprehension assessment**

Understanding the contributions of individual persons and measurement facets to test score variance is necessary to minimize measurement error and capture scores that truly represent individual differences in student ability. In the present study, we employed Generalizability theory (G theory; Cronbach, Gleser, Nanda, & Rajaratnam, 1972), which extends classical test theory, to examine sources of error in reading comprehension test scores. When applied to student reading comprehension data across multiple scores from an assessment, G theory can be used to parse total score variance into that attributable to the reader (i.e., true individual variance), versus that attributable to aspects of the measurement procedure (i.e., error variance), and interactions among those factors (i.e., additional error variance). Generalizability (G) studies use variance estimates from analysis of variance (ANOVA) to identify contributions of each source of variance, their interactions, and random error.

Variance estimates yielded in *G* studies can then be applied to decision (*D*) studies to explore measurement optimization. Specifically, *D* studies are used to identify the level of score dependability achieved with a given set of measurement conditions and the optimal number of scores necessary to attain higher levels of dependability (Shavelson, Webb, & Rowley, 1989). *D* studies produce two coefficients that index the reliability to support score-based inferences, depending on the nature of inference being made (Brennan, 1992). When scores are used to make relative inferences (e.g., rank order comparisons among individuals), the generalizability (*G*) coefficient should be evaluated. When scores are used to make absolute inferences about individuals, such as whether a child meets specified S-RCD criteria, the phi ( $\Phi$ ) coefficient is the more appropriate index. For both indices, values greater than 0.8 are considered sufficiently reliable for most inferences, though reliability greater than 0.9 is preferred for enhanced confidence in important decisions (Nunnally, Bernstein, & Berge, 1967; Webb, Shavelson, & Haertel, 2006). Findings from a given *D* study may suggest that when scores from a single measurement are not dependable, an average of scores from multiple test administrations are needed, as this reduces error by increasing the “signal to noise” ratio, thereby increasing score dependability. In the present study, we projected dependability for averaged scores across an increasing number of measurements for readers with emerging, basic, and proficient language knowledge on three response formats (open-ended questions, multiple-choice, retell), when assuming equivalence across text genre variance.

## Research questions

The present study attempted to address these issues as they apply to measurement of reading comprehension in fourth-grade students with and without disabilities. Specifically, we applied *G* theory to examine contributions of the reader (person) and test component (measurement error) variance across emerging to proficient reader profile groups, defined by language knowledge. Using data from Collins et al., (2020), we divided students into three groups according to their language knowledge: (a) < 90 emerging, (b) 90–100 basic, and (c) > 100 proficient. We then analyzed each group’s scores, partitioning variance attributable to reader, text genre, response format, and their respective interactions. Finally, we conducted decision studies with scores from three response formats to examine the relative dependability of scores across language knowledge groups. In sum, we conducted *G* and *D* studies to address the following research questions:

1. For each language knowledge group, what is the proportion of true variance attributable to differences among readers and of error variance attributable to test components (response format and text genre), as well as interactions between these facets?
2. For each language knowledge group, what is the projected dependability of scores from each response format (open-ended, multiple-choice, retell) across an average of increasing numbers of scores from 2 to 17?

In examining proportions of variance across persons and measurement facets, we would expect that for any given reliable test, the vast majority of variance would be attributable to individual reader differences, with substantially less attributable to response format and text genre. When this is the case, we have greater confidence regarding S-RCD

identification and diagnosis based on reading comprehension test scores. By contrast, large proportions of error variance attributable to test components, such as response format and text genre, may compromise our ability to obtain representative estimates of students' individual abilities across variations of these measurement procedures, thus challenging the dependability of a single score.

## Method

### Participants

We recruited participants from six fourth-grade classrooms in an urban public elementary school in the Southeastern United States, regardless of disability status, as part of a larger study (see Collins et al., 2020). The sample included a full continuum of 78 fourth-grade students, including 36 males. Of 82 initial student volunteers, three were dismissed from the study due to severe communication challenges that impeded understanding of test directions, and consequently test completion. In addition, one student did not complete a narrative multiple-choice test item and was therefore removed from analyses. No other attrition affected the final sample. Sample participants ranged in age from 9.13 to 10.74 years ( $M=9.72$ ;  $SD=0.34$ ), with the following demographics: 39% Black, 27% Hispanic/Latino, 22% White, 4% Asian, and 9% Other. Approximately 9% of the sample were ELLs, and 14% were students with high-incidence disabilities. Sample demographics generally reflected those in the schoolwide population.

### Measures

Trained graduate research assistants in education administered measures individually to students. The full battery within the larger study included various cognitive and reading measures, with all measures administered during two 60-min testing sessions (see Collins et al., 2020). In the present study, only the reading comprehension and language knowledge measures are described and reported to align with the research questions and purpose.

### Reading comprehension

Reading comprehension was assessed using six grade-level texts from level 4 of the *Qualitative Reading Inventory-Fifth Edition* (QRI-5; Leslie & Caldwell, 2011). Narrative passages included the following: (a) "Johnny Appleseed," (b) "Amelia Earhart," and (c) "Tomie dePaola." Narrativity indices for these passages ranged from 75 to 77% (Coh-Metrix; Graesser, McNamara, & Kulikowich, 2011). Expository passages included the following: (a) "Early Railroads," (b) "The Busy Beaver," and (c) "Plant Structures for Survival" (narrativity indices: 26–35%). Students read each passage aloud and then completed a short comprehension test in one of three randomly assigned response formats: open-ended, multiple-choice, or retell. Figure 1 provides examples of items from each response format. Examiners prompted students with a word if a student paused for more than 10 s when reading. This procedure was applied sparingly ( $Mdn=0$ ). Participants did not have access to passages during the comprehension assessments.

**Table 1** Means and standard deviations for language knowledge groups across text genres and response formats

Reader groups, text genre	Response formats											
	Open-ended				Multiple-choice				Retell			
	<i>M</i>	<i>SD</i>	Min	Max	<i>M</i>	<i>SD</i>	Min	Max	<i>M</i>	<i>SD</i>	Min	Max
Emerging LK ( <i>n</i> = 27)												
Narrative	3.04	1.89	0	7	5.78	1.31	3	8	10.26	4.82	1	20
Expository	2.78	2.41	0	8	4.93	1.84	1	8	7.11	3.50	0	18
Total	5.81	3.58	1	14	10.70	2.58	5	15	17.37	6.75	6	35
Basic LK ( <i>n</i> =27 <sup>a</sup> )												
Narrative	4.37	1.24	2	6	6.37	0.88	5	8	13.70	5.36	0	28
Expository	3.41	1.89	0	8	5.63	1.78	1	8	10.11	4.04	1	18
Total	7.78	2.19	4	14	12.00	1.94	8	16	23.81	7.59	7	40
Proficient LK ( <i>n</i> = 24)												
Narrative	5.46	1.41	2	8	6.71	1.30	3	8	17.58	6.20	5	29
Expository	4.79	1.77	2	8	6.63	1.28	2	8	13.33	6.18	5	27
Total	10.25	2.52	4	14	13.33	1.71	9	16	30.92	9.65	14	48

LK, language knowledge; *Emerging* defined as standard scores composite < 90; *Basic* defined as  $90 \leq$  standard scores composite  $\leq 100$ ; *Proficient* defined as standard scores composite > 100

<sup>a</sup> One student was dropped from the basic group due to missing data

**Open-ended** Each of the six QRI-5 passages included eight open-ended items (see Fig. 1 for examples of items). Four of the eight items were explicit questions that asked readers to recall literal information stated verbatim in the QRI-5 passage. The other four items were implicit questions that required students to use higher-order thinking skills to make inferences or draw conclusions from the passage. Examiners read open-ended items aloud as participants followed along to minimize potential effects of word reading difficulties. Then, participants responded orally to questions as examiners transcribed responses. Open-ended items were scored as correct or incorrect, and the QRI-5 manual reports interrater agreement as 0.98 (Leslie & Caldwell, 2011). Interrater agreement (agreements / (agreements + disagreements)) for this sample was 0.93.

**Multiple-choice** We created a second reading comprehension test with multiple-choice items to examine relations between reader abilities and response formats in the larger study. We used the eight QRI-5 open-ended items for each passage as multiple-choice item stems (see Fig. 1 for examples of how open-ended item stems were included in multiple-choice items). Using the same item stems held question types constant across the two response formats. It also ensured the content being evaluated was comparable to that tested in open-ended and retell measures. Multiple-choice answer options followed established print- and web-based guidelines for creating tests of this response format (Brame, 2013; Center for Teaching Excellence, 2013; Haladyna, 1999). These guidelines included creating alternative choices that served as functional distractors and plausible options. Every multiple-choice item had four answer choices (one correct, three distractors) presented vertically from a to d. Correct answers varied to ensure items did not follow specific patterns (e.g., c was not always the correct answer). Each item's answer choices were mutually exclusive (no overlapping content), used similar grammar stems, followed parallel forms, and were uniform in length. These parameters mitigated the potential for language and other variations (e.g., answer choices substantially

Item Type	Open Ended	Multiple Choice
<b>Narrative Passage: Amelia Earhart</b>		
<b>Explicit</b>	What was Amelia Earhart trying to do when her plane disappeared?  <i>Answer:</i> Fly around the world	What was Amelia Earhart trying to do when her plane disappeared?  a. <b>She was trying to fly around the world</b> b. She was trying to fly to Ireland and back. c. She was trying to fly passengers on a trip. d. She was trying to fly pilots home from war.
<b>Implicit</b>	What was Amelia Earhart's main goal?  <i>Answer:</i> To Fly; OR to do things that were challenging	What was Amelia Earhart's main goal?  a. To be a doctor. b. To sail the Atlantic. c. To care for pilots. d. <b>To fly a plane.</b>
<b>Expository Passage: Plant Structures for Survival</b>		
<b>Explicit</b>	What is one example of how a plant adapts to get more sunlight?  <i>Answer:</i> A vine grows up a tall plant to get more sunlight; OR lily pads grow very large to capture sunlight.	What is one example of how a plant adapts to get more sunlight?  a. A flower closes its petals to capture sunlight. b. <b>A lily pad grows very large to capture sunlight.</b> c. A tree produces more fruit to capture sunlight. d. A fern stays green all year to capture sunlight.
<b>Implicit</b>	How are a cactus and a pine tree alike?  <i>Answer:</i> A waxy substance that prevents loss of water covers them both.	How are a cactus and a pine tree alike?  a. They both have fruits that provide food for people who live nearby. b. They both have large leaves and flowers to scare away animals. c. They both have thick bark to protect their trunks from the sun. d. <b>They both have a waxy substance covering them to prevent water loss.</b>

**Fig. 1** Examples of open-ended and multiple-choice items from the Qualitative Reading Inventory-Fifth Edition. Open-ended items were scored as correct as 1 or incorrect as 0. Bolded items for multiple-choice indicate correct answers scored as correct as 1 or incorrect as 0

longer or shorter in length) in cluing readers to correct answers. No multiple-choice items used “all of the above,” “none of the above,” or combined choices that would advantage readers with partial knowledge.

Before administering tests to participants, four graduate students in education reviewed the assessments to identify answer choices that failed to conform to print- and web-based test guidelines. The research team further revised these items. In testing sessions, examiners read aloud all item stems and answer options as students followed along. Cronbach's alpha for QRI-5 open-ended and corresponding multiple-choice items was 0.80.



**Retell** Retell prompts and scoring protocols were also drawn from the QRI-5. After reading, examiners prompted students to tell everything they remembered about the passage, starting from the beginning and continuing until the end. Raw scores reflect the number of idea units recalled from the scoring checklist, either verbatim or summarized within the oral retelling. The QRI-5 organizes narrative passage idea units into four categories (setting/background, goal, events, resolution) and expository text idea units into two categories (main ideas, details). The QRI-5 manual does not report reliability statistics for retell (Leslie & Caldwell, 2011); interrater agreement for this sample was 0.82.

## Language knowledge

Three *WJ-III Tests of Achievement* (WJ-III; Woodcock, McGrew, & Mather, 2001) subtests comprised a composite measure of language knowledge: (a) *Academic Knowledge*, (b) *Oral Comprehension*, and (c) *Picture Vocabulary*. Correlations among raw scores for these measures were moderate to strong (range: 0.53–.69;  $p < 0.001$ ). In the present study, we converted subtest raw scores to standard scores, then calculated a composite score by averaging the three standard scores. Measures represented in this composite score are described below.

**General academic knowledge** The *WJ-III Academic Knowledge* subtest (Woodcock et al., 2001) assesses the academic areas of science, social studies, and humanities across a total of 78 possible items. Basal and ceiling rules were applied. Split-half reliability for ages 9 and 10 for each academic area is 0.79, 0.79, and 0.85, respectively (McGrew, Schrank, & Woodcock, 2007).

**Listening comprehension** The *WJ-III Oral Comprehension* subtest (Woodcock et al., 2001) assesses listening comprehension using cloze tasks across 34 possible items. For each item, examiners read 1–2 sentences with one word missing. Students responded by suggesting one word to complete the sentence. Basal and ceiling rules were applied. Split-half reliability for ages 9 and 10 is 0.79 (McGrew et al., 2007).

**Vocabulary** The *WJ-III Picture Vocabulary* subtest (Woodcock et al., 2001) assesses expressive vocabulary across 44 possible items. Each item is a picture, and students provide the appropriate word to represent the picture. Basal and ceiling rules were applied. Split-half reliability for ages 9 and 10 is 0.79 (McGrew et al., 2007).

## Testing procedures

Student data reflect a single data collection period spanning 2 months. Each participant was individually administered two 60-min assessment batteries within 1 week. We administered QRI-5 tests midway through each testing session, using standardized directions and order. Battery 1 assessed narrative passage comprehension, listening comprehension, and academic knowledge; battery 2 assessed expository passage comprehension and vocabulary. To ensure standardized procedures across participants, no accommodations or modifications were provided for ELLs or students with disabilities.

## Study design

The present study applied a  $3 \times 2$  design in which response format (open-ended, multiple-choice, retell) and text genre (narrative, expository) were randomized and counterbalanced across students. Applying this design, we randomly assigned students one of the three response formats for each narrative QRI-5 passage, then replicated this procedure with expository passages. This study design resulted in two scores per student for each response format, and three scores per student across each text genre. Our application of this study design aimed to minimize weight placed on a given passage, genre, or question, thus distributing any potential influence of these testing factors more generally across response formats and text genres rather than at the item level.

## Test administration and scoring

Prior to administration, examiners received training on the QRI-5 and WJ-III standardized administration and scoring procedures and met a minimum of 90% fidelity of test implementation and scoring using a researcher-developed checklist. Examiners audio-recorded all testing sessions. A second examiner selected and reviewed a random sample of 20% of audio recordings for (a) reliability of test scores and (b) fidelity of test implementation. Average reliability and fidelity exceeded 0.93. Research assistants double-entered the data, and the first author resolved discrepancies.

## Data analysis

We used a language knowledge composite score representing the average of the standard scores across all WJ-III tests to classify students among three profile groups. The groups were defined with the following composite scores: (a) *emerging* as standard scores composite  $< 90$  ( $n = 27$ ); (b) *basic* as  $90 \leq$  standard scores composite  $\leq 100$  ( $n = 27$ ); and (c) *proficient* as standard scores composite  $> 100$  ( $n = 24$ ). These sample sizes align with Webb, Rowley, and Shavelson' (1988) recommendations that generalizability studies should feature a minimum of 20 participants with at least two measurements per person.

To estimate the true observed variance among readers and error variance in scores attributable to measurement components, we conducted three separate *G* studies across the facets of the person (i.e., reader), text genre, and response format for each reader profile group (emerging, basic, proficient). Person and measurement facets were treated as fully crossed because every level of each facet was crossed with every level of all other facets (i.e., scores were available for each response format within each text genre for each student). All facets were also treated as infinite random, meaning observations in each facet were treated as a sample from an infinite universe of possible observations and allowed to vary randomly within that universe. Treating facets as infinite random provides more conservative estimates of dependability and also ensures that such estimates can be generalized beyond the observed sample of participants (Shavelson & Webb, 1991). We then conducted subsequent *D* studies using scores from each response format for each language knowledge group to project the dependability that could be achieved by

**Table 2** Mean square and percentage of variance in reading comprehension scores by reader, text genre, and response format

Source of variance	Full sample			Emerging language knowledge			Basic language knowledge			Proficient language knowledge		
	<i>df</i>	MS	Var. (%)	<i>df</i>	MS	Var. (%)	<i>df</i>	MS	Var. (%)	<i>df</i>	MS	Var. (%)
Reader	77	26.98	7.80	26	13.01	4.3	26	12.25	1.9	23	22.65	4.2
Response format	2	266.03	50.50	2	454.22	42.8	2	932.85	58.8	2	1491.58	62.4
Text genre	1	305.31	2.50	1	81.63	3.5	1	126.23	4.1	1	100.00	1.2
R × RF	154	13.89	8.40	52	9.76	9.6	52	10.39	5.1	46	14.25	1.9
R × TG	77	6.41	0	26	4.82	0	26	6.10	0	23	8.98	0
TG × RF	2	120.04	4.40	2	31.43	5.1	2	33.96	3.5	2	61.08	4.2
R × TG × RF	154	8.46	26.30	52	6.27	34.6	52	7.49	26.6	46	12.48	26.2
<i>G</i> and phi Coef.			Coef.			Coef.			Coef.			Coef.
Relative <i>G</i> Coef.			0.52			0.33			0.24			0.45
Absolute phi Coef.			0.23			0.14			0.06			0.13

*df*, degrees of freedom; *MSE*, mean square; *Var.*, variance; *R*, reader; *TG*, text genre; *RF*, response format; *Coef.*, coefficients

averaging scores from additional administrations. The number of estimated observations increased in increments of three from an average of two scores up to 17. All analyses were conducted in EduG version 6.1 (Swiss Society for Research in Education Working Group, 2012).

## Results

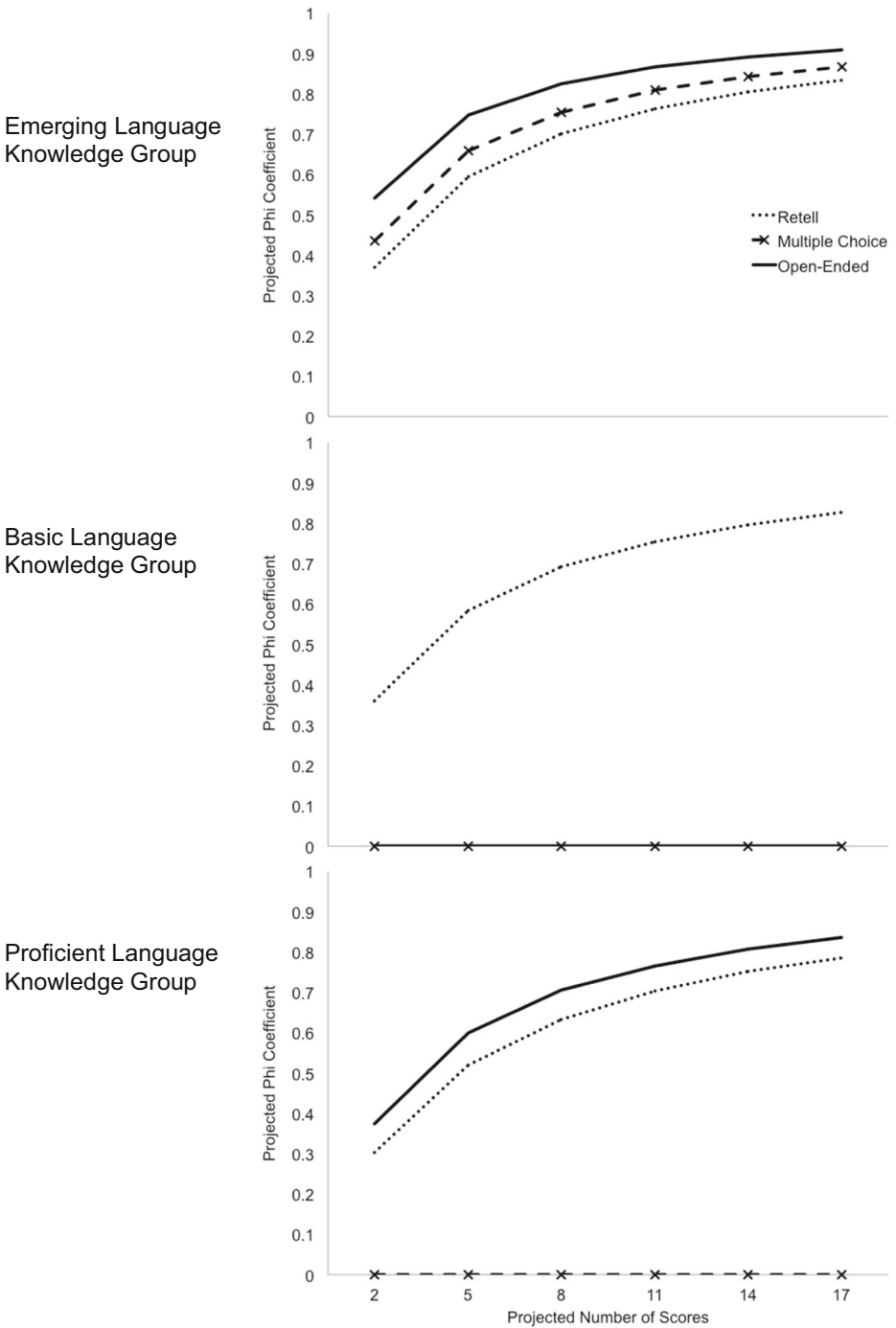
Table 1 reports descriptive statistics for QRI-5 measures across response format and text genre by language knowledge group ( $N = 78$ ). Descriptive statistics for standard score composites for each language knowledge group were as follows: (a) *emerging* ( $n = 27$ ):  $M = 83.01$ ,  $SD = 5.49$ , range: 65.33–89.33; (b) *basic* ( $n = 27$ ):  $M = 95.07$ ,  $SD = 2.80$ , range: 90.6–99.66; and (c) *proficient* ( $n = 24$ ):  $M = 106.65$ ,  $SD = 5.26$ , range: 100.33–120.33.

### Research question 1: variance attributable to reader, response format, and text genre

Table 2 indicates the results of variance partitioning across the three facets and their interactions, as well as relative  $G$  and phi coefficients. Across language knowledge groups, true person variance (i.e., variance attributable to differences among readers) accounted for only 1.9–4.3% of the observed variance in scores, indicating measurement error contributed to the vast majority of variance. Across all groups, text genre accounted for less than 5% (range: 1.2–4.1%) of all score variance, also indicating a minimal influence of this facet on scores. The interaction of text genre and reader accounted for no variance indicating that the relative ranking of participants did not vary across the two types of passages for all groups. Response format, by contrast, contributed the greatest proportion of variance in reading comprehension scores for all language knowledge groups (range: 42.8–62.4%). The interaction between response format and reader contributed 1.9–9.6% of observed variance for each group, indicating that the rank ordering of participants somewhat varied across different response formats, differentially within groups. The interaction between response format and text genre across groups contributed minimal variance (3.5–5.1%), indicating the relative ranking of response formats did not vary by genre of the passage. The remaining observed score variance (range: 26.2–34.6%) was attributable to the interaction of all three facets or other measurement characteristics not examined in this G study (e.g., examiner, rater).

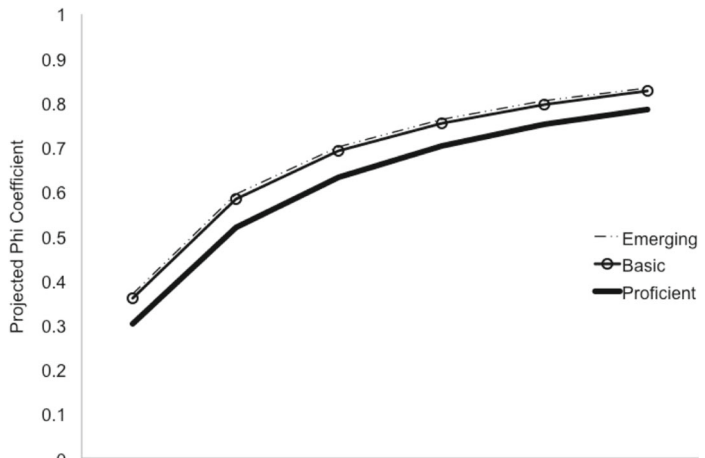
### Research question 2: dependability of scores by response format and language knowledge groups

Figures 2 and 3 reflect the projected dependability (i.e., phi coefficients) that could be achieved by averaging scores from additional administrations within each response format for each language knowledge group. Figure 2 summarizes D study results by language knowledge group to permit comparison of score dependability within each group across response formats, while Fig. 3 depicts results by response format to permit comparison of projected score dependability of a given response format across language knowledge groups. The achieved dependability of an average of two scores from open-ended responses was 0.54 for readers with emerging language knowledge, 0 for readers with basic language knowledge, and 0.37 for readers with demonstrated proficiency. The achieved dependability of an average of two scores from multiple-choice responses was 0.43 for readers with emerging language

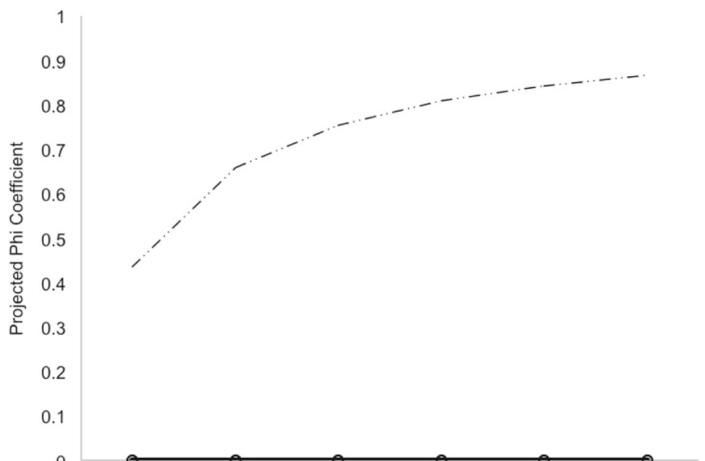


**Fig. 2** Projected dependability (phi) coefficients for language knowledge groups by response format. Graphs depict projected dependability (phi) coefficients obtained with an average of an increasing number of scores from within response format, for readers with emerging, basic, and proficient language knowledge

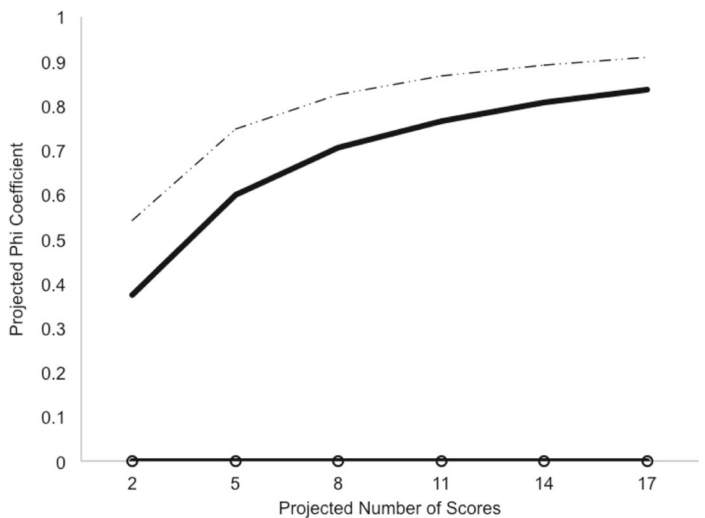
Retell



Multiple-Choice



Open-Ended



**Fig. 3** Projected dependability ( $\phi$ ) coefficients for response format by language knowledge group. Graphs depict projected dependability ( $\phi$ ) coefficients obtained with an average of an increasing number of scores within response format for readers with emerging, basic, and proficient language knowledge

knowledge, 0 for readers with basic language knowledge, and 0 for readers with demonstrated proficiency. The achieved dependability of an average of two scores from retell responses was 0.37 for readers with emerging language knowledge, 0.36 for readers with basic language knowledge, and 0.35 for readers with demonstrated proficiency. Decision studies indicated that, for readers with emerging language knowledge, projected dependability was similar for all three response formats, but highest for open-ended scores relative to multiple-choice and retell. These results suggest that to obtain a sufficiently dependable score for readers with emerging language knowledge, an average of 7 open-ended, 11 multiple-choice, or 14 retell scores may be needed. For readers with basic language knowledge, projected dependability was highest for retell, and 0 for multiple-choice and open-ended response formats. To obtain a sufficiently dependable score for this group, an average of 14 retell scores may be needed. Projections suggest that dependable scores could not be obtained by averaging scores from multiple-choice or open-ended response formats alone. For readers with demonstrated language knowledge proficiency, projected dependability was highest for open-ended scores relative to retell, and consistently zero for multiple-choice. Even so, an average of 14 open-ended or 17 retell scores would be needed to obtain sufficiently stable scores for readers with proficient language knowledge.

## Discussion

In the present study, we applied generalizability theory to explore the influence of language knowledge on the dependability of scores from three reading comprehension tests adapted from the QRI-5 that employed different response formats. Specifically, we examined fourth-graders' QRI-5 test scores to determine variance attributable to individual differences among persons (i.e., readers) and two potential sources of measurement error (response format and text genre). We investigated the proportions of variance from each of these sources separately for readers with emerging, basic, and proficient language knowledge abilities. Finally, because results indicated response format was a key source of error variance, we conducted *D* studies for each profile group to estimate the dependability of student test scores within three different response formats (open-ended, multiple-choice, retell). *D* study results estimated the number of scores necessary to obtain a dependable score across an average of an increasing number of scores from 2 to 17. Our findings have important implications for using the QRI-5 to identify students with S-RCD, and broader implications about the varied impact of response format on reading comprehension assessment in students with differing language knowledge proficiency. Therefore, we elaborate on findings that may have the strongest implications for research, practice, and future test development.

First, our *G* study findings suggest the QRI-5 tests may include a substantial amount of measurement error attributable to the administration of different response formats. Furthermore, error variance attributable to response format increased from emerging to proficient language knowledge groups, suggesting the contribution of certain test components do not necessarily generalize to students across a full continuum of abilities. These findings converge with prior research reporting differences in language knowledge contributions on QRI

response formats among varied ability levels (Collins et al., 2020; Hua & Keenan, 2017; Keenan & Meenan, 2014). Such findings, however, complicate clinical recommendations for using reading comprehension measures, such as the QRI, which depend more on language knowledge than decoding (Kamhi, 2009). In part, using various response formats may differentially introduce noise into this test, thus constraining assessment of a reader's true scores because the level of constraint depends on their language knowledge. Consequently, unintended noise may impede accurate identification of students with S-RCD, as a student's scores may vary widely by response format. As such, potential for faulty decision-making based on unreliable test scores may be heightened for this assessment, particularly when identifying students with S-RCD. Therefore, although the QRI-5 was designed for practical use to identify at-risk readers—including those with S-RCD—the extent to which this measure is reliable in distinguishing among true individual reader differences may be limited.

Our findings also suggest that although substantial variance was attributable to response format across all language knowledge groups, response format may introduce less measurement error for readers with emerging language knowledge and scores may be more stable for this profile group. Students with emerging language knowledge may generally perform less well on the QRI, regardless of response format, as their reading comprehension scores largely represent their limited language knowledge proficiency. In this regard, language knowledge deficits may increase potential for floor effects for this group, as evidenced in the minimum QRI scores measured as substantially lower, relative to the other language knowledge groups. One possible way to mitigate potential floor effects would be to increase the maximum scale on these response formats by using a composite of scores across multiple passages. Despite results suggesting *potential* for floor effects, scores indicate readers with emerging language knowledge are indeed on the QRI scale, but consistently on the lower end. When students with emerging language knowledge score consistently near 0, the stability of their scores affords more reliable detection of between- and within-individual differences. This enhanced measurement stability may allow more dependable decision-making for readers with emerging language knowledge when using the QRI-5 as a screener or in clinical diagnoses. Still, the large portion of error variance attributable to response format alone should warrant caution when using the QRI to identify students with S-RCD, as diagnoses or decisions may differ within students across response formats.

Our *D* studies extended these findings and indicated many administrations of the QRI-5 tests were required to reach a desirable level of reliability of 0.80. For readers with emerging language knowledge, eight administrations of the open-ended test may produce dependable results, and consequently more accurate decisions from these scores. Yet, for readers in the proficient language knowledge group, as many as 14 administrations of the open-ended tests were necessary to achieve the same desired level of reliability. These findings mostly held across multiple-choice and retell, though negative correlations precluded improvement in the G Coefficients for some groups (further discussed later in this section). We believe this lack of stability across tests and language knowledge groups underscores a serious concern with potential applications of this test because our results suggest these three response formats are not sufficiently reliable for decision-making when using only one score or an average of two scores, regardless of language knowledge group. Moreover, averaging several scores is needed to reach an acceptable threshold of dependability that would permit valid conclusions about a student's reading comprehension, and our results further suggest the number of scores needed varies by student language profile. However, it is not uncommon in research and practice to use one score to estimate reading



comprehension, including when making decisions regarding S-RCD. This misalignment between dependability and common assessment practice suggests further investigations of other tests are needed to determine if the large variance attributable to measurement error, particularly of different response formats, generalizes to other reading comprehension assessments. This misalignment also has clinical implications for S-RCD, and we offer recommendations for educators later in this section.

Among the three response formats, retell scores had the most stable dependability across all language knowledge groups (although far below the 0.80 threshold that permits reliable diagnostic and instructional decisions). One plausible explanation is students must engage in “knowledge telling” during a retell (Scardamalia & Bereiter, 1987), using their working memory (Spencer et al., 2019) to recall and organize passage content with their existing prior knowledge and vocabulary (Collins et al., 2020; Reed & Vaughn, 2012). In this regard, retell is perhaps a lower-order task, relying more heavily on literal recall over higher-order inferences. A reader’s prior knowledge may partially contribute to literal recall as they integrate existing information in their schema with new knowledge gained from the passage (see Priebe, Keenan, & Miller, 2012). We cannot be certain of its contribution, however, given that our research questions and study methods (i.e., counterbalancing passages—and therefore topics—with response format) precluded any conclusive evidence for how students’ prior knowledge of the passage topics influenced performance on retell. Future studies may explore the dependability of response formats across language knowledge profile groups by controlling for prior knowledge through use of narrative passages describing a fictional world (e.g., Cain, Oakhill, Barnes, & Bryant, 2001) or expository passages on topics unfamiliar to all participants.

In contrast to retell, we identified negative correlations and lack of stability for multiple-choice in the basic and proficient language knowledge groups and for open-ended scores in the basic group. The cut-points we used to classify these groups may explain these findings. In particular, the basic group comprised readers with language knowledge standard scores from 90 to 100. Previous studies indicate cut-points of low achievement as a proxy for learning disability may be relatively unstable predictors across time (Francis et al., 2005). Thus, the basic group may include “bubble” students (Booher-Jennings, 2005), who inconsistently perform above or below cut-points on occasion. Our findings suggest open-ended and multiple-choice tests administered and adapted from the QRI-5 may result in vastly different results across multiple administrations, and these tests may limit accurate identification of reading comprehension difficulties for this profile group. Therefore, diagnosis and identification of S-RCD may waiver from one administration to the next. This instability has serious implications for allocation of instructional supports and interventions, with potential revolving door effects if broader profiles are not considered. Our findings support future studies focused on readers with basic language knowledge and their performance on reading comprehension tests of different response formats to disentangle other factors contributing to observed instability.

### **Clinical implications**

Collectively, our findings suggest an average of several scores across response formats is needed to reach acceptable thresholds of dependability and permit reliable conclusions about students’ reading comprehension from the QRI-5. Moreover, the number of scores needed vary by student language knowledge group, and therefore, some response formats may be

more dependable for certain students. One encouraging finding, however, is that scores for students with emerging language knowledge were relatively comparable in dependability across all response formats. Practitioners can have some confidence in making decisions on reading comprehension for individuals in this group, as their similar dependability across response formats may reflect consistent lower performance. Conversely, the dependability of scores across response formats was highly variable for students with basic language knowledge, and this finding may have the most important implications for clinical diagnoses because it may be more difficult to identify individual students on the cusp of developing S-RCD.

One clinical implication of these findings lies in a possible solution for reducing measurement error. Practitioners who average multiple scores, specifically across key sources of error (i.e., response format), are likely to increase the signal-to-noise ratio. Averaging scores across multiple administrations increases the dependability of students' individual scores relative to a single score from a single measurement, and averaging across response formats will reduce error attributable to this facet and the number of scores needed to obtain a dependable average. This approach will ensure more reliable diagnostic and instructional decisions about students with S-RCD.

Unfortunately, it is impractical for practitioners to administer eight or more iterations of a test, each requiring up to 30 min, taking into account lost instructional time and student testing fatigue. Even in research contexts, multiple administrations may not be feasible and often require considerable resources. Alternatively, some researchers recommend that educators use tests that combine response formats to reduce measurement error introduced by this test component (García & Cain, 2014), particularly when diagnosing S-RCD (Clemens et al., 2020; Keenan et al., 2014). Few standardized assessments, however, incorporate more than one response format. Thus, our findings raise important clinically oriented questions about what changes are necessary for future test refinement and development.

Given our findings, practitioners should recognize that no single response format is optimal, and score reliability reflects the language knowledge of the reader. Importantly, we do not believe these stated issues in the assessment of reading comprehension reflect "misinterpretation" of test scores because this implies that the source of measurement error results from the assessment examiner or rater rather than the measure itself. Instead, we argue that the individual scores themselves are unreliably representative of the student's overall reading comprehension ability. To increase the extent to which practitioners can depend on reading comprehension assessments for S-RCD identification, future test development should explore how to increase variance attributable to true individual differences among persons (i.e., readers' characteristics and abilities) and decrease noise from sources of measurement error (i.e., response format) in reading comprehension.

### **Limitations and future directions**

Our findings have important implications for using tests to identify and diagnose S-RCD. Our conclusions, however, should be considered in the context of certain limitations. First and foremost, our primary goal was to manipulate reading comprehension test components by examining three response formats and two text genres and control for other potential confounds (e.g., topic, text difficulty). We thus constrained our analyses to one reading comprehension test (i.e., QRI-5). These results, however, lay the groundwork for planning an optimized study to directly evaluate the construct validity of scores from different tests (representative of different response formats), collected under optimal conditions and in

sufficient numbers to produce dependable scores. Therefore, future studies should examine if findings hold for other assessments and reader profile groups to increase our confidence in the preceding recommendations. Second, because the multiple-choice test was researcher-designed, it was not subjected to rigorous psychometric evaluation emblematic of standardized measures. Nonetheless, our procedures tend to reflect typical practice of educators, who may create their own multiple-choice tests often with very few guidelines. At the same time, the researcher-designed test results in restricted generalizability that fails to transfer to other validated reading comprehension tests, including those using multiple-choice items. This finding warrants future investigation with norm-referenced measures. Third, our analysis narrowly focused on the test components of response format and text genre. Our results indicated 26% to as much as 35% of error variance remained unexplained in our models. It is likely other test components such as text cohesion, sentence length, and question types identified in prior studies as affecting differences in test scores (e.g., Kulesz et al., 2016; McNamara, Ozuru, & Floyd, 2011) may contribute to this unexplained variance. Future studies should further examine how these test components interact with reader abilities such as language knowledge (including prior topic knowledge).

## Conclusion

The present study identifies differences in variance attributable to response format and differential contributions with respect to how this test component interacts with varied language knowledge abilities. Findings extend recent investigations on reading comprehension tests (e.g., Hua & Keenan, 2017) by suggesting test components, such as response format, introduce additional variance in test scores across different reader groups. Our findings reduce confidence in using one test—and further, one score—to identify S-RCD because too many factors may be at play in any given testing situation. That said, given the limitations of the present study, future research is warranted to precisely determine how our findings across response formats and interactions among language knowledge groups generalize to broader theoretical frameworks used in reading comprehension assessment. Further examination would inform how the reader and test component interactions should be accounted for in future test development.

A new challenge brought forth by our findings is determining how these findings translate to clinical diagnoses of S-RCD. It is reasonable to suggest that educators consider underlying reader abilities and test components during the identification of S-RCD, as they likely have other measures that may be administered during comprehensive evaluations (see recommendations by Collins et al., 2020). Yet, it is less straightforward to advise educators to analyze test components to identify measurement error across different reader groups, as these psychometric complexities likely extend beyond their professional expertise. Instead, researchers and test developers need to focus future work on ways to reduce measurement error and noise in assessments, and aligning assessments more closely with theoretical models that take test components such as response format into consideration (e.g., Francis et al., 2018). To heed this recommendation, however, test developers would need to ensure reading comprehension assessments allow quick administrations of multiple response formats, while also upholding screening and diagnostic requirements. Commitment to advancing present

testing methods will strengthen modes for diagnosing and identifying S-RCD and in turn promote more strategic allocation of intervention resources.

## Compliance with ethical standards

**Conflict of interest** The authors declare that they have no conflict of interest.

## References

- Best, R. M., Floyd, R. G., & McNamara, D. S. (2008). Differential competencies contributing to children's comprehension of narrative and expository texts. *Reading Psychology, 29*, 137–164.
- Booher-Jennings, J. (2005). Below the bubble: “Educational triage” and the Texas accountability system. *American Educational Research Journal, 42*(2), 231–268.
- Brame, C. (2013). Writing good multiple-choice test questions. Retrieved from <https://bit.ly/37BuBeO>.
- Brennan, R. L. (1992). Generalizability theory. *Educational Measurement: Issues and Practice, 11*(4), 27–34.
- 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.
- Catts, H. W., Compton, D., Tomblin, J. B., & Bridges, M. S. (2012). Prevalence and nature of late-emerging poor readers. *Journal of Educational Psychology, 104*(1), 166–181.
- Catts, H. W., Hogan, T. P., & Adlof, S. M. (2005). Developmental changes in reading and reading disabilities. In H. W. Catts & A. G. Kamhi (Eds.), *The connections between language and reading disabilities* (pp. 25–40). New Jersey: Lawrence Erlbaum.
- Center for Teaching Excellence (2013). Virginia Commonwealth University. *Writing multiple-choice questions*. Retrieved from <https://bit.ly/38NnSid>.
- Clemens, N. H., Hsiao, Y.-Y., Lee, K., Martinez-Lincoln, A., Moore, C., Toste, J., & Simmons, L. (2020). The differential importance of component skills on reading comprehension test performance among struggling adolescent readers. *Journal of Learning Disabilities* **Advanced online publication**.
- Colenbrander, D., Nickels, L., & Kohnen, S. (2017). Similar but different: Differences in comprehension diagnosis on the Neale Analysis of Reading Ability and the York Assessment of Reading for Comprehension. *Journal of Research in Reading, 40*, 403–419.
- Collins, A. A., Compton, D. L., Lindström, E. R., & Gilbert, J. K. (2020). Performance variations across reading comprehension assessments: Examining the unique contributions of text, activity, and reader. *Reading and Writing: An Interdisciplinary Journal, 33*(3), 605–634. <https://doi.org/10.1007/s11145-019-09972-5>.
- Compton, D. L., Fuchs, L. S., Fuchs, D., Lambert, W., & Hamlett, C. (2012). The cognitive and academic profiles of reading and mathematics learning disabilities. *Journal of Learning Disabilities, 45*(1), 79–95.
- Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. Hoboken: Wiley.
- Cutting, L. E., & Scarborough, H. S. (2006). Prediction of reading comprehension: Relative contributions of word recognition, language proficiency, and other cognitive skills can depend on how comprehension is measured. *Scientific Studies of Reading, 10*, 277–299.
- Eason, S. H., Goldberg, L. F., Young, K. M., Geist, M. C., & Cutting, L. E. (2012). Reader-text interactions: How differential text and question types influence cognitive skills needed for reading comprehension. *Journal of Educational Psychology, 104*, 515–528.
- Francis, D. J., Fletcher, J. M., Catts, H. W., & Tomblin, J. B. (2005). Dimensions affecting the assessment of reading comprehension. In S. G. Paris & S. A. Stahl (Eds.), *Children's reading comprehension and assessment* (pp. 369–394). New Jersey: Lawrence Erlbaum.
- Francis, D. J., Fletcher, J. M., Stuebing, K. K., Lyon, G. R., Shaywitz, B. A., & Shaywitz, S. E. (2005). Psychometric approaches to the identification of LD IQ and achievement scores are not sufficient. *Journal of Learning Disabilities, 38*(2), 98–108.
- Francis, D. J., Kulesz, P. A., & Benoit, J. S. (2018). Extending the simple view of reading to account for variation within readers and across texts: The complete view of reading (CVRI). *Remedial and Special Education: RASE, 39*(5), 274–288.
- García, J. R., & Cain, K. (2014). Decoding and reading comprehension: A meta-analysis to identify which reader and assessment characteristics influence the strength of the relationship in English. Review of *Educational Research, 1*, 1–38. <https://doi.org/10.3102/0034654313499616>.
- Graesser, A. C., McNamara, D. S., & Kulikowich, J. (2011). Coh-Metrix: Providing multilevel analyses of text characteristics. *Educational Researcher, 40*, 223–234 <https://doi.org/10/cwtd84>.
- Haladyna, T. M. (1999). *Developing and validating multiple-choice test items* (2nd ed.). New Jersey: Lawrence Erlbaum.

- Hogan, T. P., Adlof, S. M., & Alonzo, C. N. (2014). On the importance of listening comprehension. *International Journal of Speech-Language Pathology*, *16*(3), 199–207.
- Hua, A. N., & Keenan, J. M. (2017). Interpreting reading comprehension test results: Quantile regression shows that explanatory factors can vary with performance level. *Scientific Studies of Reading*, *21*(3), 225–238.
- Kamhi, A. G. (2009). Solving the reading crisis—Take 2: The case for differentiated assessment. *Language, Speech, and Hearing Services in Schools*, *40*, 212–215.
- Keenan, J. M. (2014). Assessment of reading comprehension. In C. A. Stone, E. R. Silliman, B. J. Ehren, & G. P. Wallach (Eds.), *Handbook of Language and Literacy: Development and Disorders* (2nd ed., pp. 469–484). New York: Guilford.
- Keenan, J. M., Betjemann, R. S., & Olson, R. K. (2008). Reading comprehension tests vary in the skills they assess: Differential dependence on decoding and oral comprehension. *Scientific Studies of Reading*, *12*, 281–300. <https://doi.org/10.1080/10888430802132279>.
- Keenan, J. M., Hua, A. N., Meenan, C. E., Pennington, B. F., Willcutt, E., & Olson, R. K. (2014). Issues in identifying poor comprehenders. *L'Annee Psychologique*, *114*(4), 753–777.
- Keenan, J. M., & Meenan, C. E. (2014). Test differences in diagnosing reading comprehension deficits. *Journal of Learning Disabilities*, *47*, 125–135.
- Kintsch, W., & Rawson, K. A. (2007). Comprehension. In M. J. Snowling & C. Hulme (Eds.), *The science of reading: A handbook* (pp. 209–226). Hoboken: Blackwell.
- Kulesz, P. A., Francis, D. J., Barnes, M. A., & Fletcher, J. M. (2016). The influence of properties of the test and their interactions with reader characteristics on reading comprehension: An explanatory item response study. *Journal of Educational Psychology*, *108*(8), 1078–1097.
- Leslie, L., & Caldwell, J. (2011). *Qualitative Reading inventory* (5th ed.). London: Pearson.
- McGrew, K. S., Schrank, F. A., & Woodcock, R. W. (2007). *Woodcock-Johnson III normative update: Technical manual*. Rolling Meadows: Riverside.
- McNamara, D. S., Ozuru, Y., & Floyd, R. G. (2011). Comprehension challenges in the fourth grade: The roles of text cohesion, text genre, and readers' prior knowledge. *International Electronic Journal of Elementary Education*, *4*(1), 229–257.
- Miller, A. C., Davis, N., Gilbert, J. K., Cho, S.-J., Toste, J. R., Street, J., & Cutting, L. E. (2014). Novel approaches to examine passage, student, and question effects on reading comprehension. *Learning Disabilities Research & Practice*, *29*(1), 25–35.
- Nation, K. (2019). Children's reading difficulties, language, and reflections on the simple view of reading. *Australian Journal of Learning Difficulties*, *24*(1), 47–73.
- Nation, K., & Snowling, M. (1997). Assessing reading difficulties: The validity and utility of current measures of reading skill. *British Journal of Educational Psychology*, *67*, 359–370.
- Nunnally, J. C., Bernstein, I. H., & Berge, J. M. T. (1967). *Psychometric theory* (Vol. 226). New York: McGraw-Hill.
- Perfetti, C., & Stafura, J. (2014). Word knowledge in a theory of reading comprehension. *Scientific Studies of Reading*, *18*(1), 22–37.
- Pfeiffer, S. I., Reddy, L. A., Kletzel, J. E., Schmelzer, E. R., & Boyer, L. M. (2000). The practitioner's view of IQ testing and profile analysis. *School Psychology Quarterly*, *15*(4), 376–385.
- Priebe, S. J., Keenan, J. M., & Miller, A. C. (2012). How prior knowledge affects word identification and comprehension. *Reading and Writing*, *25*, 131–149.
- RAND Reading Study Group. (2002). Reading for understanding: Toward an R&D program in reading comprehension. RAND.
- Reardon, S. F., Kalogrides, D., Fahle, E. M., Podolsky, A., & Zárate, R. C. (2018). The relationship between test item format and gender achievement gaps on math and ELA tests in fourth and eighth grades. *Educational Researcher*, *47*, 284–294.
- Reed, D. K., & Vaughn, S. (2012). Retell as an indicator of reading comprehension. *Scientific Studies of Reading*, *16*, 187–217.
- Scardamalia, M., & Bereiter, C. (1987). Knowledge telling and knowledge transforming in written composition. In S. Rosenberg (Ed.), *Advances in applied psycholinguistics* (Vol. 2, pp. 142–175) Cambridge.
- Shavelson, R. J., Webb, N. M., & Rowley, G. L. (1989). Generalizability theory. *American Psychologist*, *44*, 922–932.
- Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer* (Vol. 1). Sage.
- Spencer, M., Gilmour, A. F., Miller, A. C., Emerson, A. M., Saha, N. M., & Cutting, L. E. (2019). Understanding the influence of text complexity and question type on reading outcomes. *Reading and Writing: An Interdisciplinary Journal*, *32*, 603–637 Doi:10/cxrn.
- Spear-Swerling, L. (2004). Fourth Graders Performance on a State-Mandated Assessment Involving Two Different Measures of Reading Comprehension. *Reading Psychology*, *25*, 121–148. <https://doi.org/10.1080/02702710490435727>.

- Spencer, M., Quinn, J. M., & Wagner, R. K. (2014). Specific reading comprehension disability: Major problem, myth, or misnomer? *Learning Disabilities Research & Practice, 29*(1), 3–9.
- Stuebing, K. K., Fletcher, J. M., Branum-Martin, L., & Francis, D. J. (2012). Evaluation of the technical adequacy of three methods for identifying specific learning disabilities based on cognitive discrepancies. *School Psychology Review, 41*(1), 3–22.
- Swiss Society for Research in Education Working Group. (2012). EduG (Version 6.1) [Computer software]. Unpublished instrument. Retrieved from <http://www.irdp.ch/edumetric/englishprogram.htm>.
- Tannenbaum, K. R., Torgesen, J. K., & Wagner, R. K. (2006). Relationships between word knowledge and reading comprehension in third-grade children. *Scientific Studies of Reading, 10*(4), 381–398.
- Vellutino, F. R., Tunmer, W. E., Jaccard, J. J., & Chen, R. (2007). Components of reading ability: Multivariate evidence for a convergent skills model of reading development. *Scientific Studies of Reading, 11*(1), 3–32.
- Walker, R. (2017). *The effect of response format and presentation conditions on comprehension assessments for students with and without a reading disability* (Doctoral dissertation). Retrieved from ProQuest Dissertations & Theses Full Text. (10601892)
- Webb, N. M., Rowley, G. L., & Shavelson, R. J. (1988). Using generalizability theory in counseling and development. *Measurement and Evaluation in Counseling and Development, 21*, 81–90.
- Webb, N. M., Shavelson, R. J., & Haertel, E. H. (2006). 4 reliability coefficients and generalizability theory. *Handbook of Statistics, 26*, 81–124.
- Wixson, K. K. (2017). An interactive view of reading comprehension: Implications for assessment. *Language, Speech, and Hearing Services in Schools, 48*(2), 77–83.
- Woodcock, R. W., McGrew, K. S., & Mather, N. (2001). *Woodcock Johnson III Tests of achievement*. Rolling Meadows: Riverside.

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

## Affiliations

Alyson A. Collins<sup>1</sup> • Esther R. Lindström<sup>2</sup> • Micheal Sandbank<sup>3</sup>

Esther R. Lindström  
esl417@lehigh.edu

Micheal Sandbank  
msandbank@austin.utexas.edu

<sup>1</sup> Department of Curriculum & Instruction, Texas State University, San Marcos, TX 78666, USA

<sup>2</sup> Department of Education and Human Services, Lehigh University, Bethlehem, PA, USA

<sup>3</sup> Department of Special Education, The University of Texas at Austin, Austin, TX, USA