INTERNATIONAL LITERACY ASSOCIATION READING RESEARCH QUARTERLY

COMMENTARY

Commercially Developed Tests of Reading Comprehension: Gold Standard or Fool's Gold?

Nathan H. Clemens
The University of Texas at Austin, USA

Douglas Fuchs
Vanderbilt University, Nashville, Tennessee, USA

ABSTRACT

Many seem to believe that researcher-made tests are unnecessary, if not inappropriate, for evaluating reading comprehension interventions. We suggest that this view reflects a zeitgeist in which researcher-made (proximal) tests that align with the researchers' interventions are closely scrutinized and often devalued, whereas commercially developed (distal) tests, typically unaligned with the researchers' interventions, escape such examination and judgment. We take issue with the zeitgeist. We object to what appears as an unthoughtful rejection of proximal tests and acceptance of the distal ones. We do so first by discussing the multidimensionality of reading comprehension; then, we present evidence that commercial tests explore few of the construct's many dimensions, differ among themselves in the dimensions they address, and often have little to do with the aims and substance of researchers' comprehension interventions. We argue that these facts are reason enough to reconsider commercial tests as the gold standard in program evaluation. Finally, we make a case for the supplemental use of researcher-made tests and offer a framework to help researchers develop tests of reading comprehension that align more and less with their intervention programs.

Purported Failure to Strengthen Reading Comprehension

Researchers have learned much about reading comprehension (Berkeley, Scruggs, & Mastropieri, 2010; Elleman, 2017; Kim, Linan-Thompson, & Misquitta, 2012; McNamara & Kendeou, 2017) and have produced instructional programs to strengthen elementary and secondary students' reading for understanding. Yet, meta-analyses of formal evaluations of these programs have indicated only modest effects on student performance, with recent effects weaker than those calculated in earlier evaluations (e.g., Scammacca et al., 2016; Scammacca, Roberts, Vaughn, & Stuebing, 2015). A closer look at these meta-analyses shows that authors of the more recent evaluations were more likely to use commercially developed rather than researcher-made tests to index program impact (e.g., Edmonds et al., 2009; Hebert, Bohaty, Nelson, & Brown, 2016; Scammacca et al., 2016; Solis et al., 2012). Several research teams relied on both kinds of tests and reported students performing stronger on the researcher-made measures (e.g., Cheung & Slavin, 2016; Connor et al., 2018; D. Fuchs et al., 2018; Jones et al., 2019; Solis, Vaughn, Stillman-Spisak, & Cho, 2018; Stevens, Park, & Vaughn, 2019; Stevens, Vaughn, Swanson, & Scammacca, 2020; Swanson et al., 2017; Vaughn et al., 2019).

INTERNATIONAL LITERACY ASSOCIATION

Reading Research Quarterly, 0(0)
pp. 1–13 | doi:10.1002/rrq.415

Patton et al. (2020), for example, administered researcher-made and commercially developed measures to evaluate the efficacy of an instructional program to strengthen students' understanding of social studies and science texts. Poor readers in grades 4 and 5 were randomly assigned to two versions of the program and controls. One program variant taught strategies to promote transfer (Comp+Transfer), and the other did not (Comp-Only). The treatments were implemented in small groups three times per week for 14 weeks. There were four outcome measures: a knowledge acquisition assessment and near-transfer, mid-transfer, and far-transfer tests. Only the far-transfer test was commercially developed. On these measures, respectively, the fourth-grade Comp+Transfer group performed 2.61, 1.30, 0.71, and 0.18 standard deviations better than controls. The Comp-Only group's performance differed by 2.82, 1.23, 0.53, and −0.05 standard deviations from controls. Thus, regardless of treatment group, there was a clear stepwise pattern: strongest performance on researcher-made knowledge acquisition and near-transfer tests, weakest performance on commercial far-transfer tests (Weschler Individual Achievement Test and Gates–MacGinitie Reading Test), and moderately strong performance on a researcher-made test of mid-transfer.

Such inconsistency in students' reading performance has commonly been explained by the fact that researcher-made tests are usually aligned with the researchers' program content. Skills taught are the skills tested (Cheung & Slavin, 2016). If a program's emphasis is to help students identify the main idea of a passage, then the researcher-made test asks them to identify the main idea, often in passages reflecting similar, if not identical, subject content and formats as the instructional materials. In contrast, many commercially developed tests address skills unrelated to those targeted by a program and do so with novel formats.

The stronger performance on researcher-made tests has tended to be discounted because of the alignment between the researchers' tests and interventions. Weaker performance on the commercially developed tests has been interpreted as program participants' failure to generalize from what was learned during intervention to measures of reading comprehension unaligned with it (Slavin, 2019). Implicit is a devaluing of both the efficacy of the researchers' interventions and the validity of the self-made outcome measures.

Different View

In this commentary, we argue that how one chooses to measure reading comprehension is centrally important to the judgments one makes about program or intervention effectiveness. Whereas this may seem too obvious a proposition, we believe that its importance has been inadequately considered. We contend that researcher-made tests are indispensable for evaluating the effects of reading comprehension interventions. The effects can tell program developers how well their instruction strengthens the very skills they considered necessary to target, and can provide likely explanations of how and for whom the programs are beneficial.

Moreover, an exclusive reliance on commercially developed tests can distort an understanding of what works and does not work, partly by their focus on skills and processes not addressed by a given program. Despite this, researcher-made tests have been devalued as part of an informal but generally held perspective that promotes commercial tests as necessary and sufficient for bolstering a study's experimental rigor and importance. We start this commentary with the proposition that commercial tests are the gold standard of measurement and evaluation.

Commercially Developed Tests as a Gold Standard

During the past 20 years, the use of standardized,[1] normative, commercially developed measures to evaluate effects of reading comprehension interventions has increased greatly (Edmonds et al., 2009; Scammacca et al., 2015, 2016), no doubt in part because there are many more of them now to choose from than in years past. A more important explanation, however, is a zeitgeist that would have one view commercial tests as central to achieving methodological rigor and for producing accurate estimates of program or intervention effects. It is not difficult to find evidence of this perspective.

Best-Evidence Syntheses

Beginning in the 1980s, Slavin (e.g., 1986) argued for best-evidence syntheses, or integrative reviews of primary studies that met a set of indicators for methodological quality and rigor. Slavin (1995) offered best-evidence synthesis as an intelligent alternative to meta-analysis, which, as described by Glass (1977) and others, encouraged the inclusion of all pertinent (low- and high-quality) studies that would subsequently be organized according to the studies' substantive and methodological features, including study quality.

Best-evidence syntheses of reading programs usually relied on effect sizes calculated on students' performance on standardized, norm-referenced, commercial tests (e.g., Cheung & Slavin, 2005; Slavin, Cheung, Groff, & Lake, 2008; Slavin, Lake, Chambers, Cheung, & Davis, 2009). Studies involving researcher-made outcome measures were mostly excluded from these syntheses. Whereas Slavin and colleagues (Cheung & Slavin, 2016; Slavin, 2019; Slavin et al., 2009) noted that researcher-made tests are useful in theory building and other respects, they advised that the only data to be taken seriously as a reflection of program efficacy were produced by commercially developed tests. Slavin (2019) argued more forcefully in this regard: "Reports of effect sizes from researcher/developer measures should be treated as implementation measures, not outcomes. The outcomes emphasized should only be those from independent measures" (para. 9). Typically absent from discussions on best-evidence syntheses is consideration of what the independent measures are actually measuring.

Evidence-Based Practices

A second indication of the perspective that privileges commercial tests over researcher-made tests may be seen in various well-known initiatives to formally identify and disseminate evidence-based practices. Consider this statement in a guidance document from the respected What Works Clearinghouse (WWC; 2020):

> A third requirement of outcome measures [in efficacy studies under review by the WWC] is that they not be *overaligned* with the intervention….When outcome measures are closely aligned with or tailored to the intervention, the study findings may not be an accurate indication of the effect of the intervention. (p. 84)

The authors of the guidance document provided this example of overalignment: "an outcome measure based on an assessment that relied on reading materials or vocabulary words used in the intervention condition but not in the comparison condition" (p. 84).

We suspect that the WWC (2020) guidance document authors wanted to discourage teaching to the test, an intent we support. However, their discussion of alignment/overalignment is inadequate. Alignment is not always (or even frequently) a binary construct. Many tests are not simply aligned or unaligned with an intervention. Rather, alignment should be considered in terms of a continuum; that is, a test is more or less aligned with an intervention. Moreover, the point at which it becomes overaligned is often inherently subjective and arbitrary. How one views alignment/unalignment is probably influenced by context, including the questions researchers and others want answered by measuring student performance. So, it would have been helpful if the authors of the WWC guidance document had recognized this complexity and discussed the issue with more nuance. The absence of more thoughtful discussion in the guidance document, we believe, has encouraged incorrect and unhelpful either/or thinking about alignment and has led to an unfortunate devaluation of researcher-made tests.

It is important to consider the influence of the WWC's (2020) apparent position on test alignment. Most would probably agree that when the WWC characterizes the evidence from a developer's program evaluation as significant "without reservation," it is a notable achievement. The designation draws attention to the developer and research team for demonstrating a high level of methodological rigor and gives visibility to the team's instructional program. Our impression is that the WWC and others have contributed to a popular view of commercial tests as a most trustworthy means of helping researchers and practitioners identify best-evidence practices. Researchers aspiring to obtain grant funding to develop reading comprehension programs that will eventually be reviewed favorably by the WWC, and to publish papers about these programs in respected journals, are likely to conclude that they must use unaligned commercially developed tests as principal, if not exclusive, indicators of intervention effects.

Balanced View of Researcher-Made and Commercially Developed Measures

Whereas we are suggesting that there is a generally shared view that privileges commercially developed distal tests over researcher-made proximal ones, we recognize that scholars and professional organizations have occasionally argued for the importance of using multiple outcomes that vary in their strong-to-weak alignment with interventions. As an example, Gersten et al. (2005), in an article discussing quality standards for intervention studies, recommended that "an appropriate balance [be struck] between measures closely aligned with the intervention and measures of generalized performance" (p. 151). Similarly, in *Common Guidelines for Education Research and Development*, the Institute of Education Sciences and the National Science Foundation (2013) suggested,

> Primary outcome measures should include student outcomes sensitive to the performance change the intervention is intended to bring about (e.g., researcher-developed measures that are aligned with the experiences of the treatment group), student outcomes not strictly aligned with the intervention, and student outcomes of practical interest to educators and policymakers (e.g., test scores, grades, graduation or dropout rates). (p. 22)

Notwithstanding such support for a balanced approach to measuring intervention effects, we are skeptical of its influence on the research community where a less-than-balanced perspective appears to prevail. As noted, meta-analysts of studies of reading comprehension interventions have documented the increased use of commercial tests for evaluative purposes (e.g., Edmonds et al., 2009; Hebert et al., 2016; Scammacca et al., 2016; Solis et al., 2012), which we take as indicative of a prevailing view of them as more methodologically sound and trustworthy than researcher-made tests.

To be sure, there is good reason to critique researcher-made tests of reading comprehension. The content and format of some of these measures have indeed been over-aligned with researchers' interventions. Many such measures have lacked sufficient development work, reflected by inadequate reliability, item difficulty, and validity.

Commercial tests, by contrast, have mostly been spared such critical attention and judgment. Instead, as indicated, the research community has tended to assume that they represent a gold standard outcome, that they are trustworthy indices of program effects. Yet, there is ample evidence to question this view. These tests, we believe, are as deserving of the same degree of inspection (and skepticism) typically shown researcher-made measures.

What Do Tests of Reading Comprehension Really Test?

Complexity of the Construct

Our critique of commercial tests of reading comprehension begins with recognition of the complexity of the construct these tests are designed to measure. Reading with understanding depends on intricate (not fully understood) interactions of many cognitive processes, such as attention, working memory, reasoning, and inferential thinking; on sensitivity to the structure of language; on background knowledge and vocabulary development; on motivation; on the use of strategies such as self-monitoring; and on word-reading accuracy and efficiency (e.g., Ahmed et al., 2016; Cain & Oakhill, 2009; Cain, Oakhill, & Lemmon, 2004; Cromley & Azevedo, 2007; Cutting, Materek, Cole, Levine, & Mahone, 2009; Gough & Tunmer, 1986; Nation, 2005/2009; Peng et al., 2018; Perfetti, 1985; Perfetti & Stafura, 2014; Tighe & Schatschneider, 2014; van den Broek, White, Kendeou, & Carlson, 2009; Verhoeven & van Leeuwe, 2008). Difficulties in any of these areas may impede reading comprehension.

Moreover, these cognitive processes interact with at least two additional elements to influence the extent to which a reader understands text (RAND Reading Study Group, 2002). The first of these refers to the features and content of text, such as genre (e.g., biography, opinion), text structure (e.g., problem and solution, compare and contrast), and text complexity, all of which can influence the quality of reader's mental representation. A second element involves the activity and purpose of reading (e.g., acquiring knowledge, answering questions), which can determine the reader's motivation and engagement, which in turn can affect how deliberately the reader reads. Variations in these elements have important implications for estimating reading comprehension. Also, unlike word-reading accuracy, which can be observed in real time, the reader's mental representation and understanding of text is more covert, requiring assumptions and inferences to estimate its quality (Fletcher, 2006). Measuring such a multiply determined and opaque construct is highly prone to misinterpretation, which may lead to overly conservative and inaccurate representations of reading comprehension as a single ability that can be measured by a single test (Fletcher, 2006; Kamhi & Catts, 2017; Wixson, 2017).

Differences Among Tests of Reading Comprehension

A Press for Efficiency and an Underrepresentation of the Construct

In the absence of a consensual definition of the construct, authors of commercial tests choose which dimension(s) to measure and which to ignore. Program developers do likewise, focusing on aspects of reading comprehension they consider most important for the students whose reading performance they are hoping to influence. When operationalizing reading comprehension for their respective purposes, test developers and program developers understand that their products must be efficient because of the little time typically permitted them by practitioners. For test developers, this means creating measures with a relatively narrow focus and with simple items that can be administered quickly and scored accurately.

This press for a relatively narrow focus affects development of reading comprehension measures in at least two important ways. First, the measures tend to address only a small part of the construct, which is to say there are few comprehensive multidimensional measures of reading comprehension. Messick (1995) would likely say these measures reflect construct underrepresentation, an unreasonably narrow approach to measurement, lacking sufficient coverage of important dimensions of comprehension. Similar points about the risks of construct underrepresentation in education research were raised by Briggs (2008) and Schoenfeld (2006).

Second, the measures differ among themselves, often dramatically so (D. Fuchs et al., 2018). In the Appendix, we summarize various characteristics of commercial tests frequently used in reading comprehension research. These tests are strikingly different from one another with respect to the texts students read, questions they answer, and response modes permitted them. Unsurprisingly, these tests of reading comprehension do not correlate strongly with each other (e.g., Clemens et al., 2020; Francis et al., 2006; Keenan, Betjemann, & Olson, 2008; Keenan & Meenan, 2014). For example, in a sample of 995 younger and older students, Keenan and Meenan (2014) obtained an average correlation of .54 (range = .45–.68) among four commercial tests of reading comprehension.

Different Test Features and Implications for Performance

Variability in texts, questions, and modes of response result in unique and implicit weightings of the importance of students' skills and knowledge. Decoding and word recognition skills appear more important on comprehension measures that involve short passages or cloze tasks (Andreassen & Bråten, 2010; Clemens et al., 2020; Francis et al., 2006; García & Cain, 2014; Keenan et al., 2008; Keenan & Meenan, 2014; Nation & Snowling, 1997; Spear-Swerling, 2004). Oral language skills are more consequential on tests with longer passages (Francis et al., 2006; Keenan et al., 2008) and on tests requiring responses to questions asked verbally (Clemens et al., 2020). Struggling readers perform more like average readers on tests with response formats requiring lower level text processing (e.g., retell, sentence verification, cloze) than on tests requiring responses to open-ended and multiple-choice questions (Collins, Lindström, & Compton, 2018). Background knowledge seems to exert greater influence when students must show their understanding of informational texts rather than narrative texts (Best, Floyd, & McNamara, 2008). Tests with longer administration times appear to differentially advantage students with sustained attention and attentional control (Arrington, Kulesz, Francis, Fletcher, & Barnes, 2014; DiCerbo, Oliver, Albers, & Blanchard, 2004; Kieffer, Vukovic, & Berry, 2013; Vavassoeur, 2016).

On the Validity of Far-Transfer Tests

There is another related problem when commercial tests of reading comprehension are selected to index program effects: They often have little substantive connection to these programs. That is, by their selection of these far-transfer tests, program developers (and other program users) knowingly or otherwise have often oriented their evaluations to measure knowledge, skills, or strategies that the programs were never meant to address. Scammacca (personal communication, December 2020), reflecting on syntheses she and her colleagues completed of studies exploring the efficacy of reading comprehension programs (Scammacca et al., 2015, 2016), stated that the most commonly used commercial test among these studies was the Passage Comprehension subtest from the Woodcock–Johnson Tests of Achievement—a cloze task involving short passages that is highly influenced by decoding skills (Francis et al., 2006; Keenan et al., 2008; Keenan & Meenan, 2014). If the intent of a developer of a reading comprehension program is to strengthen inference making, say, in science and social studies texts, does it make sense for the developer to measure program efficacy by using the Passage Comprehension subtest? We suspect that many would say no. Yet, more than a few program developers do precisely this (or its equivalent).

To be clear, we are not arguing against testing the transfer of learning as part of an evaluation of program effects. In principle, it can be an important dimension of such an evaluation. Rather, our concern is that the research community has given too little thought to what represents tests of appropriate or reasonable transfer. When considering the use of a specific comprehension measure to determine the value of a given program, how does one know whether the measure's probing of students' knowledge, skills, or strategies represents a bridge too far, whether its focus is so removed from the program's intent and content as to be unfair to students and the developer?

It may be instructive to recall one of the more important messages expressed in the 1966 edition of the *Standards for Educational and Psychological Tests and Manuals* (American Psychological Association, American Educational Research Association, & National Council on Measurement in Education): The validity of any test must be based in part on consideration of the specific purpose(s) to which it will be put. So, should we consider the Passage Comprehension subtest referenced earlier as a valid measure to help determine the benefits of a program aiming to strengthen students' inference making in informational texts? We believe the general issue here requires more theoretical and empirical work.

Reconsidering How to Measure the Impact of Reading Comprehension Instruction

There is a need to rethink how we measure the efficacy of reading programs and interventions, especially those meant to strengthen students' comprehension (e.g., L.S. Fuchs & Fuchs, 1994). Program developers and interventionists should reappraise whether efficacy is best determined by an exclusive reliance on far-transfer tests. Also, if near-transfer tests are to be added to the mix of outcome measures, how closely aligned should they be to the programs targeted for evaluation?

Such rethinking might include consideration of the use of multiple measures of the construct, collectively representing a continuum of closely aligned tests to unaligned tests. This continuum of measurement might facilitate systematic exploration of whether instruction has strengthened students' comprehension and, if so, the degree to which this improvement transfers to familiar and unfamiliar learning contexts. Whereas this proposal may strike some as unorthodox, there are important reasons why measures aligned with targeted skills can benefit the identification of best-evidence practices, bridge the gap between what the field knows and does, and advance the science of reading comprehension, including program development.

Specific Skill Development

Catts and Kamhi (2014) argued that reading comprehension is not a single ability and should not be measured or targeted as if it were: "Instruction will be more effective when…tailored to students' abilities with specific texts and tasks. This instruction would involve identifying educationally relevant reading comprehension activities and *directly assessing the component skills and knowledge bases involved in these activities* [emphasis added]" (p. 74). For example, evaluating the persuasiveness of an author's argument, retelling a narrative in one's own words, writing a book report, and imposing order on details from informational texts are all specific tasks that require reading comprehension. Importantly, they represent specific and different applications of comprehension processes. An evaluation of an intervention targeting these comprehension-specific activities requires measures aligned with them.

Mediators and Moderators of Instructional Effects

Mediation and moderation analyses are infrequently used to evaluate effects of reading comprehension programs; rarely are these analyses used to determine the effects of any educational program for that matter. Nevertheless, they can be very important means of exploring theories of change underlying instructional programs, as well as identifying for whom the programs are more and less beneficial (D. Fuchs & Fuchs, 2019).

Mediation Analysis

With mediation analysis, a team of program developers who hypothesize, for example, that improving students' main idea generation is partially responsible for (i.e., causal to) an overall improvement in their reading comprehension must first demonstrate that the program led to improvements in main idea generation (the mediator). Only then can the team determine whether changes in the mediator were associated with overall improvements in reading comprehension. Mediation analyses are particularly useful when evaluating multicomponent interventions to explore the components that seem most important (or the active ingredients) in improving desired outcomes. The relevant point here is that mediation analysis relies on measures aligned with the component skills targeted by the intervention. Often, this requires use of valid researcher-developed tests of the various dimensions of the reading comprehension construct.

Consider the following study, which does not pertain to reading comprehension but illustrates the benefits of aligning measures with interventions to test hypothesized mediators. Stice, Presnel, Gau, and Shaw (2007) investigated intervention effects on the eating disorder symptoms of adolescent girls. One group of girls was assigned to a dissonance intervention. They engaged in counterattitudinal exercises to critique a "thin ideal" perception that is often associated with eating disorders. A second group was assigned to a healthy weight intervention that encouraged the girls to make sustainable healthy changes to their diet and physical exercise. To test for mediation effects on overall eating disorder symptoms, the researchers included self-developed measures aligned with one or the other intervention: a "thin ideal" internalization scale and a set of related measures of healthy eating and physical activity.

Both the dissonance and healthy weight interventions had statistically significant positive effects on reducing eating disorder symptoms. Both also exerted positive effects on their respective mediators (i.e., the dissonance intervention reduced the "thin ideal" internalization; the healthy weight intervention improved habits of healthy eating and exercise). However, there was a stronger relation between change in the "thin ideal" internalization mediator and change on desired outcomes for the dissonance intervention than there was in the relation between the healthy eating and exercise mediator and outcomes for the healthy weight intervention. In other words, there was stronger support for the dissonance intervention's hypothesized causal pathway. This exploration of the interventions' respective theories of change was made possible by researcher-made tests aligned with the attitudes and behaviors targeted for change by the interventions.

Moderation Analysis

To illustrate the importance of moderator analysis, and the usefulness of program-aligned measures to examine program effects, we discuss an evaluation of a first-grade reading comprehension program (D. Fuchs et al., 2019) for at-risk students. The researchers conducted a component analysis of the program's word decoding/fluency (DF) and reading comprehension (RC) dimensions by creating DF and DF+COMP treatments to parse the value of RC. The researchers also developed a set of measures that tested students' word-reading, nonword-reading, and reading comprehension skills. The 125 students were randomly assigned to the two treatments and controls. Treated students were tutored individually three times per week for 21 weeks in 45-minute sessions.

Students in the DF and DF+COMP groups together performed more strongly than controls on word reading and comprehension. However, the treatment students' pretreatment word reading moderated these results. Across treatments, students with weaker beginning word reading outperformed similar controls to a greater extent than students with stronger beginning word reading outperformed controls that were comparable to them. Pretreatment word reading was calculated as a factor score involving two commercially developed tests (the Sight Word Efficiency subtest of the Test of Word Reading Efficiency and the Word

Identification subtest of the Woodcock Reading Mastery Tests) and a researcher-made test (Word Identification Fluency). Thus, the moderator analysis indicated that the reading program was better for some of the at-risk students but not for all. This qualification, with important implications for program use, was based on data from a combination of measures aligned and unaligned with program content.

Relying on a Hypothesized Theory of Change to Guide the Measurement Plan

The points above speak to a need for researchers to build a measurement plan around their theory of change—their hypothesized model of how the intervention (and its components) is expected to change students' skills and which skills are more likely to be affected by it. These skills may be specific and proximal to the intervention and more general and distal to it. Whereas researchers commonly propose theories of change as part of grant proposals, they do not necessarily consider their theories when operationalizing their study methods. A theory of change can provide researchers with a map of the skills that should be measured, which in turn can lead to tests aligned with those skills. Such tests may include commercial measures or, when unavailable, researcher-made tests.

What Should Researcher-Made Tests Look Like?

We have suggested that researcher-made near-transfer measures may be (at least) as valuable as commercially developed far-transfer tests for understanding whether, how, and for whom reading comprehension programs are beneficial. If we assume this suggestion has merit, a next-order issue is, How does one construct such tests, and what should they look like? If the use of these measures is taken seriously, how should *near-transfer* be defined?

As indicated, test developers and program developers typically think about and operationalize the reading comprehension construct independently of, and differently from, each other. They may conceptualize comprehension as recalling factual information, making main idea statements, or constructing inferences. The test stimuli may be presented visually or orally, the tests may require students to respond orally or in writing to multiple-choice or open-ended questions, and the tests may be timed or untimed, administered individually or in small or large groups. On which of these (and additional) dimensions should program developers choose to coordinate their instruction with commercial or self-made outcome measures? On all? On only a few? If on a subset, on what basis would this subset be chosen? Moreover, how closely should developers match their programs/interventions to their measures without giving the treated students an undue advantage over nontreated controls? As noted earlier, measurement of program and intervention effects can be driven by researchers' hypotheses about the skills the programs are expecting to change. However, there is little guidance on how to proceed with the actual coordination between programs and their measurement and with the design of the researcher-made tests.

Design Framework

To explore the efficacy of a tutoring program developed to strengthen poor readers' understanding of informational texts, D. Fuchs et al. (2018) and Patton et al. (2020) constructed a set of comprehension measures that differed in the degree to which they aligned with the tutoring program. The matrix in Figure 1 guided the teams' development of these measures on four dimensions: strategies/skills, passage content, layout and format, and question types. The researcher-made tests described in the matrix illustrate the extent to which each was deliberately aligned to the instructional program and how collectively they ranged from most closely aligned ("Knowledge Acquisition") to least closely aligned ("Far Transfer") with the content of the instructional program.

Strategies/Skills

The first dimension, strategies/skills, refers to a test's purpose or substantive focus. If a research team develops an instructional protocol to help students infer word meaning from text, and if the team wants to determine the protocol's impact on this strategy/skill alone, then the team may develop a near-transfer test with a correspondingly narrow focus. By contrast, if the researchers wanted to evaluate their program's broader impact on inference making—for example, on recognizing an author's voice or tone, for which word meaning is necessary—they might develop a less aligned mid-transfer or far-transfer test.

A different group of researchers may be developing a program to strengthen not one but multiple strategies/skills, such as a program for improving students' comprehension of science and social studies texts. Instructional activities might include helping students identify various text structures (e.g., cause and effect, compare and contrast) and text features (e.g., titles, headings, bolded text, tables, maps) and encouraging them to monitor their understanding while reading. A closely aligned test might evaluate performance on each of these skills/strategies, whereas a less aligned test might require the coordinated use of all taught skills/strategies to find the main idea in a paragraph or book chapter.

FIGURE 1
Suggested Framework for Developing Tests of Reading Comprehension

Test features					
Level of alignment	Test	Strategy/skill required	Passage content	Layout and format	Question type/ response mode
Most aligned	Target skill acquisition	One targeted skill (e.g., inference making)	Identical to the texts used in the intervention	Passage layout and format identical to intervention materials	Multiple-choice (identical to intervention materials)
	Near-transfer	Two targeted skills (e.g., inference making, main idea identification)	Passages of the same genre (informational) and related thematically (but not identical) to the texts in the intervention	Passage layout and format identical to intervention materials	Multiple-choice (identical to intervention materials)
	Mid-transfer	All targeted strategies/skills (inference making, main idea, and prereading)	Passages of the same genre (informational) but unrelated thematically to the texts in the intervention	Less similar to intervention materials	Multiple-choice and short-answer
Least aligned	Far-transfer	All targeted skills	Passages of various genres, text structures, and themes	No/little resemblance to intervention materials	Multiple-choice, short-answer, cloze, and sequencing

Note. The figure provides a framework for how a test (or set of tests) could be constructed to measure reading comprehension depending on a desired level of alignment with an intervention program. The cells include examples of how D. Fuchs et al. (2018) and Patton et al. (2020) approached the alignment of each test feature. One or any number of tests could be included, which would be based on the skills, constructs, and degree of transfer a researcher wishes to measure in evaluating hypothesized treatment effects.

Passage Content

A second dimension of the matrix, passage content, pertains to the subject matter or topic of a test's passages, as well as the author's use of vocabulary, text structures, and genre(s). Test passages and instructional texts may be made to correspond tightly, loosely, or not at all on these features. By manipulating the extent of this correspondence, program developers can explore how well students apply learned skills/strategies to increasingly novel and complex texts. Developers may choose to deliberately vary one aspect of passage content (e.g., topic) while keeping other features (text structure, vocabulary, and genre) constant.

Layout and Format

The next dimension, layout and format, refers to the appearance and length of test passages. As with skills/strategies and passage content, layout and format may be similar or dissimilar to a given instructional text. Tests formatted to be similar will appear more familiar to students and will presumably require less transfer of learning. That is, tests and texts with similar formats may facilitate students' application of learned strategies or skills when reading the text for understanding. For example, students who learn to use paragraph breaks as cues to self-monitor for understanding may more readily apply this strategy to a test that includes multiparagraph passages than to a test that makes use of single sentences or lone paragraphs. This latter test would not offer the same prompts as the former test and would likely require greater transfer.

Question Types and Response Modes

The fourth and last dimension of the matrix is question types (e.g., cloze, multiple-choice, retell, open-ended) and response modes (e.g., written, verbal). Question types and response modes may correspond strongly or weakly with instruction. For example, a multiple-choice test may align more closely with instruction that requires students to practice responding to written questions than to instruction that asks students to respond to questions presented orally by the tutor. Tests that include question types or response formats different from the instructional protocol require greater learning transfer.

Use of the Framework

We believe that the matrix in Figure 1 offers a heuristic for how one might approach the development of researcher-made measures in connection with a specific instructional program. The matrix provides a framework for thinking about the design of measures that comport with a theory of change and a desired level (or levels) of alignment with a given program. Any number of measures can be conceptualized in the matrix. Frameworks such as this may also help researchers describe their tests, and the connections between their tests and hypothesized mechanisms of change, more clearly and persuasively. Such a framework may help researchers operationalize, clarify, and justify the use of self-made and commercial tests, thereby providing a basis for clearer and more productive explanations of a program's effects.

Psychometric Adequacy of Researcher-Made Tests

As we stated earlier, a common criticism of researcher-developed tests has been that they are less psychometrically sound than commercial tests. This is not always the case, as it is sometimes true. Weak researcher-made tests partly reflect the fact that most program developers do not have the time or money to undertake test and program development, both of which, by necessity, are almost always iterative processes. In light of this, we offer two kinds of recommendations to program developers and other researchers. The first is a perspective on testing that differs from conventional thinking in important respects. The second is suggestions about how to evaluate a test's technical properties in the context of intervention implementation.

Rethinking Test Validity

We suggest that researchers should think differently about test validity, both generally and specifically with regard to tests of reading comprehension. Validity is commonly understood as correlation coefficients between and among tests based on scores collected at the same time (concurrently) or across time (predictively). Implicit is an assumption that test validity generalizes across groups, contexts, and purposes to which one may apply test scores. Messick (1995), however, contradicted this perspective, arguing that validity is less a property of the test itself and more about what test scores mean and how they are used. Such a focus, Messick claimed, is justified (necessary, in fact) because validity is fundamentally situational. It is influenced by the sample of test takers, by context, and maybe most importantly, by how test scores are used, reflecting a reality that a test may be used for a variety of purposes and may be valid for some purposes and not for others.

In accordance with this view, we believe that establishing the validity of a test should be driven by construct representation; that, when all is said and done, it is the researcher's responsibility, not the test author's, to articulate how test scores represent the construct of the researcher's interest and how she or he interprets the scores as evidence of program effects. Our previously explained framework for test construction is but one way of attempting to establish the construct validity of a test (or tests) for a specific purpose. Moreover, we believe that commercial tests used to evaluate program effects should be subject to the same scrutiny as researcher-made tests regarding what these tests purport to measure and whether obtained scores are plausible representations of program effects.

Strengthening the Technical Properties of Tests

We suggest to program developers and other researchers that they should provide clear and comprehensive descriptions of their self-made reading comprehension tests. We recommend reporting the number, average length, and genre(s) of the reading passages. The readability of these passages can be described in terms of Lexile levels or by means of text analysis programs such as Coh-Metrix (Graesser, McNamara, Louwerse, & Cai, 2004). Researchers should also specify the types of questions asked of students and the permissible modes of response, indicate whether their tests are timed or untimed, and so forth. Also, as we argued throughout this commentary, an instructional program's degree of alignment with various measures should be explained, and the framework in Figure 1 provides one means of doing so.

Additionally, estimates of internal consistency, and when possible, test–retest reliability, should be derived from scores produced as part of an intervention trial. As previously explained, an important misunderstanding of commercial tests is that they are more reliable than researcher-made tests. The reliability coefficients reported by developers of commercial tests are estimates obtained on a sample tested under specific conditions. Because test conditions vary, users of these commercial tests may obtain different reliabilities than those reported by the test developers. Using their respective study samples, program developers and other researchers should document the reliability estimates of self-made tests and those of the commercial tests they choose to use. Interscorer agreement should also be reported, especially if a test involves rating the quality of written or oral responses.

Conclusion

Despite an apparent zeitgeist that would have one view commercial tests of reading comprehension as a gold standard, there is little evidence that they are universally trustworthy indices of program effects or that they necessarily contribute to the methodological soundness of program evaluations. In this commentary, we argued that understanding the influence of multidimensional programs on complex behaviors such as reading comprehension requires a more expansive and strategic approach to educational measurement.

We suggested that researchers should use multiple measures that vary in how they correspond to an instructional program. Some of these measures may closely resemble the program's content, materials, and task demands. Others may range from somewhat different to very different from the instructional features of the program. One or more outcome measures may be researcher developed.

Strategic use of multiple measures, which collectively may constitute a continuum of near to far program test alignment, should reduce the risk of underrepresenting (misrepresenting) the reading comprehension construct and lead to more nuanced and meaningful program evaluation. With the conceptual framework we proposed, or some variation of it, researchers may think more precisely about how their intervention components and measures align, so they may explore straightforward or more complex hypothesized effects. Moreover, the nature of program–measure alignment should be considered when developers and others weigh the quality of a program's efficacy data and whatever claims are made about its benefits.

Educators are in need of effective and practical instructional practices, including protocols to strengthen students' reading for understanding. Teachers need to know who these evidence-based practices are likely to benefit and who will probably not benefit from them. Developing evidence-based practices, and understanding their differential effects, is difficult work. It requires clever and informative evaluation. It is past time to recognize that psychometrically strong researcher-developed tests can offer unique and important information, formative and summative, on program effects. We hope this commentary contributes to a more balanced perspective on the virtues of researcher-made and commercially developed tests of reading comprehension and encourages adoption of more comprehensive and meaningful assessment frameworks to better understand program outcomes.

Toward this end, we call for more thoughtful discussion about a proper role for researcher-made tests in program evaluation. The writing team responsible for the next version of the *Standards for Educational and Psychological Tests and Manuals*, representing the American Psychological Association, American Educational Research Association, and National Council on Measurement in Education, might consider this. So, too, might the authors of the WWC. Professional organizations, such as those just mentioned, and funding agencies may separately or jointly convene meetings to review how program evaluation, especially with regard to reading, may be made stronger and more meaningful.

NOTES

Preparation of this commentary was funded in part by a grant (R324D13D0003) from the National Center for Special Education Research, Institute of Education Sciences, U.S. Department of Education, to Vanderbilt University. The authors are responsible for the content herein, which does not necessarily represent the views of the Institute of Education Sciences or the U.S. Department of Education. We thank Amy Elleman for her helpful comments on an early draft of the manuscript.

[1] The authors of the meta-analyses cited here (Edmonds et al., 2009; Scammacca et al., 2015) used the term *standardized* to describe commercial norm-referenced tests. A narrower, and perhaps more accurate, meaning of the term is that a test is administered and scored in uniform and consistent fashion by all test givers and scorers for all test takers. Accordingly, researcher-developed tests, with or without a normative population, may be considered standardized if administered and scored in an explicit and consistent manner.

REFERENCES

Ahmed, Y., Francis, D.J., York, M., Fletcher, J.M., Barnes, M., & Kulesz, P. (2016). Validation of the direct and inferential mediation (DIME) model of reading comprehension in grades 7 through 12. *Contemporary Educational Psychology*, *44/45*, 68–82. https://doi.org/10.1016/j.cedpsych.2016.02.002

American Psychological Association, American Educational Research Association, & National Council on Measurement in Education. (1966). *Standards for educational and psychological tests and manuals.* Washington, DC: American Psychological Association.

Andreassen, R., & Bråten, I. (2010). Examining the prediction of reading comprehension on different multiple-choice tests. *Journal of Research in Reading*, *33*(3), 263–283. https://doi.org/10.1111/j.1467-9817.2009.01413.x

Arrington, C.N., Kulesz, P.A., Francis, D.J., Fletcher, J.M., & Barnes, M.A. (2014). The contribution of attentional control and working memory to reading comprehension and decoding. *Scientific Studies of Reading*, *18*(5), 325–346. https://doi.org/10.1080/10888438.2014.902461

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. https://doi.org/10.1177/0741932509355988

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*(2), 137–164. https://doi.org/10.1080/02702710801963951

Briggs, D.C. (2008). Comments on Slavin: Synthesizing causal inferences. *Educational Researcher*, *37*(1), 15–22. https://doi.org/10.3102/0013189X08314286

Cain, K., & Oakhill, J. (2009). Reading comprehension development from 8 to 14 years: The contribution of component skills and processes. In R.K. Wagner, C. Schatschneider, & C. Phythian-Sence (Eds.), *Beyond decoding: The behavioral and biological foundations of reading comprehension* (pp. 143–175). New York, NY: Guilford.

Cain, K., Oakhill, J., & Lemmon, K. (2004). Individual differences in the inference of word meanings from context: The influence of reading comprehension, vocabulary knowledge, and memory capacity. *Journal of Educational Psychology*, *96*(4), 671–681. https://doi.org/10.1037/0022-0663.96.4.671

Catts, H.W., & Kamhi, A.G. (2014). Prologue: Reading comprehension is not a single ability. *Language, Speech, and Hearing Services in Schools*, *45*(3), 73–76. https://doi.org/10.1044/2017_LSHSS-16-0033

Cheung, A., & Slavin, R.E. (2005). Effective reading programs for English language learners and other language-minority students. *Bilingual Research Journal*, *29*(2), 241–267. https://doi.org/10.1080/15235882.2005.10162835

Cheung, A.C., & Slavin, R.E. (2016). How methodological features affect effect sizes in education. *Educational Researcher*, *45*(5), 283–292. https://doi.org/10.3102/0013189X16656615

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*. Advance online publication. https://doi.org/10.1177/0022219420932139

Collins, A.A., Lindström, E.R., & Compton, D.L. (2018). Comparing students with and without reading difficulties on reading comprehension assessments: A meta-analysis. *Journal of Learning Disabilities*, *51*(2), 108–123. https://doi.org/10.1177/0022219417704636

Connor, C.M., Phillips, B.M., Kim, Y.-S.G., Lonigan, C.J., Kaschak, M.P., Crowe, E., … Al Otaiba, S. (2018). Examining the efficacy of targeted component interventions on language and literacy for third and fourth graders who are at risk of comprehension difficulties. *Scientific Studies of Reading*, *22*(6), 462–484. https://doi.org/10.1080/10888438.2018.1481409

Cromley, J.G., & Azevedo, R. (2007). Testing and refining the direct and inferential mediation model of reading comprehension. *Journal of Educational Psychology*, *99*(2), 311–325. https://doi.org/10.1037/0022-0663.99.2.311

Cutting, L.E., Materek, A., Cole, C.A., Levine, T.M., & Mahone, E.M. (2009). Effects of fluency, oral language, and executive function on reading comprehension performance. *Annals of Dyslexia*, *59*(1), 34–54. https://doi.org/10.1007/s11881-009-0022-0

DiCerbo, K.E., Oliver, J., Albers, C., & Blanchard, J. (2004). Effects of reducing attentional demands on performance of reading comprehension tests by third graders. *Perceptual and Motor Skills*, *98*(2), 561–574. https://doi.org/10.2466/pms.98.2.561-574

Edmonds, M.S., Vaughn, S., Wexler, J., Reutebuch, C., Cable, A., Tackett, K.K., & Schnakenberg, J.W. (2009). A synthesis of reading interventions and effects on reading comprehension outcomes for older struggling readers. *Review of Educational Research*, *79*(1), 262–300. https://doi.org/10.3102/0034654308325998

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), 761–781. https://doi.org/10.1037/edu0000180

Fletcher, J.M. (2006). Measuring reading comprehension. *Scientific Studies of Reading*, *10*(3), 323–330. https://doi.org/10.1207/s1532799xssr1003_7

Francis, D.J., Snow, C.E., August, D., Carlson, C.D., Miller, J., & Iglesias, A. (2006). Measures of reading comprehension: A latent variable analysis of the diagnostic assessment of reading comprehension. *Scientific Studies of Reading*, *10*(3), 301–322. https://doi.org/10.1207/s1532799xssr1003_6

Fuchs, D., & Fuchs, L.S. (2019). On the importance of moderator analysis in intervention research: An introduction to the special issue. *Exceptional Children*, *85*(2), 126–128. https://doi.org/10.1177/0014402918811924

Fuchs, D., Hendricks, E., Walsh, M.E., Fuchs, L.S., Gilbert, J.K., Zhang Tracy, W., … Peng, P. (2018). Evaluating a multidimensional reading comprehension program and reconsidering the lowly reputation of tests of near-transfer. *Learning Disabilities Research & Practice*, *33*(1), 11–23. https://doi.org/10.1111/ldrp.12162

Fuchs, D., Kearns, D.M., Fuchs, L.S., Elleman, A.M., Gilbert, J.K., Patton, S., … Compton, D.L. (2019). Using moderator analysis to identify the first-grade children who benefit more and less from a reading comprehension program: A step toward aptitude-by-treatment interaction. *Exceptional Children*, *85*(2), 229–247. https://doi.org/10.1177/0014402918802801

Fuchs, L.S., & Fuchs, D. (1994). Academic assessment and instrumentation. In S. Vaughn & C. Bos (Eds.), *Research issues in learning disabilities: Theory, methodology, assessment, and ethics* (pp. 233–245). New York, NY: Springer-Verlag.

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*, *84*(1), 74–111. https://doi.org/10.3102/0034654313499616

Gersten, R., Fuchs, L.S., Compton, D., Coyne, M., Greenwood, C., & Innocenti, M.S. (2005). Quality indicators for group experimental and quasi-experimental research in special education. *Exceptional Children*, *71*(2), 149–164. https://doi.org/10.1177/001440290507100202

Glass, G.V. (1977). Integrating findings: The meta-analysis of research. *Review of Research in Education*, *5*(1), 351–379. https://doi.org/10.3102/0091732X005001351

Gough, P.B., & Tunmer, W.E. (1986). Decoding, reading, and reading disability. *Remedial and Special Education*, *7*(1), 6–10. https://doi.org/10.1177/074193258600700104

Graesser, A.C., McNamara, D.S., Louwerse, M.M., & Cai, Z. (2004). Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, & Computers*, *36*(2), 193–202. https://doi.org/10.3758/BF03195564

Hebert, M., Bohaty, J.J., Nelson, J.R., & Brown, J. (2016). The effects of text structure instruction on expository reading comprehension: A meta-analysis. *Journal of Educational Psychology*, *108*(5), 609–629. https://doi.org/10.1037/edu0000082

Institute of Education Sciences & National Science Foundation. (2013). *Common guidelines for education research and development*. Retrieved from https://ies.ed.gov/pdf/CommonGuidelines.pdf

Jones, S.M., LaRusso, M., Kim, J., Kim, H.Y., Selman, R., Uccelli, P., … Snow, C. (2019). Experimental effects of Word Generation on vocabulary, academic language, perspective taking, and reading comprehension in high-poverty schools. *Journal of Research on Educational Effectiveness*, *12*(3), 448–483. https://doi.org/10.1080/19345747.2019.1615155

Kamhi, A.G., & Catts, H.W. (2017). Epilogue: Reading comprehension is not a single ability—implications for assessment and instruction. *Language, Speech, and Hearing Services in Schools*, *48*(2), 104–107. https://doi.org/10.1044/2017_LSHSS-16-0049

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*(3), 281–300. https://doi.org/10.1080/10888430802132279

Keenan, J.M., & Meenan, C.E. (2014). Test differences in diagnosing reading comprehension deficits. *Journal of Learning Disabilities*, *47*(2), 125–135. https://doi.org/10.1177/0022219412439326

Kieffer, M.J., Vukovic, R.K., & Berry, D. (2013). Roles of attention shifting and inhibitory control in fourth-grade reading comprehension. *Reading Research Quarterly*, *48*(4), 333–348. https://doi.org/10.1002/rrq.54

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. https://doi.org/10.1111/j.1540-5826.2012.00352.x

McNamara, D.S., & Kendeou, P. (2017). Translating advances in reading comprehension research to educational practice. *International Electronic Journal of Elementary Education*, *4*(1), 33–46.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, *50*(9), 741–749. https://doi.org/10.1037/0003-066X.50.9.741

Nation, K. (2009). Connections between language and reading in children with poor reading comprehension. In H.W. Catts & A.G. Kamhi (Eds.), *The connections between language and reading disabilities* (pp. 37–47). New York, NY: Psychology. (Original work published 2005)

Nation, K., & Snowling, M. (1997). Assessing reading difficulties: The validity and utility of current measures of reading skill. *British Journal of Educational Psychology*, *67*(3), 359–370. https://doi.org/10.1111/j.2044-8279.1997.tb01250.x

Patton, S., Fuchs, D., Hendricks, E., Pennell, A., Walsh, M., Fuchs, L., Zhang Tracy, W., & Haga, L. (2020). *Explicit instruction of transfer to improve at-risk intermediate-grade students' understanding of informational texts*. Manuscript submitted for publication.

Peng, P., Barnes, M., Wang, C., Wang, W., Li, S., Swanson, H.L., … Tao, S. (2018). A meta-analysis on the relation between reading and working memory. *Psychological Bulletin*, *144*(1), 48–76. https://doi.org/10.1037/bul0000124

Perfetti, C.A. (1985). *Reading ability*. New York, NY: Oxford University Press.

Perfetti, C., & Stafura, J. (2014). Word knowledge in a theory of reading comprehension. *Scientific Studies of Reading*, *18*(1), 22–37. https://doi.org/10.1080/10888438.2013.827687

RAND Reading Study Group. (2002). *Reading for understanding: Toward an R&D program in reading comprehension*. Santa Monica, CA: RAND.

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. https://doi.org/10.3102/00346 54316652942

Scammacca, N.K., Roberts, G., Vaughn, S., & Stuebing, K.K. (2015). A meta-analysis of interventions for struggling readers in grades 4–12: 1980–2011. *Journal of Learning Disabilities*, *48*(4), 369–390. https://doi.org/10.1177/0022219413504995

Schoenfeld, A.H. (2006). What doesn't work: The challenge and failure of the What Works Clearinghouse to conduct meaningful reviews of studies of mathematics curricula. *Educational Researcher*, *35*(2), 13–21. https://doi.org/10.3102/0013189X035002013

Slavin, R.E. (1986). Best-evidence synthesis: An alternative to meta-analytic and traditional reviews. *Educational Researcher*, *15*(9), 5–11. https://doi.org/10.3102/0013189X015009005

Slavin, R.E. (1995). Best evidence synthesis: An intelligent alternative to meta-analysis. *Journal of Clinical Epidemiology*, *48*(1), 9–18. https://doi.org/10.1016/0895-4356(94)00097-A

Slavin, R. (2019, October 24). Developer- and research-made measures [Web log post]. Retrieved from https://robertslavinsblog.wordpress.com/2019/10/24/developer-and-researcher-made-measures/

Slavin, R.E., Cheung, A., Groff, C., & Lake, C. (2008). Effective reading programs for middle and high schools: A best-evidence synthesis. *Reading Research Quarterly*, *43*(3), 290–322. https://doi.org/10.1598/RRQ.43.3.4

Slavin, R.E., Lake, C., Chambers, B., Cheung, A., & Davis, S. (2009). Effective reading programs for the elementary grades: A best-evidence synthesis. *Review of Educational Research*, *79*(4), 1391–1466. https://doi.org/10.3102/0034654309341374

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. https://doi.org/10.1177/0022219411402691

Solis, M., Vaughn, S., Stillman-Spisak, S.J., & Cho, E. (2018). Effects of reading comprehension and vocabulary intervention on comprehension-related outcomes for ninth graders with low reading comprehension. *Reading & Writing Quarterly*, *34*(6), 537–553. https://doi.org/10.1080/10573569.2018.1499059

Spear-Swerling, L. (2004). Fourth graders' performance on a state-mandated assessment involving two different measures of reading comprehension. *Reading Psychology*, *25*(2), 121–148. https://doi.org/10.1080/02702710490435727

Stevens, E.A., Park, S., & Vaughn, S. (2019). A review of summarizing and main idea interventions for struggling readers in grades 3 through 12: 1978–2016. *Remedial and Special Education*, *40*(3), 131–149. https://doi.org/10.1177/0741932517749940

Stevens, E.A., Vaughn, S., Swanson, E., & Scammacca, N. (2020). Examining the effects of a Tier 2 reading comprehension intervention aligned to Tier 1 instruction for fourth-grade struggling readers. *Exceptional Children*, *86*(4), 430–448. https://doi.org/10.1177/0014402919893710

Stice, E., Presnell, K., Gau, J., & Shaw, H. (2007). Testing mediators of intervention effects in randomized controlled trials: An evaluation of two eating disorder prevention programs. *Journal of Consulting and Clinical Psychology*, *75*(1), 20–32. https://doi.org/10.1037/0022-006X.75.1.20

Swanson, E., Wanzek, J., Vaughn, S., Fall, A.M., Roberts, G., Hall, C., & Miller, V.L. (2017). Middle school reading comprehension and content learning intervention for below-average readers. *Reading & Writing Quarterly*, *33*(1), 37–53. https://doi.org/10.1080/10573569.2015.1072068

Tighe, E.L., & Schatschneider, C. (2014). A dominance analysis approach to determining predictor importance in third, seventh, and tenth grade reading comprehension skills. *Reading and Writing*, *27*(1), 101–127. https://doi.org/10.1007/s11145-013-9435-6

van den Broek, P., White, M.J., Kendeou, P., & Carlson, S. (2009). Reading between the lines: Developmental and individual differences in cognitive processes in reading comprehension. In K. Wagner, C. Schatschneider, & C. Plythian-Sence (Eds.), *Beyond decoding: The behavioral and biological foundations of reading comprehension* (pp. 107–123). New York, NY: Guilford.

Vaughn, S., Martinez, L.R., Williams, K.J., Miciak, J., Fall, A.M., & Roberts, G. (2019). Efficacy of a high school extensive reading intervention for English learners with reading difficulties. *Journal of Educational Psychology*, *111*(3), 373–386. https://doi.org/10.1037/edu0000289

Vavassoeur, L.C. (2016). *Predictive power of the Test of Everyday Attention for Children (TEA-CH) on various methods of reading comprehension assessment among low-income fourth grade children of Color* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 10238172)

Verhoeven, L., & van Leeuwe, J. (2008). Prediction of the development of reading comprehension: A longitudinal study. *Applied Cognitive Psychology*, *22*(3), 407–423. https://doi.org/10.1002/acp.1414

What Works Clearinghouse. (2020). *What Works Clearinghouse standards handbook, version 4.1*. Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Wixson, K.K. (2017). An interactive view of reading comprehension: Implications for assessment. *Language, Speech, and Hearing Services in Schools*, *48*(2), 77–83. https://doi.org/10.1044/2017_LSHSS-16-0030

Submitted October 6, 2020
Final revision received April 8, 2021
Accepted April 9, 2021

NATHAN H. CLEMENS (corresponding author) is an associate professor and Dean's Distinguished Faculty Fellow in the Department of Special Eduction at The University of Texas at Austin, USA; email nathan.clemens@austin.utexas.edu.

DOUGLAS FUCHS (corresponding author) is a professor and the Nicholas Hobbs Chair in Special Education and Human Development in the Department of Special Education at Peabody College, Vanderbilt University, Nashville, Tennessee, USA; email doug.fuchs@vanderbilt.edu.

APPENDIX

Characteristics of Commercial Tests Commonly Used in Reading Comprehension Intervention Research

			GRADE (Reading Comprehension Scale)						
Feature	**GORT–5**	**GMRT–RC**	**Passage**	**Sentence**	**SAT–10**	**TOSREC**	**WJ Passage Comprehension**	**WIAT–4**	**WRAT–4**
Oral or silent reading	Oral	Silent	Silent	Silent	Silent	Silent	Silent	Silent	Silent
Question type/ response mode	Open-ended (verbal response)	Multiple-choice	Multiple-choice	Cloze (multiple-choice)	Multiple-choice	Sentence verification	Cloze (verbal response)	Open-ended (verbal response)	Cloze (verbal response)
Sentence, single paragraph, or multiparagraphs	Single paragraph and multiparagraph	Single paragraph and multiparagraph	Single paragraph and multiparagraph	Sentence	Single paragraph and multiparagraph	Sentence	1 or 2 sentences	Single paragraph and multiparagraph	Sentence
Narrative or informational passages	Narrative	Mixed	Mixed		Mixed		Mixed	Mixed	
Answer choices require reading	No	Yes	Yes	Yes	Yes	No	No	No	No
Time limit	No	35 minutes	No	No	No	3 minutes	No	No	No

Note. Characteristics reflect test versions applicable for students in grades 4–8. GMRT–RC = Gates–MacGinitie Reading Tests–Reading Comprehension (4th edition); GORT–5 = Gray Oral Reading Test (5th edition); GRADE = Group Reading Assessment and Diagnostic Evaluation; SAT–10 = Stanford Achievement Test (10th edition); TOSREC = Test of Silent Reading Efficiency and Comprehension; WIAT–4 = Weschler Individual Achievement Test (4th edition); WJ = Woodcock–Johnson Tests of Achievement; WRAT–4 = Wide Range Achievement Test (4th edition).