[In Press citation: Arya, D.J., Clairmont, A., Katz, D. & Maul, A. (in press). Measuring reading strategy use [Special issue on validity studies]. *Educational Assessment.*]

Abstract

This study describes the development and validation of a multidimensional measure of preadolescent and adolescent readers' abilities to apply reading comprehension strategies necessary for understanding challenging academic texts. The Strategy Use Measure (SUM) was designed with the intention of being pedagogically informative to the increasingly multilingual student population in the U.S. in grades 6 through 8. The SUM aims to measure four areas of knowledge and skill that are widely purported to support the use of reading strategies: (a) morphological awareness, (b) knowledge of cognates, (c) ability to relate micro- and macro- ideas within a text, and (d) the ability to use intra- and inter-sentential context clues for defining unfamiliar words. The test was developed following a principled, iterative process to instrument development, employing Rasch models and qualitative investigations to test hypotheses related to the instrument's validity. Findings suggest promising evidence for the validity and fairness of this multidimensional measure.

Keywords. Culturally responsive assessment, reading strategies, reading assessment, Rasch measurement, validity, mixed methods

Introduction

It would be difficult to overstate the importance of the ability to read for nearly every aspect of modern life. The development of reading comprehension ability is widely recognized to be one of the primary goals of formal education, and it is small wonder, then, that its assessment is also recognized to be of critical importance. Measures of reading comprehension ability and related constructs are ubiquitous at every level of education, and are used in many contexts for a wide range of purposes.

However, despite (or, perhaps, at least in part because of) the recognized importance of reading comprehension ability and the continued proliferation of tests that claim to assess it, there remains a great deal of confusion and controversy regarding reading comprehension ability and its assessment. In the academic literature, this is evidenced by the continued lack of a clear, consensus definition of reading comprehension ability itself (Adams, 2001; Perfetti, & Stafura, 2014). The ever-changing nature of literacy surely plays a role in this: as societies become ever more multicultural and multilingual, as new technologies emerge, and as the demands of college, careers and life in general continue to evolve, what is or should be denoted by terms like “literacy” and “reading comprehension” will likely remain a moving target.

But apart from the continued need for theoretical and definitional clarity, there are also issues of direct practical relevance to educators in greater need of attention. In particular, the Common Core Standards in English Language Arts (i.e., CCSS ELA-Literacy/L/Grades 2-12) emphasize the use of *reading strategies* – activities consciously employed for facilitating comprehension of texts – especially when confronting challenging texts and when learning to read (Dewitz, Jones & Leahy, 2009; Reynolds &

Daniel, 2017). The importance of strategy use for reading complex texts (i.e., texts that are difficult to read because of the conceptual complexity of the subject matter and/or the presence of less commonly-used words and embedded clauses that link ideas together in single sentences) has been acknowledged by many literacy educators and scholars (e.g., Brown, Armbruster & Baker, 1986; Cantrell, Almasi, Carter, Rintamaa & Madden, 2010; Cho, 2013; Guthrie et al., 1998). In order to effectively promote the use of such strategies, educators require ready access to information not just about their students' overall abilities to comprehend what they read, but more specific and fine-grained information about the strategies they are able to employ and those that may need further support through explicit instruction.

However, there is at present a dearth of instruments capable of providing information about reading strategy use. Again, this state of affairs may be related to a larger lack of definitional clarity, as a theoretical definition of strategy use continues to evade full consensus in the literacy field (see, e.g., Afflerbach, Pearson & Paris, 2008). Performance-based measures face the difficulty of disentangling the many cognitive processes employed during the act of reading and understanding texts. A common mode of assessment in reading (including in purported measures of strategy use; Cromley & Azevedo, 2011) is to ask the respondent to read a passage and answer a series of questions about that passage, but given item response patterns alone it is difficult to determine which strategies were employed during the generation of these responses. For example, the extent to which a respondent reviews contextual clues before selecting the correct definition for an underlined word would depend on their prior familiarity with the underlined word, for which there is in general no independent evidence. Likewise, since strategies tend to be consciously employed only in the presence of a challenge to comprehension, it is possible that texts below the ability level of the learner do not require explicit strategies for sense-making at all. The common alternative to performance-based measurement is self report, usually via a survey (e.g., self-reported use of context clues for determining the meaning of unknown words; Wigfield & Guthrie, 2010), but the limitations of survey-based self-reports of skills are well-known; as expressed by Mokhtari and Reichard (2002), "one cannot tell from the instrument alone whether students actually engage in the strategies they report using" (p. 254).

Thus, there is a clear need for the development of instruments that can provide educators and other stakeholders with pedagogically relevant information about students' funds of knowledge and skills that support the use of reading strategies. The purpose of this paper is to introduce a new instrument built for this purpose, the Strategy Use Measure (SUM), and to present and evaluate an evidence-based argument (AERA, APA, NCME, 2014; Kane, 2006) for its validity.

Measuring Reading Strategy Use

When students confront challenging academic texts, strategies such as looking for word-level and sentential contextual clues are essential for understanding textual content (Afflerbach, Pearson & Paris, 2008; Brown & Palincsar, 1982; Wixson & Peters, 1987). As such, an actionable measure of reading strategy use should be able to provide information about a variety of skills, such as the use of prior knowledge (Wixson & Peters, 1987), previewing the text (Janzen & Stoller, 1998), the use of organizational text features for summarizing main ideas (Baker & Brown, 1984; Brown & Palincsar, 1982; Paris & Oka, 1989), the use of word-level knowledge such as morphological awareness and known cognates (Jimenez, Garcia & Pearson, 1995; Snow, August & White, 2011), and knowing when to ignore or reread a particular passage (Collins & Smith, 1980).

In recent years literacy scholars have learned much about the nature of such reading strategies and the kinds of pedagogical moves that can support their development through research programs such as Collaborative Strategic Reading (CSR; Klingner & Vaughn, 1999; 2000). CSR is a collaborative reading program originally designed for students in grades 4-12 that incorporates high-level talk and reading strategy use into an instructional approach that integrates the facilitation of reading strategies instruction and cooperative learning into a single lesson. In CSR, as students read a shared text in small groups and work together to learn the meaning of unfamiliar words and identify key ideas. Four main classes of reading strategies are explicitly taught and encouraged during group reading: a) breaking down unfamiliar words into morphemes, b) accessing cognate knowledge, c) reviewing potential contextual clues for unfamiliar words, and d) relating primary, secondary and tertiary ideas to one another.

The Strategy Use Measure (SUM) was originally designed in the context of CSR to measure students' command of these strategies, but was also intended to serve as a non-curriculum-specific measure of reading strategy use, in alignment with Common Core standards that emphasize strategy use during reading in collaborative contexts (Dewitz, Jones & Leahy, 2009; Reynolds & Daniel, 2017). The SUM was designed to be useful primarily in pedagogical (formative) contexts. The SUM assesses a student's command of four areas of knowledge and skill: (a) use of word-level knowledge such as morphological awareness (i.e., knowledge about the meaning of affixes in words; MA), (b) knowledge about English/Spanish cognates (words that are shared across two or more languages; COG), (c) surrounding context clues (the use of contextual clues within and between sentences for defining unfamiliar words; CC), and (d) use of text-based skills such as identifying micro- and macro- relationships in text (identifying primary, secondary, and tertiary ideas within a passage; MMRT). Such knowledge and skills directly support the application of strategies, which are especially useful when one is unable to automatically understand textual content (Afflerbach, Pearson & Paris, 2008).

Additionally, literacy researchers and educational psychologists have observed that some portions of the student population may overly rely on the gestalt (whole-word) approach to reading as a way of identifying words in text (e.g., Jordan, 1994). Such an approach attends to the entire shape and length of a familiar word, and is helpful for sight words that cannot be read accurately through sound-print connections (e.g., "might"). Children who overly rely on this approach may struggle with recognizing embedded morphemes within words (Carlisle, 2004), and those who struggle with such lexical analysis may also focus more on contextual clues to support textual comprehension (NRC, 1998). Such patterns thus involve strength in one kind of strategy (e.g., the use of contextual clues) but a relative weakness in another (e.g., morphological awareness).

The following sections gives a more detailed description of each of these areas.

Morphological awareness (MA). One of the biggest sources of difficulty for children engaged in school-based texts is the frequency of unfamiliar, conceptually complex words and terms (Author, 2011; Stahl, Jacobson, Davis & Davis, 1989). As students approach later elementary and middle school grades, the likelihood of encountering highly specialized, multisyllabic words increases. Literacy scholars have found that explicit instruction of morphological meanings (i.e., meanings of the smallest units of a word) can support engagement and understanding of challenging texts (e.g., Arnbak & Elbro, 2000; Berninger, Abbot, Nagy & Carlisle, 2007). For example, teaching children to see a word like "bioluminescence" as having word parts that can be featured in a myriad of other words may have a generative benefit for English language development and text comprehension. The prefix *bio* carries the meaning "life"; knowledge of this meaning, along with the featured root "lumen" that refers to the commonly known word "light", offers a young reader the opportunity to make sense of the entire word's meaning. Morphemes provide readers with familiar clues by which new vocabulary is acquired. As children develop experience in unpacking multisyllabic words and reading morphemes, challenging texts become far easier to comprehend. Instructional support in such experiences has been found to boost more than MA performance; in their systematic review of 22 intervention studies focused on explicit MA instruction, Bowers, Kirby and Deacon (2010) found that such intervention had a positive effect on reading comprehension, spelling and vocabulary for children in elementary and middle school.

Cognate knowledge (COG). The English language is filled with cognates, which are words that are shared among two or more languages. For instance, words like "problem" have close counterparts in languages such as Spanish ("problema"). Some scholars who have focused on multilingual learners--that is, students who speak languages other than English at home--view student knowledge of cognates as valuable for utilization during reading and discussion (Lubliner & Grisham, 2012; Montelongo, Hernández, Herter, & Cuello, 2011). Multilingual learners bring to the classroom a wealth of cultural and linguistic knowledge that can be leveraged for developing deeper contextual understandings about cognates embedded in academic English texts. However, learning to recognize cognates often requires explicit instruction and encouragement (Jiménez, 1997; Montelongo, Hernández, Herter, & Cuello, 2011), as well as some tacit morphological knowledge. The more multilingual learners are encouraged to use their home language in reading contexts, the more connections they are able to make across English and other linguistic systems, and the greater their ability to pick up lexical clues for comprehending (Rodríguez, 2001).

Contextual clues (CC). In addition to dissecting the unfamiliar word itself in terms of its morphemes or its similarity with words from other known languages, readers can get hints of meaning from surrounding words and sentences. For example, consider the following consecutive sentences: "Fireflies naturally produce light. We call this type of light bioluminescence. Bioluminescence does not give off much heat. This light produced by fireflies is used to communicate with each other and to find mates." These sentences might be found in a chapter of a biology textbook. Readers who do not understand the word "bioluminescence" could figure out its meaning in more than one way. In addition to unpacking the distinct morphemes within the word or referring to similar words in other known languages like Spanish ("bioluminiscencia"), readers can look for clues surrounding the word in question. Fireflies are common organisms and as such, provide a clue about the main point of the text. Knowledge about the unique behaviors of fireflies can be coupled with the strong likelihood of understanding the meaning of "light" to gain understanding of the unknown word. Using contextual clues surrounding the unknown word or phrase has been long acknowledged as an important strategy for reading comprehension and vocabulary development (Artley, 1943; Blackowicz & Fisher, 2000; Sáenz & Fuchs, 2002).

Micro and macro relationships in text (MMRT). Graphic organizers and conceptual maps have been found to support reading comprehension and text summarization (e.g., Carr & Ogle, 1987; Chang, Sung & Chen, 2002; Duke & Pearson, 2009). Texts provide information in varying degrees of importance; some ideas or propositions in text may have greater relevancy over others, and understanding this relational dynamic is critical for capturing a text-based understanding during reading (Kintsch & van Dijk, 1978; Kucan, Palicsar, Busse & Heisey, 2011). Using the text related to bioluminescence from the previous section, one can identify multiple propositions (discrete ideas); the most prominent topic is the organism (firefly) and it's unique abilities to give off light (bioluminescence). Secondarily, the reader understands that very little heat is emitted from fireflies when they light up, a supporting detail. The skill of mapping key ideas as they are represented in text has been long found to support overall text comprehension and conceptual understanding (Armbruster, Anderson & Meyer, 1991; Oliver, 2009; Robinson & Flores, 1997). For example, in a study of 74 sixth graders, Oliver (2009) found that mapping key concepts represented in science texts supported greater understanding of key terms and conceptual processes. Thus, teaching the use of such graphic tools such as conceptual maps for organizing textual information can be helpful for making sense of challenging academic texts.

Each of these areas of knowledge and skill relate to a particular kind of textual information. Two relate to information within words: MA items involve information provided in affixes, while COG items involve information provided in similar words from another language (Spanish). The other two relate to contextual information provided within and across sentences within a text: ICC items involve information that provides additional clues concerning ideas and concepts, while MMRT items involve information that signals the relative prominence of key ideas presented in text. As such, the SUM dimensions target strategies at both the lexical and contextual levels.

The intended interpretation and use of the SUM

The *Standards for Educational and Psychological Testing* (AERA, APA, NCME, 2014) represents a consensus statement by the participating organizations of the expectations for the development and use of educational tests, and emphasizes the need to provide an evidence-based argument for validity, including precision of measurement (reliability). The most recent edition of the *Standards* also foregrounds the issue of fairness – or “responsiveness to individual characteristics and test contexts so that test scores will yield the same meaning for intended uses for all test takers” (p. 50) – as a fundamental validity issue.

The *Standards* and other contemporary accounts of validity and validation emphasize the need to begin with a clear statement of the intended interpretation(s) and use(s) of an assessment. In the present case, the aspirational *interpretation* claim is as follows: when the SUM is administered to students under appropriate conditions, the four resulting scale scores[1] can be interpreted as measures of four areas of knowledge and skill that support the use of reading strategies. The aspirational *use* claim is as follows: when the SUM is administered to students under appropriate conditions, the results of the SUM can be used for pedagogical purposes, to help guide teachers' decision-making that helps support students' ability to comprehend challenging texts. As will be described in further detail, the evidence available at present is not yet sufficient to conclusively support either of these claims (thus their description as "aspirational"). However, we interpret preliminary evidence in support of the interpretation claim to be encouraging, and suggestive of further directions for research and development; there is as yet little independent direct evidence regarding the use claim, and so this represents a main priority for further study. In the following two sections we specify the intended interpretation and use claims in more detail.

The interpretation claim. The claim that the four SUM scale scores (i.e., person estimates from the multidimensional Rasch model described in more detail below) can be interpreted as measuring four areas of knowledge and skill that support the use of reading strategies implies two specific claims in need of evaluation. Each of these claims, in turn, implies a number of sub-claims and finer-grained research questions.

Subclaim 1: The production of students' (correct or incorrect) responses to individual SUM items involves the use of construct-relevant cognitive response processes – that is, the utilization of forms of knowledge and skills related to the kinds

[1] Claims are expressed here in terms of scale scores (i.e., person estimates from the multidimensional Rasch model described in more detail in later sections) rather than raw scores. Development of a computerized adapted test (CAT) for the SUM is currently underway using Concerto (Scalise & Allen, 2015), a free and open-source testing platform that interfaces with the catR library. Ultimately, the intention is that teachers will have access to a free, online version of the SUM that estimates and presents such scale scores along with confidence bands expressed in relation to construct maps, along with guidance documents and sample items to aid interpretation.

of strategy use intended to be assessed by the SUM – and minimal construct-irrelevant processes, including but not limited to those that would give an unfair advantage to individuals from different cultural or linguistic backgrounds. In the present study, this claim is supported (a) theoretically, via the articulation of the construct maps and item design strategies that guided test development, (b) qualitatively, via item paneling with multilingual content experts and cognitive interviews with both middle school- and college-aged students, and (c) quantitatively, via examination of item and model fit statistics and tests of Differential Item Functioning (DIF) by language background. Additionally, follow-up cognitive interviews and linguistic analyses helped clarify the patterns of item responses among Spanish speakers. In the language of the *Standards*, such analyses primarily involve validity evidence based on test content, examinee response processes, and the internal structure of the test.

Subclaim 2*:* When SUM items are administered appropriately to students from the intended population, the four resulting scores can be used to draw inferences about students' levels of knowledge and skill in each of the four areas described previously; further, these inferences can be expressed in quantitative terms (i.e., in terms of magnitudes). This claim is evaluated quantitatively by evaluation of item fit and model fit of a multidimensional Rasch model, and estimation of measurement precision (i.e., reliability and standard errors of measurement) for each dimension. In the language of the *Standards*, this primarily involves validity evidence based on test content and internal structure.

In addition to evidence directly targeted toward the evaluation of these two primary claims, associations were investigated between the SUM and a measure of reading comprehension, aimed at helping establish the relevance of SUM results to general reading instruction (referred to in the *Standards* as validity evidence based on relations to other variables).

The use claim. The intention of the SUM is not to replace existing summative tests of reading comprehension ability, but to help provide targeted information to teachers that can in turn suggest optimal instructional moves to support the development of areas of knowledge and skill that can ultimately contribute to reading comprehension ability. Additionally, the SUM may be used for general research purposes, such as in the exploration of pedagogical strategies that promote the development of students' competence in applying reading strategies to challenging texts, including in the context of the aforementioned CSR program.

The mere fact that the SUM provides information on four areas of knowledge and skill related to reading strategy use itself reinforces the aforementioned notion that reading comprehension ability is not a unitary skill, nor does it represent a single type of knowledge. Different students have different funds of knowledge (i.e., prior conceptual and linguistic knowledge, prior experiences, etc.) that can be leveraged to support reading comprehension, and the design of the SUM is consistent with the idea that students benefit from instruction that targets particular needs and reinforces existing strengths. Ultimately, we hope that one major way in which SUM results could be utilized is in informing the composition of small groups of students when reading collaboratively; by composing heterogeneous groups of students with complementary strengths, for example, teachers can provide students with peer support that in turn fosters collaborative learning in the classroom, thus effectively increasing the number of teachers during collaborative reading.

Given that the use of the SUM for these purposes is underwritten by the claim that the SUM scores measure what they claim to measure, all forms of evidence that support the interpretation claim are relevant to the use claim as well. In addition, preliminary evidence directly relevant to the use claim is provided by results from a collaboration with a small group of local teachers, who, with the assistance of our research team, administered the SUM and made use of its results in their classrooms, and then provided feedback regarding its utility for pedagogical applications (described further under "Study of pedagogical applications" below).

Methods

The SUM was created using an iterative, construct-centered approach to instrument development, loosely based on the Berkeley Evaluation and Assessment Research (BEAR) Assessment System (BAS; Wilson, 2005). The BAS involves iterating between phases of articulation and refinement of the model of the target constructs (*construct mapping*), the development of items and scoring guides aligned with the theories of the constructs (*items design*), checking interpretations and thought processes involved in item responses via expert feedback and cognitive interviews (exploring the *outcome space*), and psychometric investigations into the quality of the items and the consistency of student response patterns with expectations derived from theory (*measurement modeling*). The instrument development process spanned five years and involved two pilot studies with a total of 4,455 students in grades 6 through 8 attending schools within an urban district in Colorado. Each phase of this development process is described below.

Initial construct mapping and items design

Construct maps were devised to describe the nature of progressively more sophisticated levels of understanding and skill in each of the four areas described previously (morphological awareness, cognate knowledge, contextual clues, and micro and macro relationships in text), drawing from existing research and theoretical work described earlier. Figures 1 through 4 below presents the MA construct map as an example.

[Insert Figures 1-4 here]

The purpose of developing construct maps was (a) to help clarify the nature of variation in each of the four constructs, and (b) to help guide item generation, by specifying initial hypotheses about what sorts of items should be able to discriminate among individuals at lower and higher locations along each construct map.

These maps then served as initial guides for the generation of test items, each intended to elicit student behaviors that could be used to draw inferences about whether a student had mastered the knowledge and skills up to and including a given point along the construct map. All items were forced-choice (requiring the respondent to select from a series of options) and were scored dichotomously (as correct or incorrect).

The use of nonsense words for the bulk of MA and CC items helped to mitigate the use of prior knowledge; respondents would need to know the morpheme –s, for example, in order to select the plural form of *zowt*. Uses of nonsense words for reading-related assessments is a common practice for avoiding non-relevant construct knowledge (i.e., knowing the meaning of a vocabulary word within a given text precludes the need for using contextual clues; see Anderson & Freebody, 1982). Other typical assessment practices used in item development include the uses of formatting that fosters attentional gaze on key points (e.g., *select the item that is **least related*** . . .), ease of reading directions by avoiding relatively complex terminology, item stems and response choices while including a sufficient amount of context to support understanding of the task (Pearson, Hiebert & Kamil, 2007; Palincsar, Magnusson, Pesko & Hamlin, 2005). The number of items for each dimension was largely determined by the amount of time it took to respond to each item during initial cognitive interviews (e.g., CC involved lengthier texts, thus limiting the number of items accordingly) so that students could complete the SUM within a 50-minute period.

Initial item paneling and cognitive interviews

A panel of five scholars with expertise in bilingual education and assessment practices, three of whom were native Spanish speakers, reviewed each item for clarity and potential linguistic or cultural bias. Each panelist was also asked to reflect on the Spanish words featured on the cognate items and assess the extent to which they would be considered accessible by Spanish-speaking middle school students. Such consultation is increasingly necessary given the dramatic increase in cross-national communication made possible via the Internet, which has brought forth shared modern terminology, like the word "hacker," which is used in both Spanish and English. Moreover, given that language and culture change continually, consultation with native speakers is needed to avoid Spanish cognates that are rarely used or obsolete. Commentary and suggestions from item panels informed revisions to construct maps and associated items.

A series of individual cognitive interviews (Desimone & Le Floch, 2004; Jimenez, 1997) were then conducted with 15 students in grades 6 through 8, of whom 7 were female and 8 male, and of whom 5 were monolingual English-only speakers, 5 were Spanish dominant speakers, and 5 were speakers of languages other than English or Spanish. In addition to checking the amount of time respondents used to respond to items, these interviews were used to identify ambiguously-worded items and directions, and to investigate the extent to which the cognitive processes actually employed by the participants conformed to the expectations of the construct map for each dimension. Another round of revisions was then made to construct maps and items. Such interviews (also often referred to as think-alouds) were informative for checking whether students will interpret the item as intended. These interviews were conducted both in near final and final form; rounds of interviews were conducted following initial draft and each subsequent revision of the instrument.

All rounds of cognitive interviews helped in clarifying potential misinterpretations as well as unintended plausible responses. The following excerpted item from the CC dimension, for example, was found to have more than one plausible correct response:

The Galilean telescope, which is one of the many types used today, is a ***gorp*** *with two round glass lenses near each end.*

What could ***gorp*** *mean? (Choices: telescope;* ***tube****; camera)*

The intended correct response (tube) was not always deemed the most logical choice by interviewees who focused on the main topic of the item and assumed that telescopes indeed tend to have two round glass lenses on either end. In total, 88 items across the four dimensions made it through the initial review process and were included on the version of the SUM piloted in Study 1 (described below). Figure 5 below presents example items from each of the four dimensions.

Additionally, we conducted a series of cognitive interviews with seven undergraduate students, all of whom are Spanish-English bilingual speakers. Using such an older group of students helped us in clarifying response processes that may be more challenging for children to articulate.

[Insert Figure 5 here]

Study 1

Following the cognitive interviews and subsequent item revision, the revised set of items was administered online to 1,359 students (657 females, 708 males) attending one of two urban middle schools in northern Colorado. These students were in a control group of an early evaluation of the CSR program. Items were created and administered using an online, survey and assessment software program. Students had a single 50-minute class period to complete all items.

In addition to the SUM, students also completed the *Gates MacGinitie Reading Test* (on a different day within the same month). This 35-minute assessment contains a series of passages followed by items that require the respondent to draw inferences or to directly recall information presented in the passage. The GM is an established reading comprehension assessment that has been widely used for research and educational purposes (e.g., Cooter & Curry, 1989; Linn & Valencia, 1986; Slavin, Lake, Davis, Madden, 2011).

Item response patterns on the SUM were analyzed using multidimensional Rasch models (described in more detail below), along with examination of classical item statistics including reliability coefficients and item-total correlations. Results revealed low reliability estimates for the cognate (COG) and morphological awareness (MA) dimensions, which originally involved binary (yes/no; good/bad) choices, suggesting that items' discriminatory power was compromised by this response format. Further, the contextual clues (CC) dimension revealed a number of poorly-fitting items, and follow-up content analyses suggested that in many cases multiple answers could be considered correct depending on a respondent's interpretation of the item.

An additional round of revisions was made to the construct maps and existing items, and an additional set of new items was generated, leading to a total increase of 77 items across all dimensions. A second round of individual cognitive interviews was then conducted with 10 students, of whom 5 were female and 5 male, and of whom 5 were English-only speakers, 3 were Spanish dominant speakers and 2 were speakers of languages other than English or Spanish, and none of whom had been involved in any of the previous studies. The revised set of items, including both items retained and revised from the initial item set and the 77 new items, were evaluated for coherence and accessibility. A number of additional edits were made after this process, and 6 items were removed, leaving a total of 159 items.

Study 2

From the newly-revised item set, four counterbalanced test forms were created, intended to be used as pretests and posttests in the context of a second wave of evaluation of the CSR program. Each form contained 78 or 79 items. A total of 3,096 students either received form A at pretest and form B at posttest, or form C at pretest and form D at posttest (where in both cases, there were 36 overlapping items; all test forms are linked due to the joint calibration of all items).[2] Within each form, the order in which item blocks (i.e., each of the four areas of the SUM), items within blocks, and response options for each item were presented to each respondent was randomized. Information about home language use (in particular, about the primary language students used at home, number of books read in another language, and access to Spanish instruction) was collected via self-report. 52.8% (n=1,621) of the sample answered that they primarily spoke English at home, while 26.8% (n=822) of the sample answered that they primarily spoke Spanish at home. Approximately 15% of students either skipped this question or answered “other.” Speakers of languages not described above composed very small portions of the sample (see Table 1).

[2] The forms were designed such that each contained items encompassing the full range of expected difficulties.

Data from the pretest administration of these forms was utilized for the investigations of item fit, dimensionality, and differential item functioning as described in the following section. Following these analyses, two items were deleted, leaving a total of 157 items, of which 51 targeted morphological awareness, 42 targeted cognates, 33 targeted micro and marco relationships, and 31 targeted contextual clues.

Study of pedagogical applications

A revised version of the SUM was administered to students in two additional classrooms in collaboration with three teachers to gather feedback about the utility of the SUM in classroom applications. Researchers met with these teachers and interpreted SUM results for individual students from each of the classrooms. During these meetings, teachers directed the process of focusing on particular students of interest, while researchers used test results to offer informed explanations of observed performances. Data from these meetings were evaluated qualitatively, as detailed further below.

Analysis plan

The analysis plan was designed to test the key claims embedded in the validity argument, as discussed previously. Given the scope of the work, the analyses discussed here are meant to exemplify the approach taken rather than being exhaustive, and focus on the evidence relevant to the current version of the SUM, rather than the evidence that was used to guide successive revisions.

As discussed previously, it was hypothesized that the responses to each of the four sets of items included in the SUM could be modeled as measuring empirically distinguishable quantities, with individual items providing discriminating information about individuals at pre-identified locations along these quantities, as expressed in the construct maps. These hypotheses were jointly tested by examination of the fit of students' item response patterns to a series of unidimensional (Rasch, 1960/1980) and multidimensional (Adams, Wilson, & Wang, 1997) Rasch models, and inspection of item difficulty estimates.

Examination of overall model fit (i.e., the extent to which observed response patterns conformed to what would have been expected according to the model) provides evidence relevant to the hypotheses regarding the modelability of the dimensions as quantities. Inspection of correlations (disattenuated for measurement error) between dimensions and comparative tests of overall model fit between unidimensional and multidimensional models provide evidence relevant to the hypothesis regarding the empirical distinguishability of the areas of knowledge and skill represented by each of the dimensions. Comparison of individual item parameter estimates (i.e., difficulty estimates) to the construct maps provides evidence relevant to the hypotheses regarding the ordering of items--and more generally, the nature of variation--within each dimension (Wilson, 2005). Examination of item fit statistics (i.e., infit and outfit mean-square statistics, which indicate the extent to which response patterns for *each item* conformed to model expectations) provides evidence relevant to (a) the item-response-specific hypotheses as reflected in the construct maps and (b) general item quality (insofar as evidence of statistical misfit can serve to flag problematic items for further review). All Rasch models were estimated using marginal maximum likelihood via the

Test Analysis Module (TAM; Robitzsch, Kiefer, & Wu, 2017) package in the R statistical software environment.

Instances of severe deviation from theory-based expectations were flagged for further review. All items from Study 1 with any amount of misfit (according to either fit statistic) that was statistically significantly different from zero ($p < .01$) were flagged for discussion, beginning with those showing the most severe positive misfit (i.e., "underfit") and progressively moving in to the items with less severe misfit. When examining the data from Study 2, we focused on items for which either fit statistic was outside the range (.91, 1.09), and was statistically significantly different from zero. This tolerance range reflects an estimated sample-size-adjusted 99% interval for the expected value of the outfit statistic under the null hypothesis that the true value is 1 (i.e., that the item perfectly fits the Rasch model; Wu & Adams, 2013), and is more stringent than any other recommended tolerance range of which we are aware (e.g., any of the ranges suggested by Bond & Fox in 2015, which were not adjusted for sample size, the most stringent of which is .8 - 1.2).

Finally, we examined the person-separation reliability of each dimension (which is nearly identical to Cronbach's alpha, but robust to planned missing data).

For Study 1, identification of poor fit of an item was treated as a potential falsification of one or more of the hypotheses (both construct-level and item-level) implied by the construct maps. These results were triangulated with data from the cognitive interviews, and ultimately led to one or more of the following outcomes: (a) revision of the construct map of the relevant dimension, (b) revision of the item format for a set of items (e.g., moving from binary forced-choice to conventional multiple-choice options for the for the cognate (COG) and morphological awareness (MA) dimensions), (c) revision of the individual item, (e.g., when an item was found to have more than one plausibly correct answer or was poorly understood by many respondents), or (d) deletion of the item. Similar procedures were followed in Study 2, but with a reduced emphasis on the generation of new elements (e.g., new items and item formats) that would necessitate further testing.

To test the hypothesis that the items work in a comparable ways for individuals from different cultural or linguistic backgrounds, measurement invariance was evaluated via differential item functioning (DIF) analyses by language background. More specifically, for any given item, a DIF statistic (formalized as an interaction term between item and person; see Wu, Adams, Wilson, & Haldane, 2007, p. 89-102) tests the hypothesis that the item is more difficult for a member of one group compared to a member of another group, while holding constant the underlying level of the property.[3]

Results are discussed below for Study 2, comparing students who primarily spoke Spanish at home were compared to those who did not primarily speak Spanish at home. 52% of the students from the total sample primarily spoke English at home, while 26% primarily spoke Spanish at home.[4] A full breakdown of student reported home language is given in Table 1.

[insert Table 1 here]

[3] It may be worth noting that while findings of DIF can provide evidence of potential bias or systematic differences at the item level, such analyses are not well-suited to detecting systematic differences at the level of an entire test (at least in the absence of a known external criteria), given that such differences may be equally present in all items.

[4] Another 15% of students either skipped the question or selected "other" when reporting home language; these students were not included in DIF analyses. Another 17 primary home languages were reported in the dataset, the largest group of which (Arabic) comprised only 1% of the sample.

Items exhibiting DIF were then examined in greater detail, and an additional round of cognitive interviews (30 hours of think-alouds in individual and group configurations) was conducted, this time with six undergraduate native speakers of Spanish. This choice was based on the reasoning that older native Spanish speakers would be able to offer more detailed metacognitive accounts during think-alouds while retaining the tacit linguistic knowledge (*sprachgefühl*) crucial to the response processes under study.

Results

Results from Rasch analyses

Results given here are for Study 2, as discussed previously.

Item difficulty estimates. A unique and useful feature of item and person estimates derived from the Rasch model (given that the model fits the data) is that they can be directly compared to one another on a common scale. For each dimension, the scale represents both the range of person locations--that is, estimated levels of mastery of the relevant area of knowledge and skill--and the range of item difficulties, which can be directly interpreted as points along the continuum at which at about half of the respondents would answer the item correctly. This, in turn, means that an estimated location of a student along a given dimension can be interpreted by examining which items are below, near, and above the estimated location of that student: those below the student's location reflect knowledge and skills the student has likely mastered, those above the student's location represent knowledge and skills yet to be acquired, and those near the student's location represent those that are challenging but achievable for the student given their current stage of development.

A visual illustration of the ranges of estimated person locations and item difficulties is given by the multidimensional Wright map (Wilson, 2005) in Figure 6 below. The histograms on the left-hand side of the figure represent estimated distributions of person locations along each of the four dimensions, and the histogram on the right-hand side represent estimated item difficulties, with dimensions coded by color. On average, it appeared that respondents found the CC dimension to be the easiest, and the COG dimension to be the hardest (although the COG dimension also contained the easiest item on the SUM).

Wright maps can be viewed as empirical analogues to construct maps, and provide a tool for evaluating the extent to which observed item difficulties matched expectations based on theory. For example, six levels of difficulty were postulated for COG. Emergent demonstrations of such knowledge suggest the ability to identify a cognate connection between the most commonly used words across both English and Spanish (e.g, *map/mapa*). Higher levels of cognate knowledge were determined by the extent to which a word is used in academic or formal contexts (*evidence/evidencia*) and the added complexity of embedding such words in text, which in turn requires additional reflection and reasoning compared to presenting individual words to respondents.

The estimated difficulty of each item was qualitatively compared to its intended difficulty, and those with noticeable discrepancies were further discussed with the research team, including consultation with panel experts as needed. Such discussions led to either a revision of the construct map, a revision of the item, or the deletion of the item. For example, one item on the CC dimension asked respondents to select the correct meaning for a nonsense word “torp” within the sentence “mixing together different plant *torps* makes the richest compost”. This item was expected to be moderately difficult (at the fourth level of six total levels of the construct map). Empirically, however, it was the most difficult item on this dimension. Follow-up investigations suggested that this unexpected level of difficulty could be attributed to two construct-irrelevant issues. First, the item contained the word “compost,” which was less familiar to students in this population than was anticipated. Second, follow-up cognitive interviews revealed that at least one of the distractors (“food”) was plausibly also a correct answer. This finding was used to edit the item to reduce the vocabulary demand by replacing “compost” with “soil”, and to ensure a clear best response (“material”). A number of other items within unexpectedly high or low difficulty were also determined to have multiple plausibly correct answers, and were revised accordingly.

In addition, we explored potential issues with unexpected difficulty due to prior knowledge of vocabulary. An item within the MMRT dimension, for example, was inadvertently one of the most difficult items of the entire assessment. On closer inspection during cognitive interviews, the prompt to select the correct term (“gold”) that would fit as secondary to the overarching topic (“elements”) lacked sufficient contextual information, thus requiring a construct-irrelevant form of prior knowledge (i.e., about what would be considered an element).

[insert Figure 6 here]

Item fit statistics. In general, item fit statistics suggested good fit to the Rasch model. Infit and outfit mean square statistics for all but one of the items fell within the range of (.8, 1.2), which is the most conservative range of acceptability of item fit statistics given by Bond and Fox (2015). 111 out of 157 items fell within the even stricter, sample-size adjusted range of (.91, 1.09), with 17 items underfiting the model (that is, exhibiting response patterns less predictably aligned with variation in person locations than would be expected by the model), and 21 items overfitting the model (that is, exhibiting response patterns *more* predictably aligned with variation in person locations than would be expected by the model, which in general is not as great a source of concern). Table 2 shows the proportion of well-fitting, underfitting, and overfitting items for each of the four dimensions.

All misfitting items were closely examined for content and clarity, and potentially flagged for further investigation via cognitive interviews, as described further below. In all cases, misfitting items reflected technical problems such as containing more than one plausibly correct answer or a typographical error. In all, four CC items, two MA items, one MMRT item, and one cognate item was removed, yielding a final version of the SUM with 149 items.

[insert Table 2 here]

Multidimensional models. The dimensionality of the SUM was investigated by examining the fit to the data of a series of unidimensional and multidimensional Rasch models.

The first set of analyses utilized data from all SUM items. Within this set, the first model fit was a unidimensional Rasch model, which ignored the distinctions between the items on the four areas (MA, COG, CC, and MMRT), thus modeling all SUM items as measuring a single dimension (which might be termed “knowledge of reading strategies”). Following this, the second model fit was a four-dimensional Rasch model, which modeled the items from each of the four areas as measuring a distinct dimension. In principle, if the items on the four areas provide meaningfully distinct information as intended, the four-dimensional model should fit the data significantly better than a unidimensional model. This was indeed the case, as the multidimensional model fit the data better than the unidimensional model ($\chi^2(9)$=5,560, *p* <.001), as seen in Table 3. The multidimensional model also had lower estimated values of the parsimony-adjusted Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC).

[insert Table 3 here]

Estimated correlations between the dimensions of the multidimensional Rasch model are given in Table 5. The estimated correlations between MA, CC, and MMRT dimensions were all between .7 and .8.[5] As expected, the Cognate dimension exhibited much weaker correlations with the other four dimensions. Given this, a second series of models was fit to the data, this time removing the Cognates items entirely; these analyses were intended to isolate the hypotheses that the other three dimensions should

[5] Correlations among raw scores are not disattenuated for measurement error, and thus are somewhat lower than model-estimated correlations among dimensions, as reflected in Table 5.

still be empirically distinguishable from one another, and thus rule out the alternative hypothesis that the superior fit of the multidimensional model described previously was due entirely to the Cognates dimension.

[insert Tables 4 and 5 here]

Within this second set of analyses, the first model fit was again a unidimensional Rasch model, thus treating the items on the three non-COG areas (MA, CC, and MMRT), as measuring a single dimension. This model was compared to a three-dimensional model (with MA, ICC, and MMRT items modeled as each measuring separate dimensions), as well as to three intermediary two-dimensional models, in which items from two of the three areas were modeled as measuring a common dimension and items from the remaining area were modeled as measuring a distinct dimension (e.g., with MA items measuring one dimension and ICC and MMRT items combined to measure a separate dimension). All three of the two-dimensional models fit the data better than the unidimensional model, and the three-dimensional model fit better than all four of the lower-dimensional models; in addition, this model had the lowest estimated AIC and BIC values.

As previously mentioned, the ultimate intention is for the SUM to be made freely available as an online computerized adaptive test, and for model-estimated Expected A Posteriori (EAP) scale scores (rather than raw scores) to be reported to teachers (as confidence bands rather than point estimates, thus incorporating information about measurement error). These scale scores will be estimated using the four-dimensional

Rasch model described previously. Such EAP scores take into account the correlation between SUM dimensions, similar to “augmented subscores” (see, e.g., Wainer et al., 2001); thus estimates of students’ locations along each of the four dimensions are pulled slightly in the direction of their performance on the other three dimensions (“borrowing strength” across dimensions), helping safeguard (along with the presentation of confidence bands and supporting documentation) against the possibility of over-interpretation of a student’s exceptional performance on one area of the SUM (see, e.g., Briggs & Wilson, 2003).

Measurement precision. Estimated person-separation reliability coefficients for each of the four dimensions of the SUM are given in Table 6. As these estimates were based only on pretest data, as described previously, they reflect the level of measurement precision that can be expected when a student responds to roughly half of the SUM items. In general, these estimates suggest that the SUM provides a high level of measurement precision across all dimensions, with the lowest (.79) on the cognate dimension.

Measurement precision is highest for person locations closest to the middle of the scale, with somewhat lower precision at the highest and lowest ranges of the scale.

[insert Table 6 here]

Differences by language background. On average, students whose primary home language was Spanish had significantly higher estimated locations on the cognate dimension compared to those whose home language was not Spanish, but somewhat lower estimated locations on the other three dimensions, as illustrated in Table 7.

As described previously, each item was examined for differential functioning based on linguistic background, by estimating the interaction effect between language and item difficulty while holding person location constant. Given that these analyses were exploratory, and also given the inflated chance of false positives associated with the large number of individual significance tests, items were flagged using an alpha level of .01, and results of these analyses were used only as a rough guide for flagging items for further review via linguistic analysis and cognitive interviews. Each dimension was analyzed individually.

As previously noted, and as expected, Spanish speakers performed significantly better on the cognate dimension; this was true more or less uniformly for all items, and so DIF analyses for this dimension did not produce meaningful results.

DIF for MA items. Of the 49 MA items, no evidence of DIF was found for 46. One item showed evidence of being harder for Spanish speakers (by .3 logits), and two items showed evidence of being harder for non-Spanish speakers (by up to .4 logits).

DIF for CC items. Of the 27 CC items, 23 showed no evidence of DIF. Two items showed evidence of being harder for Spanish speakers (by up to .37 logits), and two showed evidence of being harder for non-Spanish speakers (by up to .61 logits, though only one item had an estimated difference of more than .5 logits).

DIF for MMRT items. Of the 30 MMRT items, 22 showed no evidence of DIF. Four items showed evidence of being harder for Spanish speakers (by up to .34 logits), and four showed evidence of being harder for non-Spanish speakers (by up to .56 logits, only one of which was above .5).

[insert Table 7 here]

Associations between SUM dimensions and reading comprehension. For the data obtained in Study 1, scores on the SUM dimensions were moderately associated with scores on the Gates-MacGinitie Reading Test, most strongly for the MA ($r = .52$, $p < .05$) and MMRT ($r = .60$, $p < .05$) dimensions, and more moderately for the CC ($r = .40$, $p < .05$) dimension. (The COG dimension was not used in this analysis).

Results from cognitive interviews

As a follow-up to the psychometric analyses described in the previous section, two additional sets of cognitive interviews were conducted (one with middle school students, and one with undergraduates) to further diagnose potential sources of DIF (Ercikan et al. 2010), as well as to further investigate the extent to which the hypothesized reading strategies were employed while making sense of and responding to items.

Cognitive interviews with middle school students. Interview responses revealed potential sources of linguistic and cultural differences in item responses among Spanish-speaking youth. A methodology was adopted that attempted to make the best use of both concurrent and retrospective judgments about items (Taylor & Dionne, 2000). After responding to each item in a concurrent verbal protocol, participants were asked retrospectively whether they believed the item would be easier for native Spanish speakers, native English speakers, or neither. For example, an item required knowledge of the non-cognate morpheme *mis-* (e.g. misanthropy) was judged to be more difficult for Spanish speakers because,

> "Mis- isn't really a clear like word? So like *contra-* at least that's like 'against' and non- is '*no*', and then mis-, the closest thing I can think of is 'mistake' I guess. If you start to think of other words that start with mis- you would get to like understanding what the meaning of that, I don't know what it's called, the part of the word?" [Isabella, Int2, 5:00-5:38]

On many items, participants employed Spanish as a resource for reading comprehension via cognate morphemes. Conversely, participants judged items as more difficult for Spanish speakers when fewer cognate morphemes were present. Referring to a prior academic testing experience, one participant explained,

> "That's what would happen to me a lot with my exams. So, when I had the CST's back then, like there was some words and it was like 'the Latin word blank', like I would feel like I was cheating in a way because sometimes the Latin roots have like some Spanish words…. and then I would read it and it would sound like Spanish and I was like 'Oh.' I mean it seemed kind of obvious" (Julia, Int2, 22:00-22:32).

According to several of the participants, more balanced items included more context clues and less unnecessary linguistic complexity, consistent with research on the construction of tests for multilingual populations (e.g., Abedi, Leon, Wolf & Farnsworth, 2008). Participants pointed to both the inherent difficulty of taking a reading assessment written in one's second language while also making adept use of metalinguistic knowledge and skills conferred by bilingualism.

In addition, qualitative information derived from cognitive interviews further suggested that respondents are indeed employing strategies while addressing items. For example, when offering a rationale for selecting a particular response from a set of choices, common expressions indexed the practice of looking at parts of a word or sentence for gaining insight to nonsense word meanings. Demonstrated reasoning during interviews also suggested that regardless of perceived Spanish language ability, students tend to use all possible knowledge (e.g., *I think we learned this in class, that "dad" means like a place or city in Spanish*) in determining the relationship between English and Spanish vocabulary.

Cognitive interviews with undergraduate students. Similar to the process used with participating children, a series of think-aloud sessions using randomly selected items were conducted with seven undergraduate students. The use of the focal strategy (MA, ICC, MMRT, COG) and the absence of construct-irrelevant strategies constitute two necessary conditions of validity. Each item response was coded for 1) the kinds of strategies employed in solving the item, 2) whether the focal strategy was used, 3) whether any unexpected strategies were used, and 4) whether or not the answer was correct. For example, a participant might use a combination of the focal strategy, such as looking for context clues, as well another acceptable test strategy, such as process of elimination. If a participant's answer was based on an unexpected strategy, such as guessing, the item was flagged for further exploration.

Detailed results from the inter- and intra-sentential context clues section illustrate the value of this method. Of the 30 items used in the section, undergraduate participants correctly answered 28, and used expected strategies for each of these. The two incorrect responses resulted from unexpected response processes. For example, in the first case, a participant explained that she was relying on mere association between the target words to respond to the item, rather than on context clues; in the second, a participant carried forward reasoning from a previous item without using context clues.

Overall results of the response process study were used to identify as problematic and ultimately remove three items not previously flagged by the psychometric analyses. In one case, an item (MA) presumed cultural knowledge about the practice of "remodeling" a house, which we discovered was not shared by all participants in the study. In another case, an item (COG) included the focal word "librería", a cognate of ambiguous meaning which, according to the dictionary of the Real Academia Española, may mean "store in which books are sold", "collection of books", and "place in which books are found" (with "biblioteca" offered as a synonym). Possible responses to the item included each of these three meanings ("books", "library", and "bookstore"). This linguistic analysis converged with the findings of the analyses of DIF, which showed that this was one of the only cognate items that actually disfavored young students who were Spanish speakers. The third item (MMRT) contained more than one logically plausible answer.

Results from linguistic analysis

In order to help test the hypothesis that the SUM correctly targeted the intended constructs for the intended populations, DIF in this study was first statistically estimated and then items showing evidence of DIF were qualitatively examined via linguistic analyses. Statistical identification of a significant interaction between item and group membership was treated as a flag for evidence of true differential item functioning. Special attention was paid to the validation of the measure for Spanish-speaking students. Differences in item functioning between groups are common, but it is important to make sure that these differences are attributable to non-arbitrary factors, such as different levels of preparation or prior opportunity to engage with the material. We found that cognate morphology (and its absence) furnish a non-arbitrary reason for differences in performance between students of varying linguistic heritage. For the MA section, the cognate morpheme hypothesis was evaluated by coding items by the number of Latinate cognate morphemes in the answer or focal word, divided by the total number of morphemes in the word, resulting in a 0-1 scale of cognate morphology. A Pearson correlation was run to test whether cognate morphology was associated with DIF values. A positive correlation emerged (r=.33, p<.05), lending support to our hypothesis. For example, a 33% in the amount of cognate morphemes in a word was associated with a .5 logit difference in item difficulty for the focal groups. For the CC section, the cognate morpheme hypothesis was evaluated using a similar procedure, scoring items by the number of cognate words in the prompt and responses, divided by the total word length of the passage. A similar positive correlation emerged (r=.4, p<.02). On this section, passages with DIF favoring English monolinguals had roughly half the proportion of cognate words as corresponding passages favoring emergent Spanish bilinguals. In both cases, analysis of residuals served as a guide to which items followed the expected pattern.

Pedagogical applications for teachers

Results from the SUM have been used to provide a picture of the competencies and gaps in reading strategy knowledge for teachers. Researchers met with three teachers and interpreted SUM results for individual students from two classrooms. During these meetings, teachers directed the process of focusing on particular students about whose capacities they wanted to know more, while researchers used results from each of the four dimensions of the SUM to offer informed explanations of observed performances. The teachers reported that these results deepened their understanding of the literacy competencies of their students. "It was really helpful when I saw the results and I started figuring out how to help my students," attested one teacher during a training session. The teacher went on to request that test results be shared with teachers of the same students in subsequent grades.

The students from these two classrooms who took the SUM now participate in a collaborative strategic reading curriculum offered as part of our university-school partnership. This program is similar to the original CSR program, with the added components of critical reading and civic engagement within interdisciplinary learning contexts (Arya & Maul, 2016; Arya et al., 2017; McBeath, Harlow, Arya & Longitin, in press). The program emphasizes small-group configurations involving both heterogeneous reading discussion groups (Spanish bilingual speakers teaching English-only speakers related cognates) and differentiated instructional practice (e.g., identifying those students who would benefit from word games related to morphemes), based on research demonstrating the benefits of differentiated instruction for developing specific knowledge and skills (Tieso, 2003; Steenbergen-Hu, Makel & Olszewski-Kubilius, 2016) as well as fostering heterogeneous peer groups for building confidence and classroom community (Belfi, Goos, De Fraine, & Van Damme, 2012; Oakes, 2008). SUM results helped inform the composition of these groups.

Perhaps most tellingly, all three teachers (and their principal) requested that the SUM be administered once again to guide instruction, expressing the view that results of the SUM are easier for them to interpret than those of state-level reading assessments. Teachers stated that the SUM and the collaborative reading program have prompted them to learn more about the multifaceted nature of reading and current scientific knowledge about reading instruction.

In our practice, we have used scores on the cognate dimension to make instructional recommendations for teachers of emergent bilinguals. When Spanish-speaking students score in the lower and middle range of the cognate dimension, we have interpreted this to mean that they may benefit from practice in using cognates to gain a foothold in inferences about the meanings of words. Such exercises can involve vocabulary review of words and morphemes that function as Spanish-English cognates, such as with flashcards or with texts adapted specifically for bilinguals.

Discussion

The purpose of this paper was to introduce a new measure of reading strategy use, designed primarily for formative purposes, and to present and evaluate multiple sources of evidence for its validity, especially as related to the interpretation and use claims described previously.

Using the language of the *Standards for Educational and Psychological Testing* (AERA, APA, NCME, 2014), we have presented (a) *evidence based on test content* in the form of a discussion of the theories on which the instrument was built (illustrated in particular by the construct mapping process) and description of the procedures for item creation and revision, as well as post-hoc linguistic analyses of items, (b) *evidence based on response processes* in the form of description of results from cognitive interviews, triangulated with results from psychometric analyses (in particular, identification of DIF), (c) *evidence based on the internal structure of the test* in the form of psychometric analyses, and in particular evidence for the fit of response data to a multidimensional Rasch model and estimation of its precision/reliability of measurement, and (d) *evidence of relations with other variables* in the form of associations with the Gates-MacGinite reading test. *Fairness* has also been foregrounded throughout, with a particular focus on fairness to individuals from non-English-dominant backgrounds, though extensive investigation of the role of language in the production of item responses. *Evidence based on the consequences of testing* is, at this point, limited, as the test has not yet seen operational use beyond the two major studies involved in its construction and validation. This said, every effort has been made to anticipate possible negative consequences of the use of the test, including (as has been described in this paper) carefully investigating potential sources of bias due to linguistic background, and other construct-irrelevant forms of variance.

Findings indicate tentative support for the measurability and empirical distinguishability of each of the four dimensions, though further work is needed to develop optimal score-reporting systems that provide an appropriate balance between effective presentation of potentially pedagogically useful information and safeguards against possible over-interpretation. Measurement precision is high, though there is a relative absence of easy items for this population, which may reflect the novelty of reading strategy instruction within the participating school district. This finding suggests the need for consideration of items reflecting more basic or emergent forms of reading strategies, particularly if the instrument is used for younger populations in the future.

The relatively low correlation between the Cognates dimension and the other three dimensions was to be expected, as was the fact that individuals who speak Spanish at home tended to perform more successfully on this dimension, based on the requirement of Spanish language knowledge for successful responses to items. Cognitive interviews with native speakers of Spanish yielded further evidence that multilingual learners utilize cognate morphology as a resource to respond to items targeting morphological awareness.

Previously, we proposed aspirational interpretation and use claims regarding this instrument. Preliminary evidence has been provided in support of the interpretation claim, as summarized here, though further work is needed in particular to evaluate the use claim and develop tools to help support the effective use of the SUM for pedagogical purposes.

Future investigations and applications

The SUM has been created and tested for students in grades 6 through 8. As our understanding of the development of reading strategy use continues to improve, further work will aim to make the SUM appropriate for students both younger and older than this range. In particular, further research into early reading practices and related instructional scaffolds may inform the nature and format of items targeted to younger (elementary) students, as well as some students in the high school range. Similarly, further research on the growing complexities of strategy knowledge and use as students approach graduation and higher education opportunities will help to inform the development of a measure that could be used across the lifespan of learning.

Future work on the SUM will emphasize sustainability and efficiency, and aims to eventually result in a pre-calibrated item bank sufficient to support a computerized adaptive testing (CAT) format. Such a format will help maximize measurement precision at the student level while reducing the amount of time necessary for testing. Even without this goal, ongoing item analysis and refinement will be continually necessary; as societies, and particularly learning communities within societies, continue to evolve culturally, socially, and linguistically, the SUM (like all reading assessments) will require continual monitoring to ensure fairness (Bond, 1995; Solano-Flores & Soltero-González, 2011). Additionally, as the SUM is increasingly used in the classroom, close early monitoring and solicitation of feedback from teachers, students, and other stakeholders will help to collect further evidence of the validity of the test based on the consequences of its use, and to develop clearer and more specific guidelines concerning optimal instructional moves based on SUM results.

Given the demonstrated value of explicit instruction in reading strategies for students from linguistically diverse backgrounds and the paucity of existing performance-based measures of reading strategy use, there is a clear need for accessible, pedagogically-informative, rigorously validated measures in this area. We hope that the SUM begins to address this need.

References

Abedi, J., Leon, S., Wolf, M. K., & Farnsworth, T. (2008). Detecting test items differentially impacting in the performance of ELL students. In M.K.Wolf, J.L.Herman, J.Kim, J.Abedi, S.Leon, N.Griffin, P.L. Bachman, S. M. Chang, T. Farnsworth, H. Jung, J. Nollner, & H. W. Shin (Eds.), *Providing validity evidence to improve the assessment of English language learners* (pp. 55–75). CRESST Report 738. Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing.

Adams, N. J. (2001). On the lexile framework. In National Center for Education Statistics (Ed.), Assessing the lexile framework. Results on a panel meeting NCES 2001-08 (pp. 15–21). Washington, DC: National Center for Education Statistics.

Adams, R. J., Wilson, M., & Wang, W. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement*, 21(1), 1–23.

AERA, APA, & NCME. (2014). *Standards for educational and psychological testing.* Washington, D.C.: American Psychological Association, Inc.

Afflerbach, P., Pearson, P. D., & Paris, S. G. (2008). Clarifying differences between reading skills and reading strategies. *The Reading Teacher*, *61*(5), 364-373.

Anderson, R. C., & Freebody, P. (1982). Reading comprehension and the assessment and acquisition of word knowledge. *Center for the Study of Reading Technical Report; no. 249*.

Armbruster, B. B., Anderson, T. H., & Meyer, J. L. (1991). Improving content-area reading using instructional graphics. *Reading Research Quarterly*, 393-416.

Arnbak, E., & Elbro, C. (2000). The effects of morphological awareness training on the reading and spelling skills of young dyslexics. *Scandinavian Journal of Educational Research*, *44*(3), 229-251.

Artley, A.S. (1943). Teaching word meaning through context. *The Elementary English Review*, *20*(1), 68–74.

Arya, D.J. Harlow, D. Hansen, Harmon, L., A. McBeath, J. & Pulgar, J. (2017). Innovative youth: An engineering and literacy integrated approach. *Science Scope*, 40(9), 82-88.

Arya, D. & Maul, A. (2016). The building of knowledge, language, and decision-making about climate change science: a cross-national program for secondary students. *International Journal of Science Education*, 1-20.

Baker, L., & Brown, A. L. (1984). Metacognitive skills and reading. *Handbook of reading research*, *1*, 353-394.

Belfi, B., Goos, M., De Fraine, B., & Van Damme, J. (2012). The effect of class composition by gender and ability on secondary school students' school well-being and academic self-concept: A literature review. Educational Research Review, 7, 62–74.

Berninger, V. W., Abbott, R. D., Nagy, W., & Carlisle, J. (2010). Growth in phonological, orthographic, and morphological awareness in grades 1 to 6. *Journal of Psycholinguistic Research, 39*(2), 141-163.

Blackowicz, C. Z., & Fisher, P. (2000). Vocabulary instruction. *ML Kamil, PB*.

Bond, L. (1995). Unintended consequences of performance assessment: Issues of bias and fairness. *Educational Measurement: Issues and Practice*, *14*(4), 21-24.

Bond, T., & Fox, C. M. (2015). Applying the Rasch model: Fundamental measurement in the human sciences (Third). New York: Routledge.

Borsboom, D., & Mellenbergh, G. (2004). Why psychometrics is not pathological: A comment on Michell. *Theory and Psychology, 14,* 105-120.

Bowers, P. N., Kirby, J. R., & Deacon, S. H. (2010). The effects of morphological instruction on literacy skills: A systematic review of the literature. *Review of educational research*, *80*(2), 144-179.

Brown, A. L., Armbruster, B. B., & Baker, L. (1986). The role of metacognition in reading and studying. *Reading comprehension: From research to practice*, 49-75.

Brown, A. L., & Palincsar, A. S. (1982). Inducing strategic learning from texts by means of informed, self-control training. *Topics in Learning & Learning Disabilities*.

Cantrell, S. C., Almasi, J. F., Carter, J. C., Rintamaa, M., & Madden, A. (2010). The impact of a strategy-based intervention on the comprehension and strategy use of struggling adolescent readers. *Journal of Educational Psychology*, *102*(2), 257.

Carlisle, J. F. (2004). Morphological processes that influence learning to read. *Handbook of language and literacy: Development and disorders*, 318-339.

Carr, E., & Ogle, D. (1987). KWL Plus: A strategy for comprehension and summarization. *Journal of reading*, *30*(7), 626-631.

Chang, K. E., Sung, Y. T., & Chen, I. D. (2002). The effect of concept mapping to enhance text comprehension and summarization. *The Journal of Experimental Education*, *71*(1), 5-23.

Cho, B. Y. (2013). Adolescents' constructively responsive reading strategy use in a critical internet reading task. *Reading Research Quarterly*, *48*(4), 329-332.

Collins, A., & Smith, E. (1980). Teaching the process of reading comprehension (Tech. Report No. 182.). Urbana, IL: Illinois University, Center for the Study of Reading (ED 193 616).

Cooter Jr, R. B., & Curry, S. (1989). Gates-MacGinitie Reading Tests. *Reading Teacher*, *43*(3), 256-58.

Cromley, J., & Azevedo, R. (2011). Measuring strategy use in context with multiple-choice items. *Metacognition and Learning*, *6*(2), 155-177.

Desimone, L.M. & Le Floch, K.C. (2004). Are we asking the right questions? Using cognitive interviews to improve surveys in education research. *Education Evaluation and Policy*, 26(1), 1-22.

Dewitz, P., Jones, J., & Leahy, S. (2009). Comprehension strategy instruction in core reading programs. *Reading Research Quarterly*, *44*(2), 102-126.

Duke, N. K., & Pearson, P. D. (2009). Effective practices for developing reading comprehension. *Journal of education*, *189*(1-2), 107-122.

Ercikan, K., Arim, R., Law, D., Domene, J., Gagnon, F., & Lacroix, S. (2010). Application of think aloud protocols for examining and confirming sources of differential item functioning identified by expert reviews. *Educational Measurement: Issues and Practice*, 29(2), 24-35.

Guthrie, J. T., Van Meter, P., Hancock, G. R., Alao, S., Anderson, E., & McCann, A. (1998). Does concept-oriented reading instruction increase strategy use and conceptual learning from text? *Journal of Educational Psychology*, *90*(2), 261.

Janzen, J. and Stoller, F.L. (1998). Integrating strategic reading in L2 instruction. *Reading in a Foreign Language 12 (1)*, 251-269.

Jiménez, R.T. (1997). The strategic reading abilities and potential of five low-literacy Latina/o readers in middle school. *Reading Research Quarterly,* 32(3), 224-243.

Jiménez, R. T., García, G. E., & Pearson, P. D. (1995). Three children, two languages, and strategic reading: Case studies in bilingual/monolingual reading. *American Educational Research Journal*, *32*(1), 67-97.

Jordan, N. C. (1994). Developmental perspectives on reading disabilities. *Reading & Writing Quarterly: Overcoming Learning Difficulties*, *10*(4), 297-311.

Kane, M. T. (2006). Validation. In R.L. Brennan (Ed.), *Educational measurement, 4th edition* (pp. 17-64). Santa Barbara: Greenwood Publishing Group.

Kintsch, W., & Van Dijk, T. A. (1978). Toward a model of text comprehension and production. *Psychological Review*, *85*(5), 363.

Klingner, J. K., & Vaughn, S. (1999). Promoting reading comprehension, content learning, and English acquisition through Collaborative Strategic Reading (CSR). *The Reading Teacher*, *52*(7), 738-747.

Klingner, J. K., & Vaughn, S. (2000). The helping behaviors of fifth graders while using collaborative strategic reading during ESL content classes. *TESOL Quarterly*, *34*(1), 69-98.

Kucan, L., Palincsar, A. S., Busse, T., Heisey, N., Klingelhofer, R., Rimbey, M., & Schutz, K. (2011). Applying the Grossman et al. theoretical framework: The case of reading. *Teachers College Record*, *113*(12), 2897-2921.

Linn, R. L., & Valencia, S. (1986). *Reading assessment: Practice and theoretical perspectives*. Center for the Study of Evaluation, Graduate School of Education, University of California, Los Angeles.

Lubliner, S., & Grisham, D. L. (2012). Cognate strategy instruction. *Learning from Culturally and Linguistically Diverse Classrooms: Using Inquiry to Inform Practice*, 105.

McBeath, J., Harlow, D., Arya, D., & Longtin, M. (in press). From becoming to being scientists: Developing STEM programming for girls using design-based implementation research approaches, *Afterschool Matters*.

Montelongo, J. A., Hernández, A. C., Herter, R. J., & Cuello, J. (2011). Using cognates to scaffold context clue strategies for Latino ELs. *The Reading Teacher*, *64*(6), 429-434.

National Research Council. (1998). *Preventing reading difficulties in young children*. National Academies Press.

Oakes, J. (2008). Keeping track: Structuring equality and inequality in an era on accountability. *Teachers College Record,* 110, 700–712.

Oliver, K. (2009). An investigation of concept mapping to improve the reading comprehension of science texts. *Journal of Science Education and Technology*, *18*(5), 402-414.

Palincsar, A. S., Magnusson, S. J., Pesko, E., & Hamlin, M. (2005). Attending to the nature of subject matter in text comprehension assessments. In S.G. Paris and S.A. Stahl (Eds), *Children's reading comprehension and assessment* [pp. 257-278] NJ: Lawrence Erlbaum.

Paris, S. G., & Oka, E. R. (1989). Strategies for comprehending text and coping with reading difficulties. *Learning Disability Quarterly*, 32-42.

Pearson, P. D., Hiebert, E. H., & Kamil, M. L. (2007). Vocabulary assessment: What we know and what we need to learn. *Reading research quarterly*, *42*(2), 282-296.

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

Pressley, M., & Afflerbach, P. (1995). *Verbal protocols of reading: The nature of constructively responsive reading.* New York: Routledge.

Rasch, G. (1960/1980). *Probabilistic models for some intelligence and attainment tests*. Chicago, IL: University of Chicago Press.

Reynolds, D., & Daniel, S. (2017). Toward contingency in scaffolding reading comprehension: Next steps for research. *Reading Research Quarterly*. DOI: 10.1002/rrq.200

Rodríguez, T.A. (2001). From the known to the unknown: Using cognates to teach English to Spanish-speaking literates. *The Reading Teacher*, *54*(8), 744–746.

Robinson, J. A., & Flores, T. P. (1997). Novel techniques for visualizing biological information. *ISMB*, 5, 241-249.

Robitzsch, A., Kiefer, T., & Wu, M. (2017) TAM: Test Analysis Modules. CRAN. https://CRAN.R- project.org/package=TAM R package version 2.5-14.

Sáenz, L. M., & Fuchs, L. S. (2002). Examining the reading difficulty of secondary students with learning disabilities: Expository versus narrative text. *Remedial and Special Education*, *23*(1), 31-41.

Scalise, K., & Allen, D. D. (2015). Use of open-source software for adaptive measurement: Concerto as an R-based computer adaptive development and delivery platform. *British Journal of Mathematical and Statistical Psychology*, *68*(3), 478-496

Slavin, R. E., Lake, C., Davis, S., & Madden, N. A. (2011). Effective programs for struggling readers: A best-evidence synthesis. *Educational Research Review*, *6*(1), 1-26.

Snow, C. E., August, D., & White, C. E. (2011). Spanish-speaking students' use of cognate knowledge to infer the meaning of English words. *Bilingualism: Language and Cognition*, *14*(2), 243-255.

Solano-Flores, G., & Soltero-González, L. (2011). Meaningful assessment in linguistically diverse classrooms. *Teacher preparation for bilingual student populations: Educar para Transformar*, 146-163.

Stahl, S. A., Jacobson, M. G., Davis, C. E., & Davis, R. L. (1989). Prior knowledge and difficult vocabulary in the comprehension of unfamiliar text. *Reading research quarterly*, 27-43.

Steenbergen-Hu, S., Makel, M. C., & Olszewski-Kubilius, P. (2016). What one hundred years of research says about the effects of ability grouping and acceleration on K–12 students' academic achievement: Findings of two second-order meta-analyses. *Review of Educational Research*, *86*(4), 849-899.

Taylor, K. & Dionne, J. (2000). Accessing problem-solving strategy knowledge: The complementary use of concurrent verbal protocols and retrospective debriefing. Journal of Educational Psychology, 92(3), 413-425.

Tieso, C. L. (2003). Ability grouping is not just tracking anymore. *Roeper Review*, 26, 29–36.

Wainer, H., Vevea, J. L., Camacho, F., Reeve, B. B., Rosa, K., Nelson, L., et al. (2001). Augmented scores—"borrowing strength" to compute scores based on small numbers of items. In D. Thissen & H. Wainer (Eds.), Test scoring (pp. 343–387). Hillsdale, NJ: Lawrence Erlbaum Associates.

Wigfield, A., & Guthrie, J. T. (2010). The impact of Concept-Oriented Reading instruction on students' reading motivation, reading engagement, and reading comprehension. In J. Meece & J. S. Eccles (Eds.), Handbook on schools, schooling, and human develop- ment (pp. 463–477). Mahwah, NJ: Erlbaum.

Wilson, M. (2005). *Constructing measures*. Mahwah, NJ: Lawrence Erlbaum.

Wixson, K. K., & Peters, C. W. (1987). Comprehension assessment: Implementing an interactive view of reading. *Educational Psychologist*, *22*(3-4), 333-356.

Wu, M., & Adams, R. J. (2013). Properties of Rasch residual fit statistics. *Journal of Applied Measurement, 14*, 339-355.

Wu, M., Adams, R.J., Wilson, M. R., & Haldane, S. A. (2007). ACER ConQuest 2.0: Generalised Item Response Modelling Software. Victoria: ACER Press.