Educational and Psychological Measurement
2019, Vol. 79(1) 65–84
© The Author(s) 2018
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/0013164418763255
journals.sagepub.com/home/epm
SAGE

Constructing Subscores That Add Validity: A Case Study of Identifying Students at Risk

Gina Biancarosa[1], Patrick C. Kennedy[1], Sarah E. Carlson[1], HyeonJin Yoon[1], Ben Seipel[2,3], Bowen Liu[4], and Mark L. Davison[4]

Abstract

Prior research suggests that subscores from a single achievement test seldom add value over a single total score. Such scores typically correspond to subcontent areas in the total content domain, but content subdomains might not provide a sound basis for subscores. Using scores on an inferential reading comprehension test from 625 third, fourth, and fifth graders, two new methods of creating subscores were explored. Three subscores were based on the types of incorrect answers given by students. The fourth was based on temporal efficiency in giving correct answers. All four scores were reliable. The three subscores based on incorrect answers added value and validity. In logistic regression analyses predicting failure to reach proficiency on a statewide test, models including subscores fit better than the model with a single total score. Including the pattern of incorrect responses improved fit in all three grades, whereas including the comprehension efficiency score only modestly improved fit in fourth and fifth grades, but not third grade. Area under the curve (AUC) statistics from receiver operating characteristic (ROC) curves based on the various models were higher for models including subscores than those without subscores. Implications for using models with and without subscores are illustrated and discussed.

Keywords

diagnostic testing, formative assessment, at-risk screening, reading comprehension, subscores

[1]University of Oregon, Eugene, OR, USA
[2]University of Wisconsin–River Falls, River Falls, WI, USA
[3]California State University, Chico, CA, USA
[4]University of Minnesota, Minneapolis, MN, USA

Corresponding Author:
Mark L. Davison, Department of Educational Psychology, University of Minnesota, 56 East River Road, Minneapolis, MN 55455, USA.
Email: mld@umn.edu

Guidelines for educational assessment often recommend reporting diagnostic information to guide students and teachers in addressing specific needs (e.g., the *Elementary and Secondary Education Act* of 2001 and 2015), but how can we increase the diagnostic information provided by assessments? One way is by reporting subscores as well as a total score, although researchers have found that subscores often provide relatively little added value over a single total score (Haberman, 2008a; Lyren, 2009; Puhan, Sinharay, Haberman, & Larkin, 2010; Sinharay, 2010). For instance, Puhan et al. found that in a mathematics test for beginning teachers containing four subcontent areas (concepts, integrate knowledge, models, real-life problems), one could estimate a person's true score in one of the content areas (say concepts) more accurately from their total test score than from their subscore in concepts. In these studies, subscores most often corresponded to the number of items correct in subareas of the broader content domain (e.g., real-life problems in the test of mathematics). However, this is not the only way to create subscores. For instance, subscores can be used to quantify the number of incorrect responses of a given type that a student makes or the efficiency with which a student can correctly answer items.

In addition, past evaluations of subscore utility have been based primarily on the internal structure of the subscores: factor structure or internal consistency reliability (e.g., Puhan et al., 2010; Sinharay, 2010). They typically do not consider the incremental validity provided by subscores over and above a total score. The premises of this research are that, given the previous findings on subscores, the field needs to explore other ways of constructing them and needs to explore properties of subscores beyond their internal structure. To support this argument, results for a reading comprehension assessment are presented as support for two other methods for constructing subscores.

In this study, we evaluated reading comprehension subscores based, not on content subareas such as the distinction between literal and inferential comprehension, but on the number of incorrect responses of various types committed by students. We also evaluated a score reflecting the temporal efficiency with which students arrive at correct answers: minutes per correct response. In addition to analyzing the internal structure of the total score and subscores, we examined the validity of the overall score and the incremental validity of subscores, over and above the overall score, in identifying at-risk students. For this purpose, we employed the logistic regression counterpart (Davison, Jew, & Davenport, 2014) of the linear regression procedure proposed by Davison and Davenport (2002; Davison, Davenport, Chang, Vue, & Su, 2015)

Subscore Types

The term *subscore* has at least two different meanings. First, it can refer to a score on a test within a test battery (e.g., the SAT Verbal, Quantitative, and Analytic subtests). Second, it can refer to subareas from a single test (e.g., literal and inferential comprehension scores from a reading comprehension test). The former typically have the advantage that they reflect a wider variety of content areas. The latter have the advantage that they extract additional information without requiring students to take more than one test. Our reading of the literature on subscore utility leads to the conclusion that subscores from a battery are more likely to have value over and above a total score than are subscores from a single test (Haberman, 2008a; Lyren, 2009; Puhan et al., 2010; Sinharay, 2010). However, it could be that the limitations of subscores drawn from a single test were related to how the subscores were defined. For a single test, defining subscores in terms of items from subcontent areas may be ineffective.

The subscore types in the present research were drawn from theories of reading comprehension processing and observations from narrative comprehension think-aloud research (Carlson, Seipel, & McMaster, 2014; Graesser, Singer, & Trabasso, 1994; Kintsch & van Dijk, 1978; McMaster et al., 2012, Rapp, van den Broek, McMaster, Kendeou, & Espin, 2007). This research has established that some who struggle with reading narrative passages can be considered specific poor comprehenders in that they struggle not with decoding (i.e., reading the words off a page), but rather with the comprehension process specifically (e.g., Rapp et al., 2007). Think-aloud research also indicates that poor comprehenders can be differentiated by at least two types of cognitive processes (e.g., paraphrasing and elaboration) on which they over rely often leading to incorrect responses (Carlson et al., 2014; McMaster et al., 2012; Rapp et al., 2007). Some students show a strong predilection to paraphrasing information from a narrative even when the correct answer requires an inference. Others show a predilection for elaborations, evaluations, predictions, or associations that go beyond the literal information in the story but do not relate to the narrative sequence. Most authors refer to this second process as elaboration, because most such responses are elaborations of story information. However, because this second category includes more than just elaborations, it is here called lateral connection. In narrative comprehension, paraphrases and lateral connections represent important comprehension processes and can lead to correct answers. For instance, if an item is a literal comprehension task, a paraphrase will be the correct answer. In the context of inferential comprehension, however, which require inferences that complete a narrative sequence or make connections between two disparate pieces of text, neither paraphrases nor lateral connections complete the sequence of events.

Unlike most traditional multiple-choice tests that have two types of responses for each item, correct and incorrect, the test on which this research is based has three types: correct, paraphrase, and lateral connect. The paraphrase and lateral connect types were created because they mimic the cognitive processes that appear in think-aloud responses, and because of their relationship to documented intervention outcomes. In classroom instruction, poor comprehenders responded differentially to intervention based on their preferred cognitive processes during reading

(McMaster et al., 2012; van den Broek et al., 2006). ‘‘Paraphrasers’’’ comprehension skills improved more in a general questioning condition (e.g., ‘‘Make a connection to what you previously read.’’), whereas ‘‘lateral connectors’’’ comprehension skills improved more from questioning about causal sequence (e.g., ‘‘Why?’’). To date, this instructional finding has been found in classroom settings, but not small group, individualized instruction (McMaster, Espin, & van den Broek, 2014).

Other tests have included reports of misconceptions or errors committed by students (delMas, Garfield, Ooms, & Chance, 2007; Hermann-Abell & DeBoear, 2011; Hestenes, Wells, & Swackhamer, 1992; Sadler, 1998). They are called *distractor-driven assessments* by Hestenes et al. and *concept inventories* by Sadler. However, in these assessments, any particular error type typically appears in only a few items, and hence the frequency with which the error occurs is not a reliable score. For subscores to be useful, they must be reliable (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014; Sinharay, 2010). In contrast, each incorrect response type on the assessment studied here appeared as a response option for every item. Thus, the paraphrase and lateral connect subscores were based on a sufficient number of items to yield reliable subscores.

Another of our subscores is based on the efficiency with which students produce correct answers. It is conceptually similar to the index of fluency used to measure oral reading proficiency (i.e., correct words per minute), but is calculated and reported as the inverse of correct words per minute: i.e., minutes per correct response. Comprehension efficiency is determined by dividing the total testing time by the number of items answered correctly. In the reading literature, automaticity theory (LaBerge & Samuels, 1974), efficiency theory (Perfetti, 1985), or dual processing theory (Goldhammer et al., 2014) posit that as students become better readers, the cognitive reading process becomes more automatic, less effortful, less consciously controlled and thereby less time consuming. Consequently, one would expect that as reading comprehension improves, students would be able to reach correct responses at a more rapid efficiency. A faster comprehension efficiency is not a goal in and of itself, but rather it is evidence that the reading comprehension process is becoming more automatic. For the assessment used in this study, computerized administration makes it feasible to record the total testing time for each student, from which one can compute a students’ comprehension efficiency.

Evaluating Validity via Diagnostic Accuracy

Receiver operating characteristic (ROC) curves have become one of the standards for statistically evaluating the diagnostic accuracy of educational assessments for their utility as screening measures (Smolkowski & Cummings, 2015, 2016). ROC curve analyses test how well a measure predicts performance-level classification on a criterion measure (Silberglitt & Hintze, 2005). ROC curves visually depict the proportion of individuals who *actually belong to a group* who are correctly identified by the screening measure as *being in that group* (i.e., *sensitivity*) relative to the proportion of individuals who do not *actually belong in that group* who are incorrectly designated as *belonging* (i.e., 1 −*specificity*). ROC curves plot (1 − specificity) on the *x*-axis and sensitivity on the *y*-axis, resulting in a curve that begins in the lower left corner where both proportions are zero, and rises toward the upper right corner, where both are 1. A curve that rises steeply toward the upper left corner of the plot indicates a more accurate predictor measure, because it represents a higher proportion of correct classifications relative to incorrect classifications. Measures with poor diagnostic accuracy have an ROC curve close to the diagonal line on the plot, which indicates a 50% probability of correct classification, the probability of assigning a correct classification at random (Swets, Dawes, & Monahan, 2000).

The area under the curve (AUC) statistic serves as an indicator of the overall diagnostic accuracy of an assessment. AUC values range from .50 to 1.0, where higher values indicate higher classification accuracy. An AUC of .50 indicates chance accuracy, and an AUC of 1.0 indicates perfect classification. Traditionally, an AUC between .50 and .70 indicates low accuracy, an AUC of .70 to .90 is good, and an AUC greater than .90 is excellent (Swets, 1988). Alternatively, the AUC value represents the proportion of time a screener correctly identifies individuals with a certain condition (e.g., a learning disability) versus individuals without in a randomly selected pair (e.g., a student with learning disability and a student without learning disability). For example, an AUC of .80 means that in a randomly selected pair of students, one with a disability and one without, the measure correctly identifies the one with a learning disability in 80% of the trials.

In this research, we first examined the added value of mistake subscores over and above the total number of mistakes using Haberman's (2008a) value added analysis. Then we used a series of logistic regression and ROC curve analyses to evaluate the predictive validity of multiple types of subscores from a reading test in the identification of students at-risk for not reaching proficiency on a statewide test in Grades 3 to 5. Our primary hypothesis was that the subscores would add value over and above a total score and that subscores would improve the prediction of proficiency over and above the total score alone, but that individual subscores may not be equally useful at all grades. That is, the more items answered incorrectly, the more information one has about the types of processes (i.e., paraphrase vs. lateral connect) to which the student is prone. Therefore, we hypothesized that incorrect answer propensity scores would be more useful among less proficient students in lower grades (i.e., when students tend to get more items incorrect). In contrast, automaticity theory suggests that automaticity emerges later in the reading development process (Goldhammer et al., 2014; LaBerge & Samuels, 1974; Perfetti, 1985), suggesting that comprehension efficiency may be a better predictor among more proficient students and in later stages of reading development, when individual differences in automaticity are more pronounced. Thus, we hypothesized that comprehension efficiency may be more predictive in fifth grade than in third or fourth grade.

Method

Participants

Participants were 625 elementary students from 13 schools in two school districts in two western states in 2015-2016: 245 in third grade, 210 in fourth grade, and 170 in fifth grade. The sample included 337 (53.9%) females, 240 (38.4%) ethnic minority students, 233 students eligible for free and reduced lunch (FRL; 47.8% of the 487 students for whom FRL data were available), 48 (7.7%) English language learners, and 45 (7.2%) students who received special education services. Our sample was predominantly White (61.6%) and Hispanic (24.5%) with smaller percentages of African American (3.4%), American Indian (1.8%), Asian (5.4%), Hawaiian (1.3%), and two or more races (1.4%). Whites were overrepresented and African Americans were underrepresented but all other groups were represented roughly in proportion to their representation in the K-12 population (U.S. Department of Education, National Center for Education Statistics, 2015).

Assessments

Smarter Balanced (SBAC) English Language Arts (ELA) Assessment. Developed by a consortium of 15 states, the SBAC ELA assessment (Smarter Balanced Assessment Consortium, 2016) is one of the Common Core State Standards aligned measures for Grades 3 to 8. It has two components, a computer-adaptive test and performance tasks that combine traditional assessment questions with interactive activities to assess students' abilities to apply critical thinking and solve problems. In Grades 3 to 5, the SBAC ELA contains between 43 and 47 items related to reading, writing, speaking, and research. The assessment is untimed, but the estimated total testing time is about 3.5 hours. SBAC proficiency was chosen as a criterion because educators, policy makers, and parents are concerned about students achieving proficiency on statewide tests. The assessment was expected to be predictive of SBAC ELA proficiency because both tests are in the area of English language literacy and both cover reading, although reading is only one component of the SBAC ELA assessment.

As the criterion variable, we used district-provided achievement level information on SBAC ELA student classifications, which were available for a greater number of students than scale scores. The SBAC ELA measure provides four achievement levels based on the corresponding scale scores: 1 = not meeting the state ELA achievement standards; 2 = nearly meeting the state ELA achievement standards; 3 = meeting the state ELA achievement standards; and 4 = exceeding the state ELA achievement standards. We classified students at Levels 1 to 2 as not proficient and those at Levels 3 to 4 as proficient. For the analysis, SBAC performance was coded as 1 = not proficient and 0 = proficient.

Multiple-choice Online Causal Comprehension Assessment (MOCCA). The Multiple-choice Online Causal Comprehension Assessment (MOCCA) is a multiple-choice, online assessment designed to identify comprehension processes for students in

Grades 3 to 5. It has nine computer-administered, 40-item forms, with three at each grade level. Participants were randomly assigned to take one of the three forms at their grade level. Each MOCCA item consists of a seven-sentence story in which the sixth sentence is removed. For each item, three response types are presented: a causally coherent inference, a paraphrase, and a lateral connection. The causally coherent inference represents the correct answer (i.e., the original sixth sentence) and indicates full comprehension of the item-story. The other two incorrect responses are used to identify patterns in the types of processes students tend toward when not comprehending fully (i.e., paraphrase or lateral connection). The test itself imposes no time limit, but testing typically occurred during one class period (i.e., 30-60 minutes), with a mean student testing time of 35 minutes across grades.

MOCCA provides a total of six scores of interest to this research, five of which were used as predictors. The first two are the number correct (NC) out of 40 possible items, and its inverse, the number of incorrect responses (NI) out of 40 possible items (i.e., NI = 40 – NC). There are three ways in which a student can fail to answer an item correctly: choosing the paraphrase response option, choosing the lateral connect response option, or failing to reach an item in the time allowed. This leads to three subscores, each with a range of zero to 40: the number of paraphrase responses (NP), the number of lateral connect responses, (NL), and the number of items not-reached (NR). The first five scores are related, such that NI = 40 – NC = NP + NL + NR. In the analyses below, NI and NC are interchangeable, but inversely related measures of a student's overall performance on the test. We emphasize NI rather than NC, because it represents the sum of the three incorrect responses NP + NL + NR, and this summative relationship leads directly to methods for comparing the performance of total score NI vs. the three incorrect response scores (Davison & Davenport, 2002; Davison et al., 2015; Haberman, 2008a). The sixth score of interest is the comprehension efficiency (CE), the student's total testing time divided by the number of correct responses, in number of minutes and seconds.

Procedures

Schools were recruited through local connections and the DIBELS Data System (DDS) at the University of Oregon. Students with parental consent took MOCCA in groups either in their school computer lab or in their classrooms on computers or tablets between February and June of 2016. At the end of the school year, participating districts shared students' state assessment scores for that year. To evaluate our hypotheses regarding the incremental utility of process propensity and comprehension efficiency subscores, we conducted logistic regression and ROC curve analyses for each grade. For the logistic regression analyses, we tested a series of three increasingly complex models aligned with our hypotheses. Our first two models were chosen because comparing the fit of the two models directly addresses the question of whether subscores add to validity over a total score. Model 1 includes only one predictor, the student's total number incorrect (NI). A key feature of Model 1 is that it assigns the same expected probability to every student with the same total score. This provides a baseline estimate of the extent to which the total score is predictive of SBAC outcomes, without the inclusion of any subscale scores. Given the relationship between number correct (NC) and total number incorrect (NI = 40 – NC), it would be redundant to repeat this analysis using NC as the criterion.

Model 2 contains the three incorrect score predictors: NP, NL, and NR. A key feature of Model 2 is that it assigns different expected probabilities to people with the same total score NI but different patterns of incorrect responses. The model of Equation (2) is a hierarchically embedded submodel of Equation (1) in which all of the linear coefficients in Equation (1) are constrained to be equal. That is, if π is the probability of being not proficient for a predictor vector (NP, NL, and NR), the logit for Model 2 can be expressed as:

$$\mathrm{Ln}\left(\frac{\pi}{1-\pi}\right) = b_1\mathrm{NP} + b_2\mathrm{NL} + b_3\mathrm{NR} + a \tag{1}$$

In the hierarchically embedded submodel with all three weights equal to b, Equation (1) becomes:

$$\mathrm{Ln}\left(\frac{\pi}{1-\pi}\right) = b(\mathrm{NP} + \mathrm{NL} + \mathrm{NR}) + a \tag{2}$$

$$= b\mathrm{NI} + a \tag{3}$$

because NI = NP + NL + NR. Equation (3) represents Model 1 with only a single predictor, NI.

Model 3 allows us to test whether taking efficiency into account improves prediction. It represents a further extension of Model 2, adding a fourth predictor to the three incorrect response types in Model 2: comprehension efficiency CE. Model 2 can thus be considered a hierarchically embedded submodel of Model 3 in which the weight for CE is constrained to 0.

Results

Subscores

For subscores to be useful, they must be reliable. Across forms and grades, the reliability (alpha) of the NC (or NI) scores ranged from .93 to .94, the NP reliabilities from .86 to .89, the NL score reliabilities from .72 to .82, and the NR score reliabilities from .96 to .97. The NR reliabilities are almost certainly inflated by a lack of independence of the NR response variable across items at the end of the test, but to date, the only available reliability estimates are internal consistency estimates.

For subscores to be useful, they must also be distinct. For validity purposes, this means that their intercorrelations should not be too high. Lyren (2009) and McPeek, Altman, Wallmark, and Wingersky (1976) suggest an upper limit of .90 for the disattenuated correlations. Sinharay (2010) proposed an upper limit of .80 for the average disattenuated correlation among the subscores, adding that ‘‘it is possible to find unique tests for which these figures do not provide accurate guidance’’ (p. 169). For our subscores, the average (over three forms) disattenuated correlation estimates for NP and NL were .88, .82, and .81 in third, fourth, and fifth grades, respectively. For NP and NR, the corresponding figures were −.35, −.17, and −.51, and for NL and NR, they were −.51, −.22, and −.22.

Added Value of Subscores

Haberman (2008a) proposed an analysis that uses the reliability of a subscore, the reliability of the total score, and the correlation of the subscore with the total score to determine if subscores add value over and above a total score. Since our subscores NP, NL, and NR add to a total incorrect score NI, the analysis can be applied to these subscores. To describe the analysis, consider one of the subscores NP. For each person, one can conceive a true score on a paraphrase propensity dimension manifested in the subscore NP. If one wants to estimate a person's true paraphrase score, there are three ways to estimate that true score: (1) estimate it from the observed NP score, (2) estimate it from the observed total number of errors NI, or (3) estimate it using both NP and NI. In Haberman's analysis, one asks the question: Which of these three methods of estimating the true paraphrase score will yield the most precise true score estimate where precision is measured by the root mean square error (RMSE) of estimation? The method with the smallest RMSE gives the most precise estimate. A subscore can be said to add value if the RMSE estimating with the subscore alone or the subscore in combination with the total score is smaller than the RMSE estimating with the total score alone.

Table 1 shows the RMSE for each method of estimating the true score for each of our three incorrect types in all three grades. For instance, for the paraphrase incorrect type in third grade, the RMSE was 4.57 if the true score is estimated from the total observed score NI, 2.13 if the true score is estimated from the observed subscore NP, and 2.12 if estimated from both NI and NP. Because the RMSE estimating with NP or both NP and NI is smaller than the RMSE for estimating with the total score NI alone, the subscore can be said to add value over the total score. In every grade, the RMSE is smaller for estimation with the observed subscore or with the observed subscore and total score NI than the RMSE for estimation with the total score NI alone. Estimating with the observed subscore yields a smaller RMSE than estimating from the total score and adding the total score to the subscore improves estimation very little. As added value is defined by Haberman, all of our subscores add value over and above the total incorrect score NI at every grade.

Means of Proficient and Nonproficient Students

Table 2 shows the mean scores of proficient and non-proficient students by grade on the six scores described above: NC, NI, NP, NL, NR, and CE. Proficient and non-

Table 1. Value Added Analysis: Root Mean Square Errors (RMSE) for Estimating Subscore True Scores From Observed Total Score, Observed Subscore, and Both.

Grade	Subscore	RMSE from total score	RMSE from subscore	RMSE from both
3	# Paraphrase	4.57	2.13	2.12
3	# Lateral Cnct.	3.30	1.92	1.87
3	# Not Reached	7.40	1.42	1.39
4	# Paraphrase	4.03	1.89	1.88
4	# Lateral Cnct.	3.15	1.76	1.72
4	# Not Reached	6.76	1.36	1.34
5	# Paraphrase	3.27	1.79	1.78
5	# Lateral Cnct.	2.70	1.82	1.68
5	# Not Reached	5.68	1.11	1.11

Note. Lateral Cnct. = lateral connect.

proficient students differ significantly on all six measures at every grade. For the two overall scores, NC and NI, effect sizes are large, ranging between 1.3 and 1.6 in absolute value over the grades. For NP and NL, effect sizes are somewhat smaller, but still generally large in absolute value, ranging from −0.71 to −1.01. The effect sizes for NR are somewhat smaller in absolute value, ranging from −0.33 to −0.74. Of the three types of incorrect responses, the number of items not reached is less highly associated with proficiency at each of the grades. The effect sizes for CE are also relatively large, ranging from -0.64 to -1.02.

Logistic Regression

Table 3 shows the fit measures for the logistic regression models with the dichotomous SBAC proficiency variable as the criterion at each grade. At every grade, the likelihood ratio tests comparing Models 1 and 2 lead to rejection of Model 1, which assigns equal expected probabilities to everyone with the same total score, in favor of Model 2, which distinguishes among students with the same total score based on their pattern of incorrect responses. In addition, the AIC and the BIC are lower for Model 2 than Model 1 in every grade, suggesting that Model 2 fits the data better than Model 1, even after accounting for the two additional parameters in Model 2. As one measure of effect size, Table 3 also contains Nagelkerke's pseudo-R^2 (Nagelkerke, 1991), a measure of the extent to which the addition of parameters improves the prediction of a logistic regression model. We chose to report Nagelkerke's pseudo-R^2 because it ranges between 0.0 and 1.0 and thus has the same range as the familiar R^2. However, unlike the familiar R^2, it does not have a proportion of variance interpretation, and it improves as a function of the likelihood rather than variance accounted for. Using three predictors improves Nagelkerke's pseudo-R^2 by .05 to .10, depending on the grade. These results support the hypothesis that Model 2, with separate

Table 2. Means, Standard Deviations, and Effect Sizes for MOCCA Scores by Grade and Proficiency.

	M		*SD*		
Score	Proficient	Not Proficient	Proficient	Not Proficient	*g*
	Third grade (number proficient = 120; number not proficient = 125)				
NC	26.87	14.96	8.57	7.87	1.44**
NI	13.13	25.04	8.57	7.87	−1.44**
NP	3.28	8.17	4.23	5.48	–0.99**
NL	3.10	6.94	2.54	4.57	−1.03**
NR	6.75	9.93	8.84	10.37	–0.33*
CE	1.46	2.63	1.09	1.70	–0.81**
	Fourth grade (number proficient = 119; number not proficient = 91)				
NC	31.08	16.65	8.84	9.69	1.56**
NI	8.92	23.35	8.84	9.69	−1.56**
NP	1.97	6.37	3.11	5.54	−1.01**
NL	2.23	5.10	2.53	4.21	–0.85**
NR	4.72	11.88	8.38	11.45	–0.73**
CE	1.06	2.71	0.56	3.85	–0.64**
	Fifth grade (number proficient = 101; number not proficient = 69)				
NC	31.89	20.49	8.15	9.56	1.30**
NI	8.11	19.51	8.15	9.56	−1.30**
NP	1.88	4.17	2.63	3.94	–0.71**
NL	1.91	4.55	2.32	3.96	–0.85**
NR	4.32	10.78	7.87	9.81	–0.74**
CE	0.96	1.84	0.40	1.25	−1.02**

Note. MOCCA = Multiple-choice Online Causal Comprehension Assessment; NC = number correct; NI = number incorrect; NP = number of paraphrases; NL = number of lateral connects; NR = number not reached; CE = comprehension efficiency (minutes/correct).
$^{*}p < .05$. $^{**}p < .01$.

subscores, better accounts for the data than the simpler Model 1, with only total incorrect score as a predictor.

As hypothesized, the results for Model 3 vary by grade. In third grade, the likelihood ratio test (Model 2 vs. Model 3) failed to reject ($p \geq .05$) the more parsimonious Model 2, with just response-based subscores. In fourth and fifth grades, however, the likelihood ratio tests led to rejection of Model 2 that does not include comprehension efficiency in favor of the model that does. Although the improvement is small, adding comprehension efficiency improves Nagelkerke's pseudo-R^2 by .01 in fourth grade and .05 in fifth grade. In addition, the AIC and the BIC are lower for Model 3 than Model 2, suggesting that Model 3 better fits the data than either Model 1 or 2, even after taking into account the additional parameter in Model 3.

Table 4 shows the regression weights and their standard errors by grade for Models 2 and 3. For each of the three incorrect response variables, the unit is the same, one item. Thus, for each incorrect response variable, the unstandardized regression weight indicates the amount by which the expected logit increases with a

Table 3. Logistic Regression Fit Statistics, Nagelkerke R^2, and Likelihood Ratio Test (LRT) for Comparing Each Model to the Model Above It.

Model	−2LL	AIC	BIC	R^2	LRT
			Third grade		
Model 1	242.34	246.34	253.34	0.437	
Model 2	212.99	220.99	234.99	0.538	29.35**
Model 3	212.99	222.99	240.50	0.538	0.00
			Fourth grade		
Model 1	197.45	201.45	208.15	0.467	
Model 2	184.15	192.15	205.53	0.521	13.30**
Model 3	178.82	188.82	205.53	0.537	5.25*
			Fifth grade		
Model 1	175.97	179.97	186.25	0.365	
Model 2	166.61	174.61	187.15	0.418	9.36**
Model 3	158.84	168.84	184.52	0.460	7.77**

Note. LL = log likelihood; AIC = Akaike information criterion; BIC, Bayesian information criterion.
*$p < .05$. **$p < .01$.

one-item increase in the predictor. In third grade, the three incorrect response variables are significant at $p < .01$ in both models, but the addition of comprehension efficiency in Model 3 is not. In fourth grade, all of the predictors in both models are significant at $p < .05$. In fifth grade, NP is not significant in either model. In Model 3, CE is significant at $p < .05$, but NR is not.

Out of concern for multicollinearity, variance inflation factors (VIF) were computed for every predictor in each of our models. The largest VIF was 2.32 for NP in the model with four predictors (NP, NL, NR, CR) in third grade. Kutner, Nachtsheim, Neter, and Li (2005, p. 409) suggest that VIF $>$ 10 be taken as evidence for serious multicollinearity.

At the level of the individual student, the choice of model can make a major difference if the assessment is used to identify which students are at risk and in need of remediation. For instance, in the logistic regression analysis and using a predicted probability of .5 as the cut-score separating those who are and are not at risk, 31 or 13% of third graders would be classified differently depending on whether Model 1 or Model 2 probabilities were used. If these probabilities were used in deciding who was eligible for remediation, there would be 31 children for whom eligibility recommendation would depend on the model chosen. In fourth and fifth grade, 22 (10%) and 15 (9%) of students' eligibility recommendations would depend on the model chosen. If the program is an effective program, eligibility represents a high stakes decision for the child.

ROC Curve Analyses

The three panels of Figure 1 show the results of the ROC curve analyses by grade. In the ROC analysis, model predicted probability of being not proficient was used to

Table 4. Unstandardized Logistic Regression Weights with Standard Errors for Models 2 and 3 by Grade.

	Model 2		Model 3	
Variable	*B*	*SE B*	*B*	*SE B*
		Third grade		
NP	0.132**	0.042	0.132**	0.048
NL	0.421**	0.077	0.422**	0.078
NR	0.147**	0.023	0.147**	0.025
CE			–0.003	0.156
		Fourth grade		
NP	0.263**	0.063	0.212**	0.067
NL	0.155*	0.070	0.136*	0.069
NR	0.119**	0.019	0.083**	0.025
CE			0.710*	0.347
		Fifth grade		
NP	0.112	0.074	0.040	0.080
NL	0.310**	0.083	0.264**	0.086
NR	0.112**	0.021	0.045	0.032
CE			1.495*	0.595

Note. Dependent variable was coded 1 = not proficient, 0 = proficient. NP = number of paraphrases; NL = number of lateral connects; NR = number not reached; CE = comprehension efficiency (minutes/correct).
*p < .05. **p < .01.

predict actual not proficient status. In all grades, the AUC exceeds .80 for every model, indicating that, for MOCCA, the total score alone provides a relatively accurate diagnosis of SBAC proficiency. In third grade, the AUC for Model 1 was .84 (CI = .79-.89), the AUC for Model 2 was .88 (CI = .84-.92), and the AUC for Model 3 was .88 (CI = .84-.92). In fourth grade, the AUC for Model 1 was .86 (CI = .80-.91), the AUC for Model 2 was .88 (CI = .83-.92), and the AUC for Model 3 was .88 (CI = .84-.93). In fifth grade, the AUC for Model 1 was .81 (CI = .75-.88), the AUC for Model 2 was .83 (CI = .77-.90), and the AUC for Model 3 was .85 (CI = .79-.91).

The addition of subscore data consistently improved the accuracy of the diagnosis, both across grades and across the distribution of student skill, as measured by both the AUC statistic and a visual inspection of the curve for each model. That is, in each grade, there are large sections of the plot where the curve for Model 2 is clearly above the curve for Model 1. In third grade, the curves for Models 2 and 3 are nearly identical, emphasizing that the addition of comprehension efficiency in third grade does not improve the model. In contrast, the plots for both fourth and fifth grade have areas (i.e., ranges of student scores) where the curve for Model 3 is clearly above the curve for Model 2.

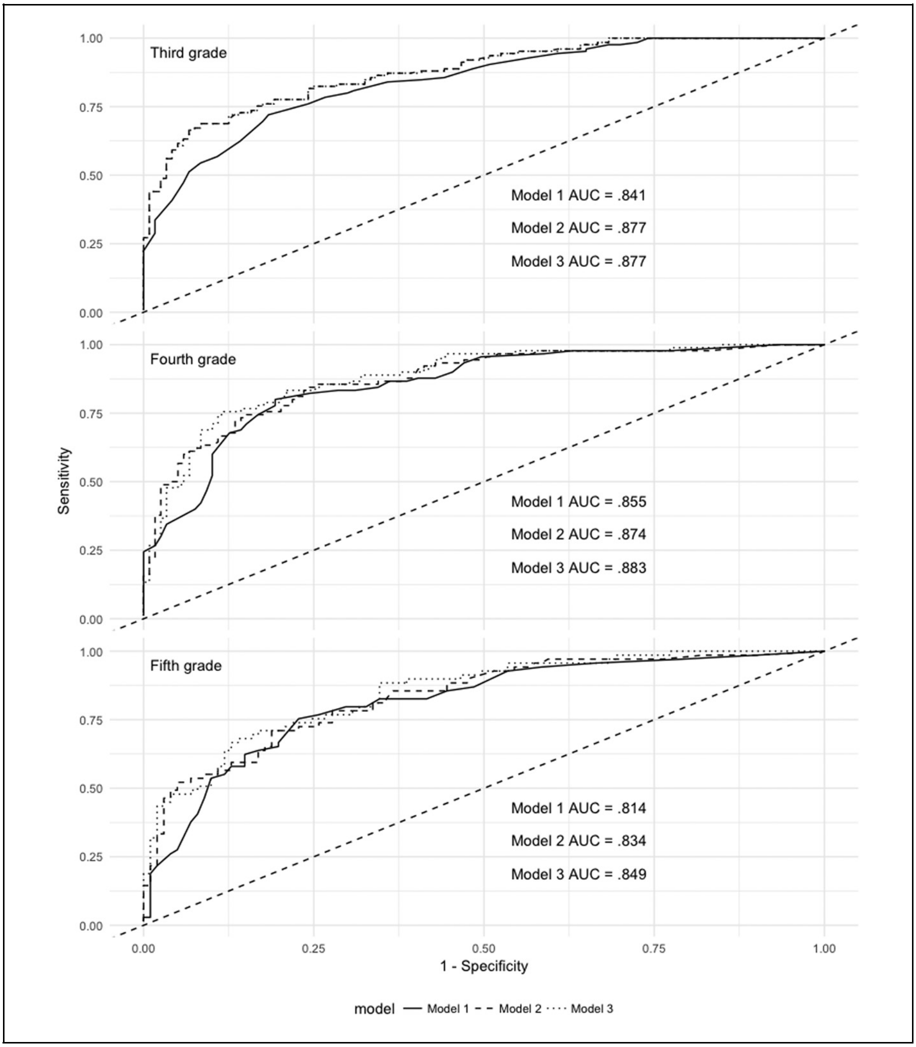

Figure 1. Receiver operating characteristic (ROC) curve results for three models in each of three grades.

Model Differences for Individual Students

The impact of Model 1 versus Model 2 differences for individual students is best illustrated by examining the expected probability for a group of students with the same overall score and different incorrect response patterns. Table 5 illustrates this difference for 13 third grade students, all of whom had the same number of items correct

Table 5. Subscores, Model Probabilities, and Proficiency for 13 Third Grade Students With the Same Overall Score.

Subscore							Model 1	Model 2		Model 3	
NC	NI	NP	NL	NR	CE	Actual class.	Prob.	Prob.	At risk	Prob.	At risk
17	23	0	0	23	1.01	0	.6473	.3409	0	.3415	0
17	23	0	2	21	0.73	0	.6473	.4726	0	.4735	0
17	23	20	3	0	2.86	0	.6473	.4681	0	.4683	0
17	23	1	3	19	1.86	1	.6473	.5376	1	.5376	1
17	23	2	5	16	1.53	1	.6473	.6650	1	.6653	1
17	23	17	6	0	3.94	0	.6473	.6771	1	.6765	1
17	23	15	8	0	1.94	1	.6473	.7891	1	.7896	1
17	23	14	9	0	1.79	1	.6473	.8333	1	.8337	1
17	23	6	10	7	1.83	1	.6473	.8809	1	.8811	1
17	23	11	12	0	1.58	1	.6473	.9225	1	.9228	1
17	23	11	12	0	1.60	1	.6473	.9225	1	.9228	1
17	23	9	14	0	5.10	1	.6473	.9551	1	.9547	1
17	23	8	15	0	2.36	1	.6473	.9660	1	.9660	1

Note. CE = comprehension rate; NC = number correct; NI = number incorrect; NP = number of paraphrase responses chosen; NL = number of lateral connection options chosen; NR = number of items not reached; Actual class = observed proficiency classification (0 = not at risk, 1 = at risk) based on Smarter Balanced Assessment Consortium (SBAC) performance; Prob. = model estimated probability of student risk status; At risk = model predicted risk status based on subscale scores.

(17, in Column 1), and the same number of incorrect answers (23, in Column 2). However, these students differ in their patterns of incorrect responses, and they have been rank ordered from lowest to highest based on their number of NL responses, the predictor with the largest weight in third grade. Column 7 shows the Model 1 predicted probability of being at -risk (.6473) for each student, all of which are the same, because for the only predictor in Model 1 (i.e., NI), these students all have the same score, 23. Because this predicted value is greater than .5, the logistic regression model classified all 13 in the at-risk group.

Column 8 shows the predicted probabilities for Model 2. Unlike the previous column, they are not equal, and range from .3409 to .9660. Because the predicted probability of risk was less than .5 for three of the 13, three students were classified as not at risk by Model 2, but not Model 1. All three students did in fact score at the proficient level on the SBAC, making the not-at-risk designation the correct classification. If one were placing these students in an intervention on the basis of these two models, the results would be very different for these three students. Similarly, Column 10 shows the predicted probabilities for Model 3, which range from .3415 to .9960. Because CE did not improve the model in third grade, the same students are predicted to be at risk by Model 2 and Model 3. As illustrated here, Model 1 assigns equal at-risk probabilities to every student with the same total score. Models 2 and 3 distinguish between students with the same total score based on their pattern of incorrect responses. In this example, ranking students by their NL scores ranks them from low to high on their Model 2 predicted at-risk probabilities, because the NL responses have the highest regression weight. In all grades, given equal number incorrect scores, students with a predominance of NR incorrect responses were at comparatively small risk. Especially in third and fifth grades, those with a predominance of NL incorrect responses were at highest risk. This can be seen in Table 5 by comparing predicted probabilities for those with a strong predominance of NR responses (first two rows) with the probabilities for those having a predominance of LC responses (last two rows).

Discussion

The simple means presented in Table 2 make it clear that proficient and nonproficient students (as measured by SBAC) differed in multiple ways on their performance on MOCCA. They differed in how many items they answered correctly, in their selection of both types of incorrect responses, and in the numbers of items they failed to reach. Moreover, they differed not only in how accurately they answered questions and the incorrect responses they preferred but also in the efficiency with which they correctly answered (i.e., the number of minutes per correct item). This latter finding is consistent with the hypothesis that as reading improves, it becomes more automatic, and with automaticity comes faster efficiency (LaBerge & Samuels, 1974; Logan, 1997; Perfetti, 1985; Samuels, Ediger, Willcutt, & Palumbo, 2008; Samuels & Flor, 1997). This latter finding is also consistent with earlier ones on oral reading comprehension rate (e.g., Neddenriep, Hale, Skinner, Hawkins, & Winn, 2007; Skinner, Neddenriep, Bradley-Klug, & Ziemann, 2002; Skinner et al., 2009). However, the contribution of comprehension efficiency, over and above that of the incorrect response scores, is small as evidenced by the similarity of the curves for Models 2 and 3 in Figure 1.

Including subscores in prediction models, rather than just an overall score, had three effects. First, it led to models that better accounted for the data. Compared to the total score only model, including the three incorrect response scores resulted in a better fit: lower AIC and BIC values, a higher Nagelkerke pseudo-R^2, and a significant likelihood ratio test in all three grades. These findings support the conclusion that, given carefully constructed alternatives, the pattern of student incorrect responses can add information over and above that provided by the total score alone. These subscores indicate that students with the same total score may not be equally at-risk. As evidenced by the ROC graphs, incorrect responses were especially informative in third grade, where students tend to make more mistakes overall.

Similarly, adding comprehension efficiency improved prediction and model fit in fourth and fifth grade: lower AIC and BIC values, as well as significant likelihood ratio tests. That this finding did not hold for third grade seems consistent with automaticity theory, which posits that the reading comprehension process is initially slow and controlled, but with practice becomes more rapid and automatic (LaBerge &

Samuels, 1974; Logan, 1997; Perfetti, 1985; Samuels et al., 2008; Samuels & Flor, 1997). However, automaticity theory is not clear about when the transition to more automatized processing occurs. Practically important differences in automaticity may not occur until somewhat later in reading comprehension development, in which case, comprehension efficiency might not be a good predictor until somewhat later in the developmental process. Second, adding the incorrect response scores changed the ROC curves. In third grade, the ROC curve for Model 2 is higher than that for Model 1 at almost every point on the curve. In fourth and fifth grade, the ROC curve is, with few exceptions, as high or higher at every point on the curve. The AUC is lowest for Model 1 at every grade. Judging by the ROC curves, models with more than a single predictor seemed to do as well as or better than those with only a total score at almost all points along the sensitivity continuum.

The third effect of including subscores occurred at the level of the individual student. When an intervention is highly effective, the decision to provide an intervention (or not) can have a large impact on student outcomes. In the grades evaluated here, the decision of whether to provide an intervention changed for approximately 10% to 15% of students when based on a single total score rather than the pattern of incorrect responses. Adding comprehension efficiency had a smaller impact, but for a given student, the consequences could be substantial.

Investigating Subscores

Just as reliability limits the validity of a single test, the reliability of the total score and its subscores limits the incremental validity of subscores, and Haberman (2008b) derives an estimate of that upper limit. However, we were not interested in the upper limit of subscore incremental validity, but rather the actual increment to validity for a specific criterion variable. Davison et al. (2015) describe a linear model for this purpose given a continuous criterion variable. Here we illustrated how their procedure can be extended to a categorical criterion variable based on a logistic model. Our analysis also illustrates how, given the linear relationship between number correct and number incorrect, the procedures of both Davison et al. (2015) and Haberman (2008a) can be extended to an analysis of subscores based on incorrect responses. Most importantly, we have described a method of constructing incorrect responses to multiple-choice items that, at least in this case study, yielded subscores with added value and validity

Although the current study is a case study of a single assessment, it does point to the possibility of constructing subscores that add information over and above a total score by using carefully constructed distractors each corresponding to a cognitive process leading to an identifiable processing preference. It also suggests that efficiency measures may add information over and above a single total score. Efficiency measures may be particularly valuable in the case of computer-administered assessments where testing times can be recorded with no additional effort on the part of the test administrator. For individually administered tests or paper/pencil tests, it may be impractical to compute efficiency indices. In addition, the use of incorrect answer patterns and efficiency scores may also apply beyond the assessment of reading. Like the conclusions from any single study, however, these suggestions must be viewed with caution pending replication with other reading assessments and in other domains.

Authors' Note

The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305A140185 to the University of Oregon.

ORCID iD

Mark L. Davison https://orcid.org/0000-0003-3656-9672

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Carlson, S. E., Seipel, B., & McMaster, K. (2014). Development of a new reading comprehension assessment: Identifying comprehension differences among readers. *Learning and Individual Differences*, *32*, 40-53.

Davison, M. L., & Davenport, E. C., Jr. (2002). Identifying criterion-related patterns of predictor scores using multiple regression. *Psychological Methods*, *7*, 468-484.

Davison, M. L., Davenport, E. C., Jr., Chang, Y.-F., Vue, C. K., & Su, S. (2015). Criterion-related validity: Assessing the value of subscores. *Journal of Educational Measurement*, *52*, 263-279.

Davison, M. L., Jew, G., & Davenport, E. C., Jr. (2014). Patterns of SAT scores, choice of STEM major, and gender. *Measurement and Evaluation in Counseling and Development*, *47*, 118-126.

delMas, R., Garfield, J., Ooms, A., & Chance, B. (2007). Assessing students' conceptual understanding after a first course in statistics. *Statistics Education Research Journal*, *6*(2), 28-58.

Goldhammer, F., Naumann, J., Stelter, A., Tóth, K., Rölke, H., & Kleine, E. (2014). The time on task effect in reading and problem solving is moderated by task difficulty and skill: Insights from a computer-based large-scale assessment. *Journal of Educational Psychology*, *106*, 608-626.

Graesser, A. C., Singer, M., & Trabasso, T. (1994). Constructing inferences during narrative text comprehension. *Psychological Review*, *101*, 371-395. doi:10.1037/0033-295x.101.3.371

Haberman, S. J. (2008a). When can subscores have value? *Journal of Educational and Behavioral Statistics*, *33*, 204-229.

Haberman, S. J. (2008b). *Subscores and validity* (ETS RR-08-64). Princeton, NJ: ETS.

Hermann-Abell, C. F., & DeBoear, G. E. (2011). Using distractor-driven standards-based multiple-choice assessments and Rasch modeling to investigate hierarchies of chemistry misconceptions and detect structural problems with individual items. *Chemistry Education Research and Practice*, *12*, 184-192. doi:10.1039/C1RP90023D

Hestenes, D., Wells, M., & Swackhamer, G. (1992). Force concept inventory. *The Physics Teacher*, *30*, 141-158. doi:10.1119/1.2343497

Kintsch, W., & van Dijk, T. A. (1978). Toward a model of text comprehension and production. *Psychological Review*, *85*, 363-394. doi:10.1037/0033-295X.85.5.36.

Kutner, M. H., Nachtsheim, C. J., Neter, J., & Li, W. (2005). *Applied linear statistical models* (5th ed.). New York, NY: McGraw-Hill.

LaBerge, D., & Samuels, S.J. (1974) Toward a theory of automatic information processing in reading. *Cognitive Psychology*, *6*, 293-323.

Logan, G. D. (1997), Automaticity of reading: Perspectives from the instance theory of automatization. *Reading and Writing Quarterly*, *13*, 123-146.

Lyren, P. (2009). Reporting subscores for college admissions tests. *Practical Assessment, Research, and Evaluation*, *14*, 1-10.

McMaster, K. L., Espin, C. A., & van den Broek, P. (2014). Making connections: Linking cognitive psychology and intervention research to improve comprehension of struggling readers. *Learning Disabilities Research & Practice*, *29*, 17-24.

McMaster, K. L., van den Broek, P., Espin, C. A., White, M. J., Rapp, D. N., Kendeou, P., . . . Carlson, S. (2012). Making the right connections: Differential effects of reading intervention for subgroups of comprehenders. *Learning and Individual Differences*, *22*, 100-111.

McPeek, M., Altman, R., Wallmark, M., & Wingersky, B. C. (1976). *An investigation of the feasibility of obtaining additional subscores on the GRE Advanced Psychology Test* (GRE Board Professional Report No. 74-4). Princeton NJ: ETS. (ERIC Document No. ED163090)

Nagelkerke, N. J. D. (1991). A note on a general definition of the coefficient of determination. *Biometrika*, *78*, 691-692.

Neddenriep, C. E., Hale, A. D., Skinner, C. H., Hawkins, R. O., & Winn, B. D. (2007). A preliminary investigation of the concurrent validity of reading comprehension rate: A direct, dynamic measure of reading comprehension. *Psychology in the Schools*, *44*, 373-388. doi:10.1002/pits.20228

Perfetti, C. A. (1985). *Reading ability*. London, England: Oxford University Press.

Puhan, G., Sinharay, S., Haberman, S., & Larkin, K. (2010). The utility of augmented subscores in a licensure exam: An evaluation of methods using empirical data. *Applied Measurement in Education*, *23*, 266-285.

Rapp, D. N., van den Broek, P., McMaster, K. L., Kendeou, P., & Espin, C. A. (2007). Higher-order comprehension processes in struggling readers: A perspective for research and intervention. *Scientific Studies of Reading*, *11*, 289-312.

Sadler, P. M. (1998). Psychometric models of student misconceptions in science: Reconciling qualitative studies and distractor-driven assessment instruments. *Journal of Research in Science Teaching*, *35*, 165-396.

Samuels, S. J., Ediger, K.-A. M., Willcutt, J. R., & Palumbo, T. J. (2008). Role of automaticity in metacognition and literacy instruction. In S. E. Israel, C. C. Block, K. L. Bauserman, & K. Kinnucan-Welsch (Eds.), *Metacognition in literacy learning: Theory, assessment, instruction, and professional development* (pp. 41-59). New York, NY: Routledge.

Samuels, S. J., & Flor, R.F. (1997). The importance of automaticity for developing expertise in reading. *Reading & Writing Quarterly*, *13*, 107-121.

Silberglitt, B., & Hintze, J. (2005). Formative assessment using CBM-R cut scores to track progress toward success on state-mandated achievement tests: A comparison of methods. *Journal of Psychoeducational Assessment*, *23*, 304-325.

Sinharay, S. (2010). How often do subscores have added value? Results from operational and simulated data. *Journal of Educational Measurement*, *47*, 150-174.

Skinner, C. H., Neddenriep, C. E., Bradley-Klug, K. L., & Ziemann, J. M. (2002). Advances in curriculum-based measurement: Alternative rate measures for assessing reading skills in pre- and advanced readers. *The Behavior Analyst Today*, *3*, 270-281.

Skinner, C. H., Williams, J. L., Morrow, J. A., Hale, A. D., Neddenriep, C. E., & Hawkins, R. O. (2009). The validity of reading comprehension rate: Reading speed, comprehension, and comprehension rates. *Psychology in the Schools*, *46*, 1036-1047.

Smarter Balanced Assessment Consortium. (2016). *2014-2015 Technical report*. Retrieved from http://portal.smarterbalanced.org/library/en2014-15.

Smolkowski, K., & Cummings, K. D. (2015). Evaluation of diagnostic systems: The selection of students at risk of academic difficulties. *Assessment for Effective Intervention*, *41*, 41-54.

Smolkowski, K., & Cummings, K. D. (2016). Evaluation of the DIBELS diagnostic system for the selection of native and proficient English speakers at risk of reading difficulties. *Journal of Psychoeducational Assessment*, *34*, 103-118.

Swets, J. A. (1988). Measuring the accuracy of diagnostic systems. *Science*, *240*, 1285-1293.

Swets, J. A., Dawes, R. M., & Monahan, J. (2000). Psychological science can improve diagnostic decisions. *Psychological Science in the Public Interest*, *1*, 1-26.

U.S. Department of Education, National Center for Education Statistics (2015). *Common core of data. State Nonfiscal Public Elementary/Secondary Education Survey data*. Retrieved from https://nces.ed.gov/ccd/stnfis.asp

van den Broek, P., McMaster, K., Rapp, D. N., Kendeou, P., Espin, C., & Deno, S. (2006, June). Connecting cognitive science and educational practice to improve reading comprehension. Paper presented at the Institute of Education Sciences Research Conference, Washington, DC.