Please quote from or cite the published version: Liu, B., Kennedy, P. C., Seipel, B., Carlson, S. E., Biancarosa, G., & Davison, M. L. (2019). Can we learn from student mistakes in a reading comprehension assessment? *Journal of Educational Measurement, 56,* pp. 815 – 835. doi.org/10.1111/jedm.12238

Can We Learn from Student Mistakes in a Formative, Reading Comprehension Assessment?

Bowen Liu, University of Minnesota
Patrick C. Kennedy, University of Oregon
Ben Seipel, University of Wisconsin, River Falls and California State University, Chico
Sarah E. Carlson, Georgia State University
Gina Biancarosa, University of Oregon
Mark L. Davison, University of Minnesota

Correspondence concerning this manuscript should be addressed to Mark L. Davison, Department of Educational Psychology, University of Minnesota, 56 E. River Rd., Minneapolis, MN 55455. The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305A140185 to the University of Oregon. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

Abstract

This paper describes an on-going project to develop a formative, inferential reading comprehension assessment of causal story comprehension. It has three features to enhance classroom use: equated scale scores for progress monitoring within and across grades, a scale score to distinguish among low-scoring students based on patterns of mistakes, and a reading efficiency index. Instead of two response types for each multiple-choice item, correct and incorrect, each item has three response types: correct and two incorrect response types. Prior results on reliability, convergent and discriminant validity, and predictive utility of mistake subscores are briefly described. The three-response-type structure of items required re-thinking the IRT modeling. IRT-modeling results are presented, and implications for formative assessments and instructional use are discussed.

Can We Learn from Student Mistakes in a Formative Reading Comprehension Assessment?

Thorndike and Thorndike-Christ (2010, p. 68) define formative evaluation as assessment to guide future classroom instruction. In what follows, we describe the construction of an inferential reading comprehension assessment that has three characteristics designed to help guide classroom instruction on reading comprehension: multiple, equated forms for monitoring student progress over time; diagnostic scores describing a low-scoring student's predominant incorrect answer type (if there is one), and a reading rate score to monitor the development of reading efficiency (i.e., automaticity, comprehension fluency).

The test measures story causal sequence comprehension and is designed to be administered at one or more points before or during the instructional process. Results can then be used to design classroom lessons and individualize student instruction. For instance, the test might be administered at the beginning of the school year. If students generally seem to display a predominant type of error, the teacher may want to increase the use of instructional strategies to address that form of mistake, for instance questioning strategies such as those described in McMaster et al. (2012, 2014) or Rapp (2007). Or, if the assessment indicates that a particular student reads inefficiently, the teacher may want to design reading activities to improve the efficiency. While the test can also be administered as an outcome or screening measure, it is designed to be administered as a pre-test with subscores that can be used in data-based design and individualization of instruction. It is also designed to be administered in the midst of instruction for mid-course modification of instruction at the classroom or individual student level. Indeed, many of the test's innovative features are designed to inform instruction and cannot be fully utilized if the assessment is given only as an outcome measure post-instruction.

This paper begins with a description of the test itself and the literature on which we based its development. We present new data used to select IRT models appropriate for the data and purposes of the assessment. The selection of an IRT model was complicated by design decisions made to make the test diagnostic of student mistake patterns. Lastly, we discuss implications for the development of classroom use, as well as standardized and formative assessments.

MOCCA. The test is the Multiple-choice Online Causal Comprehension Assessment (MOCCA). There are nine 40-item forms of MOCCA, three each in grades three, four, and five. Each MOCCA item is a seven-sentence narrative story with a causal sequence organized around a goal structure, in which the sixth sentence is missing from the story. Readers are prompted to choose a sentence that best fits where the missing sentence is in the sequence of the story. Rather than retrofitting an existing test for diagnostic purposes using an IRT model or constructing the test to comply with the assumptions of an existing IRT model, the goal was to construct the test for diagnostic purposes from the start and then choose or adapt an IRT model to the data and the intended assessment purposes.

In its construction, MOCCA differs from other reading comprehension assessments in at least three respects. First, it is administered online using tablets or computers. This enables us to precisely measure item response times with the goal of using those response times to monitor student progress toward reading efficiency. Second, drawing from the curriculum-based measurement (CBM) literature (e.g., Van Norman, Christ, & Newell, 2017; Deno, 1985), it uses a modified maze task. In the familiar maze reading task, students read a sentence with a missing word, and are asked to select the response that best fills-in the missing word. CBM tasks are good for repeated measures of student progress within an academic year, whereas published standardized achievement assessments are not (Shin, Deno, & Espin, 2000). However, traditional

maze tasks may only be good at measuring sentence-level processing and comprehension (January & Ardoin, 2012). As illustrated in Figure 1, MOCCA uses a similar maze approach, except that students are asked to select a sentence from three alternatives that best fills in the sentence missing from the paragraph. This story-level maze approach requires discourse-level processing rather than simple word integration (i.e., semantic analysis; Kintsch & Rawson, 2005) at a sentence level. Third, whereas most multiple-choice tests have two kinds of answers, correct and incorrect, each MOCCA item (story) has three kinds of responses, one correct answer and two types of incorrect answers: paraphrase and lateral connection. For students who make a number of mistakes, one can assess whether the student exhibits a predominant incorrect response pattern when choosing the sentence choices to complete the missing sentence.

The correct answer, termed the causal coherent response, is the response that best completes the causal sequence of the story. The first type of incorrect response, paraphrase, simply paraphrases prior information (generally the goal, subgoal, or a combination of the two) without advancing the story or its causal sequence. The second type of incorrect response, the lateral connection response, is an elaboration of, evaluation of, or association with information in the story. That is, the response goes beyond the information in the story but does not complete the causal sequence. It may be an inference, and it may be accurate, but it does not fully complete the story (i.e., there is still a causal gap in the story). Paraphrase and lateral connections are different styles of incorrect responding. In the example item shown in Figure 1, the main character's goal is to go to the store with her dad. Moving down the alternatives, the responses represent the lateral connection, paraphrase, and causal coherent (correct) answers, respectively. The paraphrase response type simply restates the goal of the story; the lateral connection response type moves beyond the information in the story with an inference, but there is still

missing causal information; and finally, the causally coherent response type shows how her choice completes the story in a causal way because she is happy at the end.

Equating. In the test construction process, one goal was to develop three forms at each grade, each with an overall comprehension scale score equated within and across grades so that teachers could monitor student comprehension skill within grades and across grades without administering a given form of the test more than once. Our plan was to use a familiar IRT equating design (Lord, 1980; Kolen & Brennan, 2014) and an IRT model consistent with the three-response type structure of items. Furthermore, in addition to the overall comprehension score, the test design required a score that could be used in classifying low-scoring students by their predominant incorrect answer type, if there was one. In the research reported below, we compared IRT models for a dimension of overall comprehension accuracy that could be used to equate forms within and across grades, and a second dimension that could be used to classify students by the predominant incorrect answer choice, where applicable.

Incorrect Alternatives. Our two types of incorrect responses were drawn from think-aloud research on inferential reading comprehension (e.g., Coté, Goldman, & Saul, 1998; McMaster et al., 2012; Trabasso & Magliano, 1996a, 1996b; Wolfe & Goldman, 2005). In think-aloud tasks, students identified as poor comprehenders using criterion measures have been found to have a tendency to rely either on paraphrase or elaboration processes that correspond to our paraphrase and lateral connection response types. In the think-aloud research, many researchers use the term “elaborations,” rather than “lateral connection”. In our work, however, we use the term lateral connection because the lateral connection options include responses, such as associations or evaluations that involve judgements and inferences that go beyond simple elaborations.

Importantly, neither paraphrases nor lateral connections are incorrect in an absolute sense. Both are processes that facilitate comprehension, and in the context of literal comprehension, the correct answer would be a paraphrase. Rather, the paraphrase and lateral connection processes, despite being generally supportive of comprehension, are incorrect in the context of MOCCA because they do not provide the necessary information to close the causal gap of the story.

There is research indicating that in classroom instruction, poor comprehenders who predominantly paraphrase the text ("paraphrasers") and those who predominantly make lateral connections ("lateral connectors") respond differently to instruction (McMaster et al., 2012; Rapp et al., 2007). In these studies, paraphrasers benefitted more from a questioning strategy emphasizing general connection making (e.g., "Make a connection to what you previously read."), whereas lateral connectors benefitted more from a questioning strategy more narrowly focused on causal connections (e.g., "Why was Janie happy?"). However, a more recent study using small group instruction did not replicate these earlier results, perhaps because students were receiving optimal feedback about their understanding or lack of understanding of the text (McMaster, Espin, & van den Broek, 2014).

The two types of poor comprehenders identified in think-aloud research demonstrate fundamentally different approaches to comprehension as a process with the paraphrasing poor comprehenders exhibiting a tendency to be overly reliant on the text for meaning and the lateral connection poor comprehenders exhibiting a tendency to indiscriminately make elaborations about the text. MOCCA was developed to identify such tendencies in a more efficient manner.

Structuring incorrect alternatives around common mistakes or misconceptions is hardly new (e.g., Delmas, Garfield, Ooms, & Chance, 2007; Hermann-Abell & DeBoear, 2011;

Hestenes, Wells, & Swackhamer, 1992; Sadler, 1998). Assessments that do so have been called distractor-driven assessments (Hestenes et al., 1992) or concept inventories (Sadler, 1998), and many of these assessments are found in the sciences (e.g., the *Force Concept Inventory,* Hestenes et al., 1992; the *Genetics Concept Assessmen*t; Smith, Wood, & Knight, 2008). However, these inventories generally contain examples of many different types of misconceptions, all of which appear as an option for only a few items, so the inventories do not yield reliable scores pertaining to any particular misconception. In contrast, MOCCA focuses on only two types of mistakes, paraphrase and lateral connection; includes both of these mistake types as an option for every item; and yields a reliable score for each response type: the number of items for which a paraphrase was chosen, and the number of items for which a lateral connection was chosen.

Automaticity. MOCCA has also been influenced by the literature on reading automaticity, also called efficiency, fluency, or dual processing (Goldhammer, Naumann, Stelter, Tóth, Rölke, & Klieme, 2014; Laberge & Samuels, 1974; Perfetti, 2010; Perfetti & Lesgold, 1979; Posner & Snyder, 1975; Samuels & Flor, 1997). The National Reading Panel (Rand, 2002) defined fluency in terms of accuracy, appropriate rate, and good expression. While the definition refers to appropriate rate, rather than a fast rate per se, measures of fluency use scores such as correct words per minute in which faster is better, other things being equal (e.g., Cianco et al., 2015; Hale et al., 2011; McCane-Bowling et al., 2014; Skinner et al., 2002; Skinner et al., 2009). As a result, we were interested in whether response rate on comprehension items answered correctly could be used to monitor progress in attaining automaticity. We believe reading automaticity is necessary for purposes of reading to learn so that reading processes do not interfere with attention to content. While our progress on measures of automaticity is limited, a preliminary study showed that a measure of correct response rate was reliable (marginal

reliability of .87 - .90 depending on form and grade), and that it added modestly over and above response accuracy to prediction of which students will and will not attain proficiency on a statewide test (Biancarosa, et al., in press) in fourth and fifth grades, but not third.

Item/Story Development

Of the more than 500 stories written, 480 were selected for the pilot phase. All stories were reviewed for cultural and developmental appropriateness, among other things, by an external panel of six teachers who worked with Grades 3-5, including a special education teacher and a Title 1 specialist from a Spanish-English dual-language school. Items flagged by the teachers were reviewed and revised or dropped, with fewer than a dozen being dropped. Stories were then selected to balance forms within grade by readability as measured by Flesch-Kincaid Grade Level (Kincaid, Fishburne, Rogers, & Chissom, 1975) and other story features such as the gender of the main character, the explicitness of the goal in a story, and whether the end of the story satisfied the main goal or not.

In the first year of the study, pilot data were collected in 23 schools and five districts from two states with third, fourth, and fifth grade students (n = 360, n = 307, n = 263 respectively). Results demonstrated that although there were differences in mean performance by ethnicity and gender, very few items demonstrated evidence of differential item functioning (DIF), suggesting little evidence of potential bias in the test items. As a result, 10 of the 480 piloted items/stories demonstrating DIF were dropped. No apparent causes (e.g., the content of the story) could be discerned as an obvious reason behind the DIF of these items. Also important to note is that story statistics generated through Cohmetrix analyses (Graesser, McNamara, & Kulikowich, 2011) (H), such as multiple readability formula estimates and vocabulary load indices (e.g., lexical diversity, age of acquisition, polysemy), did not correlate with proportion

correct, indicating that test performance was unlikely predominantly a function of decoding or vocabulary ability. As a final validity check, classifications of poor comprehenders as paraphrasers and lateral connectors were triangulated with think-aloud data for a diverse subsample of students from one district. Results suggested that MOCCA was identifying the two poor comprehender types well.

Item statistic results from the pilot study were then used to revise the remaining 470 stories, predominantly with a focus on ensuring that lateral connections were consistent with the final emotion of the story and the paraphrases were consistent with any updating of the original goal. By design, we had more items than necessary. Thus, of the 480 stories piloted, we retained 360 to allow for three forms of 40 items per grade level. Forms were again constructed to balance readability and story features across forms within grade, but also with a new focus on balancing for difficulty as measured by proportion correct.

Pilot Research Results

Reliability. Simple reliabilities, alpha, have been good to excellent for the raw number correct (NC) score and the number paraphrase (NP) score, but lower for the number lateral connection (NL) score. In year 1 pilot data, reliabilities for the NC score ranged from .92 to .95 across grades and forms. Those for the NP score ranged from .71 - .89, and those for the NL score ranged from .49 - .74. In year 2 field test data, the NC score alphas ranged from .92 - .94, NP scores from .86 - .89, and NL scores from .72 - .82. While the scores based on incorrect answers have lower internal consistency reliabilities, perhaps due in part to their more restricted variances, nevertheless the NP score showed good to excellent reliabilities in both years and the NL score had consistently good reliabilities, at least in year 2.

Do Incorrect Responses Matter? Having shown that incorrect answer scores can be reliable, we turned to the question of whether those scores provide additional information not available from a simple NC score. To address this question, Biancarosa et al. (in press) employed a logistic regression analysis of incorrect answer profiles (Davison, Davenport, Chang, Vue, & Shiyang, 2015). For purposes of this analysis, students could be scored as incorrect for one of three reasons: a paraphrase response, a lateral connection response, or not completing the item. Using a subset of the Year 2 field test data for which statewide test data was available (*Smarter Balanced Assessment Consortium*, 2017), two logistic regression models were fit within each grade (Biancarosa, in press). The criterion variable was the same for both models: a dichotomous indicator of whether the student reached proficiency on the statewide exam. In Model 1, there was only one predictor, the total score. In Model 2, there were three predictors, the three incorrect answer scores: *NP*, *NL*, and not-reached (*NR*). In all three grades, Model 2, with the mistake types as predictors, fit the data significantly better ($p < .05$) than Model 1 with only the total score as the predictor.

Areas under (AUC) receiver operating characteristic (ROC) curves generated from Model 1 and Model 2 predicted probabilities exceeded .8 for all grades and both models, and in each grade, the ROC for Model 2 was as high or higher than that for Model 1 at almost every level of specificity, with the exception of the extreme ends of the curve. That is, holding specificity constant, the sensitivity was almost always as high or higher for Model 2 than Model 1, although the differences were most notable in third grade. These results suggest that the student profile of mistakes in Model 2 (*NP*, *NL*, and *NR*) carries information that can improve model fit and prediction over and above that contained in the total score (Model 1). Further, in fourth and fifth grade, but not third grade, it was found that an index of rate, minutes per correct response,

improved model fit over and above the number correct score *and* the incorrect answer propensity scores.

Convergent and Discriminant Validity. Using Year 1 data, Davison et al. (2018) found that, in seven samples ranging from 36 – 112 students, MOCCA was significantly correlated with other standardized reading tests. Even though MOCCA is focused on inferential, story causal sequence comprehension, it is correlated with other reading and language arts tests with a broader content coverage. Furthermore, in those same samples, MOCCA was more highly correlated with reading test scores than with math test scores, although it was consistently correlated significantly with the math scores as well. The evidence in these analyses supports the convergent and discriminant validity of MOCCA.

In the current study, two research questions were examined: What is the best item response model on which to base a measure of overall accuracy for the purpose of equating accuracy scores across forms and grades, and what is the best item response model on which to base an index indicating the student's predominant error type, if indeed the student has one?

Methods

Participants

The sample was a national convenience sample from 59 schools in 32 districts and 14 states, including 1,577 students in third grade, 1,498 students in fourth grade, and 1,215 students in fifth grade. Across grades and forms, the sample was 51% male, 10% English language learners, 51% free and reduced meal status, and 11% special education students. In ethnicity, 7% Black, 3% Asian, 23% Hispanic, and 64% White. Thus, the sample was quite representative of US demographics, with only moderate under-representation of Black students.

Measure

Each MOCCA form contains 40 items consisting of a seven-sentence story in which the sixth sentence is missing. From three alternatives, students must select the sentence that best completes the story. Each story has only one item, so MOCCA does not have a testlet structure. Within grade, stories were assigned to forms so that the average story reading level and number of words per story was as nearly equal as possible. Within the reading level and number of words constraint, stories were randomly assigned to forms within grade. For each grade, story reading levels range from one level below grade to one level above grade. For instance, in third grade, forms contain stories with reading levels from second through fourth grade, with a mean of 3.0 on the Flesch-Kincaid scale.

Procedure

In MOCCA, directions are shown to the student on a screen with two sample items. By selecting a button, the student can choose to have the directions read. Students were randomly assigned to forms, and each student took the form they were assigned on a laptop or tablet in a computer lab or classroom. The test is untimed, but teachers often limited the amount of time to approximately one period, about 45 minutes with a range of 30 – 60 minutes. Students were required to answer each item before they could move to the next item. As a result, the only items left blank (if any) were ones at the end of the test that the student did not reach.

Comprehension Dimension

Given the unusual nature of the response options, we began our efforts to model comprehension by plotting empirical test option response functions (Figure 2). To create these graphs, items were scored dichotomously, a two-parameter logistic model was fit to the dichotomously scored items, and then for each of 15 intervals along the 2PL θ continuum, we plotted the mean number of items endorsed in each response category. Since each option type

appears in all 40 items, these means could range from 0 to 40. While the 2PL model was used to estimate θ, the main results in the graphs are not sensitive to the choice of dichotomous model, because the correlations of the θ scores for various dichotomous models (not shown) are so high.

The response variables in Equation 1 fall between an ordered polytomous variable and a nominal response variable in that the response options are partially ordered. Conceptually, the correct answer is above the two incorrect answers, but the incorrect answers are not ordered. However, Figure 2 shows that the lateral connection response curve has a unimodal, nonmonotonic empirical test option response function, and so it behaves somewhat like a middle category in an ordered polytomous variable. Thus, for the purposes of estimating an IRT-based comprehension score, we coded the response of person *i* on item *j* as:

x_{ij} = 0 if paraphrase response (1)

= 1 if lateral connection response

= 2 if correct (causally coherent) response

Given the conceptual partial ordering and the test option response functions shown in Figure 2, we decided to fit both ordered and nominal models for polytomous data. The following decision rule was adopted regarding fit: on the basis of the AIC and BIC, select the model that performs best across all forms and grades from among those with acceptable *RMSEA*s (*RMSEA* ≤ .09; Browne & Cudeck, 1992; Hu & Bentler, 1999; Maydeu-Olivares & Joe, 2014). Also, we preferred to use the same model for all forms.

Incorrect Response Propensity (IRP) Dimension

A second goal was to develop an IRT-based score that could be used to identify low-scoring students with a strong propensity toward either the lateral connection or the paraphrase response. Initially, we examined a simple raw score indicator, the number of paraphrase

responses (*NP*) chosen minus the number of lateral connection responses (*NL*) chosen, *NP – NL*. The *NP – NL* score can be considered a measure of a bipolar dimension in which students who choose only paraphrase responses are at the upper extreme, and students who choose only lateral connection responses are at the lower extreme. The intuitive appeal of this simple raw score indicator led us to consider a rather unusual coding of the item data such that, if a student answered every item, the total item score would be within an additive constant of *NP – NL* (i.e., *NP – NL* + 40) and also a sufficient statistic for estimating the latent variable θ, assuming the data satisfied the partial credit model. That is, if x_{if} is the score of person *i* on item *j*, then in this second coding,

$x_{ij} = 2$ if the paraphrase response is chosen (2)

= 1 if the correct, causally coherent response is chosen

= 0 if the lateral connection response is chosen.

With 40 items, if the student answers every item, the simple total score will range from 0 to 80.

We are all familiar with simple rating scales with a neutral category for the measurement of attitudes, such as the following:

□ Disagree □ Neither Agree Nor Disagree □ Agree

In the coding above, the multiple-choice responses are conceived as a quasi-rating scale with the correct response as a neutral point. Rather than being "Neither Agree nor Disagree," however, the neutral point is "Neither Paraphrase nor Lateral Connection." This pseudo-rating scale is bipolar, with "Lateral Connection" at one end and "Paraphrase" at the other.

□ Lateral Connection □ Neither Paraphrase Nor Lateral Connection □ Paraphrase

The partial credit model is a model for ordered polytomous responses categories. If the partial credit model holds, the total of the item scores is a sufficient statistic for estimating the underlying θ parameters (DeAyala, 2009, p. 169; Masters, 1982; Wright & Masters, 1982). With the coding in Equation 2 and given that a student answers every item, then with a little algebra (see Appendix), the total score can be shown to be within an additive constant of the difference (*NP – NL)*. To explain this conceptually, one can conceive of computing the total score by initially giving each person a score of 40 points prior to beginning the test and then subtracting one point for every lateral connection response and adding one point for every paraphrase response. Then the person's total score would equal 40 + (*NP – NL*). As this expression shows, the person's total score is within an additive constant of the difference (*NP – NL*), and the total score is a sufficient statistic for estimating θ in the partial credit model. Therefore, the difference (*NP – NL*) would also be a sufficient statistic for estimating θ in the partial credit model. (*NP – NL*) is a difference or contrast between *NP* and *NL*, and variables that reflect such contrasts have been called style variables or within-person contrast dimensions (Messick, 1994). In our context, "style" means the student's predominant style of reasoning and/or response when providing an incorrect response.

The sufficiency of 40 + (*NP – NL*) led us to code the data as in Equation 2, and to fit the partial credit model and two competing models that do not assume equal item discriminations, the generalized partial credit model and the graded response model. In these models, the θ dimension is conceived as a bipolar dimension, with students who predominantly choose the lateral connection response at the negative end, and students who predominantly choose the paraphrase response at the positive end. In the middle are students who choose the two types of

incorrect responses (approximately) equally often and includes students who get most items correct.

Results

In our norming data, 84% of 3rd graders completed all of the items. For 4th and 5th grades, the corresponding figures are 74% and 90%. The average number of items completed was 37, 34, and 38 for 3rd, 4th, and 5th graders respectively. The average number of minutes spent on the test was 39, 34, and 39 for 3rd, 4th, and 5th graders.

Comprehension Dimension

Table 1 shows the fit measures for the ordered and nominal polytomous models: graded response model (Samejima, 1969), generalized partial credit (Muraki, 1992), and nominal models (Bock, 1972). Comparing the AIC and BIC, the nominal and graded response models tended to have the lowest values, but the choice between these two models was not entirely clear. Of the three models, the AIC for the nominal model was lowest for all nine forms. For the BIC, the graded response model had the lowest value for seven forms, and the nominal model had the lowest BIC for the remaining two forms. Further, differences in the AIC and BIC for the two models was not always large. The *RMSEA* is meaningless for the nominal model (Maydeu-Olivares & Joe, 2014), so is not reported. The *RMSEA* for the graded response model ranged from .00 - .50, and was at least acceptable (*RMSEA* < .09) for eight of the nine forms.

Close examination of the discrimination parameters (not shown) for the nominal model, helps explain why that model tended to have better fit measures, at least better AIC, but also why an ordered polytomous model, the graded response model was a close second. If the nominal model is fit to three ordered categories, one would expect the ordering of the nominal model discrimination parameters to correspond with the ordering of the categories. On all nine forms,

the discrimination parameters for the correct alternative was the highest. For most, but not all, of the items, the category discrimination orderings were paraphrase < lateral connection < causally coherent, the ordering in Figure 1, and so the graded response model based on this ordering fit reasonably well for most forms, as reflected in the BICs. However, at the item level, there were exceptions to this ordering of discrimination parameters, which resulted in the nominal model fitting somewhat better, at least as measured by the AIC.

In developing an overall measure of comprehension, does the choice of model matter practically? Table 2 shows the correlations of the θ estimates from the nominal model and the graded response model. To two decimal places, the θ correlations are 1.00 for the graded and nominal models across all nine forms. To allow for a comparison of the θ estimates from a more familiar dichotomous model (i.e., correct/incorrect), Table 2 shows the correlations of score estimates from polytomous models with those from the 3PL model with guessing parameters constrained equal across items (3PLC), the best fitting of several dichotomous models. These ranged from .98 to 1.00. Marginal reliability estimates for the polytomous models ranged from ranged from .86 to .92.

It should be noted that the data to which we have applied the nominal model is somewhat different from the multiple-choice data in most other applications (e.g., Sadler, 1998). In our data, the incorrect option categories represent meaningful categories. That is, category 1 was the same type of option, paraphrase, for every item; category 2 was a lateral connection option for every item. In most other applications, the content of the option category varies unsystematically from item to item. Nevertheless, given the high correlation between the 3PLC and polytomous model θs and given that dichotomous models are more commonly used with achievement data,

some of our colleagues have questioned whether switching to a nominal model for the causally coherent dimension is warranted.

Incorrect Response Propensity (IRP) Dimension

Table 3 shows the fit measures for the IRP models. For all nine forms, the AIC was lowest for the graded response model, and the BIC was lowest for the partial credit model. Given skepticism regarding the equal discrimination parameter of the partial credit model, our tentative plan is to use the graded response model for measuring the IRP dimension, although that may change once we have compared the fit of the graded and partial credit models using the complete norming and equating sample data now being collected. The marginal reliability estimates for θ of the graded response model ranged from .59 to .70.

Table 2 shows the correlation of the IRP dimension scores (using the graded response model) with comprehension scores based on the 3PLC, graded response, and nominal models. These correlations display similar trends across the grades and forms. Using the IRP and nominal model comprehension scores to illustrate the trends, the correlations were generally negative and decreasing by grade. Across the three forms within each grade, the correlations ranged from -.494 to -.363 in third grade, from -.403 to -.229 in fourth grade, and -.255 to .127 in fifth grade. These correlations provided support for the discriminant validity of the IRP dimension, in that the absolute values of these correlations suggest that the IRP dimension is distinct from the comprehension dimension. In third and fourth grade, and to a lesser extent in fifth grade, the correlations are negative, suggesting that a propensity toward the paraphrase end of the IRP dimension is associated with lower comprehension scores. As shown in Figure 2, those at the lowest levels of comprehension show a strong predominance of paraphrase over lateral

connection responses, but at higher levels of comprehension, lateral connection responses become slightly more predominant.

Discussion

Results to date provide evidence for the reliability of the raw comprehension score (number correct), the IRT comprehension dimension score, and raw scores for the error propensities (*NP*, *NL*, and *NR*). Familiar IRT models seem appropriate for MOCCA response variables and will serve as the basis for equating across forms and grades. While MOCCA was not designed as a summative assessment, correlations of the total correct score with standardized reading and math measures display a pattern that support both the convergent and discriminant validity of MOCCA with respect to existing reading and math measures. The MOCCA total score may be useful as a summative measure or a screening measure, but its efficiency and error propensity scores were designed for formative application in the design of classroom and individual instruction. Specifically, MOCCA is designed to provide error propensity and efficiency scores useful for instructional planning when applied formatively without sacrificing information about overall comprehension ability similar to that provided by existing assessments. It should be noted, however, that in predicting summative SBAC scores (Biancarosa et al., 2018), error propensity scores were found to add predictive validity over and above that provided by a single comprehension score. Our student samples have been diverse, and items have been screened for differential item functioning by gender and, to a lesser extent, ethnicity (only Hispanics vs. Whites due to sample size limitations). This diversity enhances generalization to diverse populations, but it does not ensure that results generalize equally well to every subpopulation.

As we complete development of MOCCA, we are using IRT to equate the causally coherent dimension across forms to facilitate tracking student growth within and across grades. Then, we plan to fit a separate unidimensional graded response model for the incorrect response variables as a way to identify students who predominantly favor the paraphrase or lateral connection responses and identify items that best discriminate between students who predominantly choose one or the other type of incorrect response. In student reports, we do not plan to report the incorrect response propensity score, but we do plan to identify students with scores at least one standard deviation below the mean as possible *lateral connectors*, and students with scores at least one standard deviation above the mean as possible *paraphrasers*.

Classroom Application

MOCCA has both practical and technical aspects that make the test data pertinent to classroom application. From a technical perspective, as stated in the introduction, the design of the MOCCA forms has three critical features to make the test data useful formatively. The first is an IRT-based comprehension score that can be used to equate forms within and across grades so that student progress can be monitored longitudinally on up to three occasions within a grade and across the three grades without using the same form more than once for any given student. We plan to use an anchor item design and the normative sample data currently being collected for the equating.

The second design feature involves using a cut-score on the comprehension dimension to identify poor comprehenders who show a predominant incorrect response type. Poor comprehenders with a score one or more standard deviations below the mean on the IRP dimension will be flagged as possible *lateral connectors*. Those poor comprehenders with an

IRP dimension score one or more standard deviations above the mean will be classified as possible *paraphrasers*.

Third, the MOCCA design also allows for the development of scores to reflect automaticity or efficiency of response. To date, we have reported only a reading comprehension efficiency measure, minutes per correct response: the total amount of testing time divided by number of correct responses. Early results (Biancarosa, 2018) indicate that, at least at some grades, the efficiency index reflects information with incremental validity in predicting proficiency on a statewide test. The efficiency index is at an earlier stage in development and implementation than are the other two major design features, although none are at the final stage of implementation.

Practically, teachers have identified four useful implications of administering MOCCA in their classrooms. First, MOCCA can diagnose a student's comprehension issue (i.e., paraphrases or lateral connections). Traditional standardized comprehension assessments cannot do this. Instead, those assessments can only indicate whether a student is struggling or not with comprehension; they cannot identify *why* the student struggles. Second, as a result of the diagnosis, teachers have indicated that they can better identify students and form appropriate reading groups. Classroom teachers use a variety of instructional techniques and settings to maximize learning. This includes whole class, small group, and individual instruction. MOCCA subscores and diagnoses enable teachers to appropriately group students with similar issues for efficient instruction.

Third, also as a result of diagnosis, student groupings, and previous research, teachers can appropriately select texts, reading strategies, and interventions for students. As previously indicated, teachers can utilize appropriate questioning techniques while teaching to aid student

cognition (e.g., McMaster et al., 2012). With “paraphrasers,” teachers (or reading partners) can encourage them to make a connection between the just-read text and something that they previously read in the text. With “lateral connectors” or “elaborators,” teachers can prompt them to make a causal connection within the text (i.e., “Why did …?”). Finally, based on recent survey data asking teachers for feedback based on the MOCCA results from their classrooms, some teachers found MOCCA data to be useful for triangulating with other reading measures and monitoring student growth.

No single reading assessment can measure and identify all issues associated with reading. Nor do all struggling readers struggle the same way. Some reading assessments are appropriate for identifying issues that struggling readers have with vocabulary. Other reading assessments are appropriate for identifying issues with fluency or decoding. MOCCA is appropriate for, and was systematically designed to, identify issues of poor and slow comprehension. In combination with other assessments, MOCCA could help identify a reader who gets the gist of a text (i.e., sufficient vocabulary and decoding skills) but is unable to make appropriate inferences while reading.

We conceptualize reading like learning to play the piano or shoot free throws in basketball: doing any of these things well requires both instruction and practice. In the literature on automaticity, structured practice is the most commonly mentioned intervention (LaBerge & Samuels, 1974; Logan, 1997; Samuels & Flor, 1997). Hence, structured practice would seem to be an appropriate intervention for students with good comprehension but poor efficiency. To date, work based on think-aloud measures suggests that poor comprehension can be addressed through questioning interventions and that in classroom settings (but perhaps not in small group tutoring interventions), paraphrasers and lateral connectors respond differentially to such

questioning interventions (McMaster et al., 2012; McMaster et al, 2014; Rapp et al., 2007). It is a matter for future research whether individualizing instruction for poor comprehenders based on comprehension scores and IRP-based classifications (i.e., paraphraser vs. lateral connector) will improve instruction. Like most, if not all tests designed for formative classroom use, the instructional effects of MOCCA are a matter for future research.

Conclusion

In conclusion, we return briefly to the question posed in the title of this manuscript: "Can we learn from student mistakes in a formative reading comprehension test?" In two senses, the answer is a tentative "yes," pending future research. The data reported in Biancarosa et al. (2018) suggest that information about student mistakes can be useful in identifying students at risk of failing to reach proficiency on a statewide exam. Among students with equal numbers of mistakes, those with a predominance of lateral connection errors were more at risk, especially in third and fifth grades; whereas those with a predominance of items not reached were less at risk. Second, the findings by McMaster et al. (2012) and Rapp et al. (2007), tempered by the results of McMaster et al. (2014), suggest that information about mistakes may be useful in individualizing interventions for struggling comprehenders.

However, as in most test development projects, extensive work on validity (construct, criterion-related, and instructional) must wait until after the norming and calibration phases. Such validation can take an extensive period of time. Beyond MOCCA, which has a unique item design, the question remains as to whether and what extent subscores based on meaningful distractors and efficiency can be developed for other reading tests and tests in other content areas.

The unique features of MOCCA cannot be fully utilized unless it is administered prior to or early in the instructional process, so that the information provided by those features can inform the instructional design and student individualization processes.

Appendix

With the point assignment in Equation 2, the total score T will be

$$T = 0*NL + 1*NC + 2*NP \quad \text{(A1)}$$

$$= NC + 2*NP \quad \text{(A2)}$$

Given a student who answers every item, the sum of the three subscores will be 40:

$$NL + NC + NP = 40 \quad \text{(A3)}$$

Utilizing the relationship in A2 and A3

$$T = NC + 2*NP \quad \text{(A4)}$$

$$= NC + 2*NP + [40 - (NL + NC + NP)] \quad \text{(A5)}$$

$$= NP - NL + 40 \quad \text{(A6)}$$

Hence, given the item coding in Equation 2, the total of the item scores will be within an additive constant of the difference $NP - NL$. Since the total score is a sufficient statistic for estimating θ in the partial credit model, the difference $NP - NL$ will also be a sufficient statistic for estimating θ. Hence, the θ estimate can be considered an index of the same dimension as is $NP - NL$. Maris and van der Maas (2012) use a similar line of reasoning to justify their IRT model based on a scoring rule that leads to a sufficient statistic for estimating θ just as we have justified the partial credit model based on the fact that it leads to a total score that is within an additive constant of a scoring rule that provides a sufficient statistic for the model and an intuitively plausible index of the construct, propensity to favor paraphrase or lateral connection responses.

References

Biancarosa, G., Kennedy, P. C., Carlson, S. E., Yoon, H.-J., Seipel, B., Liu, B., & Davison, M. L. (in press). Constructing subscores that add validity: A case study of identifying students at-risk. *Educational and Psychological Measurement*.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika, 37,* 29 – 51.

Browne, M. W. & Cudeck, R. (1992). Alternative ways of assessing model fit. *Sociological Research and Methods, 21(2),* 230 – 258.

Cianco, D., Thompson, K., Schall, M., Skinner, C., & Foorman, B. (2015). Accurate reading comprehension rate as an indicator of broad reading in students in first, second, and third grades. *Journal of School Psychology, 53,* 393-407. doi: 10.1016/j.jsp.2015.07.003

Cote, N., Goldman, S. R., & Saul, E. U. (1998). Students making sense of informational text: Relations between processing and representation. *Discourse Processing, 25(1),* 1 – 53.

Davison, M. L., Biancarosa, G., Carlson, S. E., & Seipel, B. (2018). Preliminary findings on the computer-administered Multiple-choice Online Causal Comprehension Assessment, a diagnostic reading comprehension test. *Assessment for Effective Intervention, 43(3),* 169 – 181.

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(3),* 263 – 279.

deAyala, R. J. (2009). The theory and practice of item response theory. New York: Guilford.

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.

Deno, S. L. (1985). Curriculum-based measurement: The emerging alternative. *Exceptional Children*, *52*, 219–232. doi: 10.1177/001440298505200303

Goldhammer, F., Naumann, J., Stelter, A., Tóth, K., Rölke, H., & Klieme, 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*(3), 608–626. doi: 10.1037/a0034716

Graesser, A. C., McNamara, D. S., & Kulikowich, J. M. (2011). Coh-Metrix providing multilevel analyses of text characteristics. *Educational Researcher, 40*(5), 223-234.

Hale, A. D., Henning, J. B., Hawkins, R. O., Sheeley, W., Shoemaker, L., Reynolds, J. R., & Moch, C. (2011). Reading assessment methods for middle-school students: An investigation of reading comprehension rate and Maze accurate response rate. *Psychology in the Schools*, *48*(1), 28–36. doi: 10.1002/pits.20544

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*(2), 184–192. doi:10.1039/C1RP90023D

Hestenes, D., Wells, M., & Swackhamer, G. (1992). Force Concept Inventory. *The Physics Teacher*, *30*(3), 141–158. doi:10.1119/1.2343497

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural equation modeling: A multidisciplinary journal, 6*(1), 1-55.

January, S. A., & Ardoin, S. P. (2012). The impact of context and word type on students' maze task accuracy. *School Psychology Review, 41*(3), 262-271.

Kincaid, J. P., Fishburne, L. R. P., Jr., Rogers, R. L., & Chissom, B. S. (1975). Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for Navy enlisted personnel. (Research Branch Report No. 8-75). Millington, TN: Naval Air Station Memphis (75). Retrieved from http://library.ucf.edu/Web/purl.asp?dpid=DP0008946

Kintsch, W. & Rawson, K. A., (2005). Comprehension. In M. J. Snowling & C. Hulme (Eds.) *The science of reading: A handbook* (pp. 209-226). Malden, MA: Blackwell.

Kolen, M. J., & Brennan, R. L. (2014). Test equating, scaling and linking: Methods and practices (3rd ed.). Springer: New York.

LaBerge, D., & Samuels, S. J. (1974). Toward a theory of automatic information processing in reading. *Cognitive Psychology, 6*(2), 293–323. doi:10.1016/0010-0285(74)90015-2

Logan, G. D. (1997), Automaticity of reading: Perspectives from the instance theory of automatization. *Reading and Writing Quarterly, 13*(2), 123 – 146.

Lord, F. M. (1980). Application of item response theory to practical testing problems. Routledge: New York.

Maris, G. & van der Maas, H. (2012). Speed-accuracy response models: Scoring rules based on response time and accuracy. *Psychometrika, 77*, 615 – 633.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*, 149 – 174.

Maydeu-Olivares, A. & Joe, H. (2014). Assessing approximate fit in categorical data analysis. *Multivariate Behavioral Research, 49,* 305 – 328.

McCane-Bowling, S. J., Strait, A. D., Guess, P. E., Wiedo, J. R., & Muncie, E. (2014). The utility of maze accurate response rate in assessing reading comprehension in upper elementary and middle school students. *Psychology in the Schools, 51*(8), 789–800.

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*(1), 17-24.

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

Messick, S. (1994). The matter of style: Manifestations of personality in cognition, teaching, and learning. *Educational Psychologist, 29,* 121 – 136.

Perfetti, C. (2010). Decoding, vocabulary, and comprehension. In M. G. McKeown & L. Kucan (Eds.), *Bringing reading research to life* (pp. 291–302). New York: Guilford Press.

Perfetti, C. A., & Lesgold, A. M. (1979). Coding and comprehension in skilled reading and implications for reading instruction. *Theory and practice of early reading, 1,* 57–84.

Posner, M. I., & Snyder, C. (1975). Attention and cognitive control. In R. Solso (Ed.), *Information processing and cognition: The Loyola Symposium* (pp. 55–85). Hillsdale, NJ.: Erlbaum.

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

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. doi: 10.1080/10888430701530417

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*(3), 165–396.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement,* No. 17.

Samuels, S. J., & Flor, R.F. (1997). The importance of automaticity for developing expertise in reading. *Reading & Writing Quarterly, 13(2),* p. 107 — 121.

Scientific Software International (undated). *IRTPRO Guide*. Chicago: Author.

Shin, J., Deno, S. L., & Espin, C. (2000). Technical adequacy of the maze task for curriculum-based measurement of reading growth. *The Journal of Special Educaiton 34*(3), 164-172.

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*(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*(10), 1036-1047.

Smarter Balanced Assessment Consortium: 2014-2015 Technical Report (2016). Retrieved from http://portal.smarterbalanced.org/library/en2014-15, July 7, 2017.

Smith, M. L., Wood, W. B., & Knight, J. (2008). The genetics concept assessment: A new concept inventory for gauging student understanding of genetics. *CBE – Life Sciences Education, 7(4),* 422 – 430.

Su, S., Davison, M. L., Liu, B., Seipel, B., Biancarosa, G., & Carlson, S. E. (2017). *Item response models incorporating item response times: Measuring reading efficiency and automaticity.* Manuscript submitted for publication.

Thorndike, R. M., & Thorndike-Christ, T. (2010). *Measurement and evaluation in psychology and education* (8th ed.). Pearson: New York.

Trabasso, T., & van den Broek, P. (1985). Causal thinking and the representation of narrative events. *Journal of Memory and Language, 24,* 612–630.

Van Norman, E. R., Christ, T. J., & Newell, K. W. (September, 2017). Curriculum-based measrement of reading progress monitoring: The importance of growth magnitude in decision making, *The School Psychology Review, 46(3),* 320 – 328.

Wright, B. D. & Masters, G. N. (1982). *Rating scale analysis*. Chicago: MESA Press.

Table 1

Fit Measures for Comprehension Dimension Polytomous Models by Form

Form	Measures	Nominal	GRM	GPC	PC
	AIC	30254.84	30406.65	30590.85	30797.58
Form 3.1	BIC	30941.19	30921.41	31105.62	31145.04
	RMSEA		0.00	0.00	0.00
	AIC	30893.37	30980.82	31121.42	31266.31
Form 3.2	BIC	31573.37	31490.82	31631.42	31610.56
	RMSEA		0.00	0.00	0.17
	AIC	27507.57	27703.47	27903.56	28168.89
Form 3.3	BIC	28188.80	28213.93	28414.02	28513.45
	RMSEA		0.04	0.08	0.00
	AIC	25751.35	25838.98	25994.56	26095.49
Form 4.1	BIC	26431.65	26349.21	26504.79	26439.89
	RMSEA		0.07	0.00	0.03
	AIC	24797.85	24947.38	25111.55	25199.78
Form 4.2	BIC	25472.19	25453.14	25617.31	25541.17
	RMSEA		0.05	0.00	0.05
	AIC	23801.40	23865.53	24045.86	24243.11
Form 4.3	BIC	24468.87	24366.13	24546.46	24581.01
	RMSEA		0.03	0.00	0.10
	AIC	22521.90	22633.01	22749.84	22940.31
Form 5.1	BIC	23174.33	23122.33	23239.16	23270.6
	RMSEA		0.50	0.10	0.00
	AIC	18494.37	18613.78	18740.07	19020.99
Form 5.2	BIC	19130.99	19091.25	19217.54	19343.28
	RMSEA		0.05	0.08	0.11
	AIC	18694.55	18766.24	18912.48	19183.6
Form 5.3	BIC	19326.66	19240.31	19386.55	19503.6
	RMSEA		0.02	0.00	0.00

Note: GRM = graded response model, GPC = generalized partial credit model, PC = partial credit model, AIC = Akaike information criterion, BIC = Bayesian information criterion, and RMSEA = root mean square of approximation.

Table 2

Theta Correlations of Comprehension Dimension (3PLC, GRM, NRM Models) and IRP Dimension

Form	3PLC NRM	3PL GRM	NRM GRM	3PLC IRP	NRM IRP	GRM IRP
3.1	0.989	0.984	0.996	-0.473	-0.494	-0.533
3.2	0.989	0.986	0.997	-0.418	-0.464	-0.480
3.3	0.990	0.987	0.996	-0.361	-0.363	-0.398
4.1	0.994	0.992	0.998	-0.389	-0.395	-0.408
4.2	0.990	0.988	0.997	-0.389	-0.403	-0.422
4.3	0.995	0.992	0.998	-0.218	-0.229	-0.239
5.1	0.991	0.992	0.997	-0.257	-0.255	-0.266
5.2	0.993	0.993	0.997	-0.241	-0.254	-0.241
5.3	0.998	0.992	0.998	0.053	0.127	0.139

Note: 3PLC = 3 parameter logistic model of comprehension dimension with equality constrained guessing parameters; NRM = nominal response model of comprehension dimension; GRM = graded response model of comprehension dimension; IRP = graded response model of incorrect response propensity dimension.

Table 3

Fit Measures for Incorrect Response Propensity Models by Form

Form	Measures	GRM	GPC	PC
Form 3.1	AIC	34832.54	34918.16	34928.11
	BIC	35347.30	35432.93	35275.58
	RMSEA	0.00	0.00	0.04
Form 3.2	AIC	34783.59	34821.34	34826.35
	BIC	35293.58	35331.34	35170.60
	RMSEA	0.00	0.00	0.00
Form 3.3	AIC	31632.16	31668.06	31675.98
	BIC	32142.62	32178.52	32020.54
	RMSEA	0.10	0.09	0.00
Form 4.1	AIC	29987.25	30018.02	30024.94
	BIC	30497.48	30528.25	30369.35
	RMSEA	0.00	0.00	0.00
Form 4.2	AIC	28849.58	28890.29	28886.68
	BIC	29355.34	29396.04	29228.06
	RMSEA	0.05	0.00	0.03
Form 4.3	AIC	27552.85	27586.93	27622.59
	BIC	28053.45	28087.53	27960.50
	RMSEA	0.00	0.00	0.00
Form 5.1	AIC	25864.50	25890.99	25915.30
	BIC	26353.81	26380.31	26245.59
	RMSEA	0.09	0.00	0.15
Form 5.2	AIC	21500.10	21519.53	21565.05
	BIC	21977.56	21997.00	21887.34
	RMSEA	0.02	0.00	0.00
Form 5.3	AIC	22186.52	22216.06	22249.61
	BIC	22660.59	22690.13	22569.61
	RMSEA	0.05	0.02	0.01

Note: GRM = graded response model, GPC = generalized partial credit model, PC = partial credit model, AIC = Akaike information criterion, BIC = Bayesian information criterion, and RMSEA = root mean square of approximation.

Figure Captions

Figure 1. Screen shot of practice item with, from top to bottom, the lateral connection, paraphrase, and causal coherent (correct) answers respectively.

Figure 2. Average numbers of responses by theta and response type.

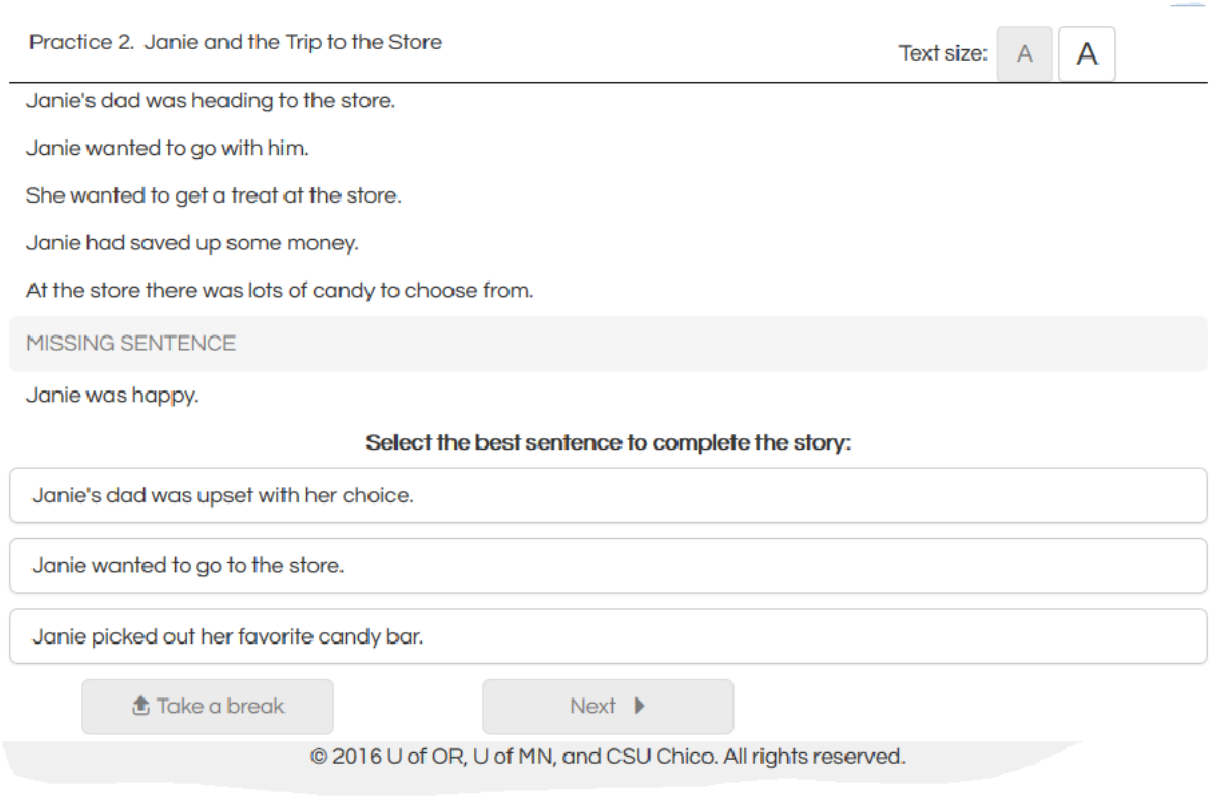

Figure 1. Screen shot of practice item with, from top to bottom, the lateral connection, paraphrase, and causal coherent (correct) answers respectively.

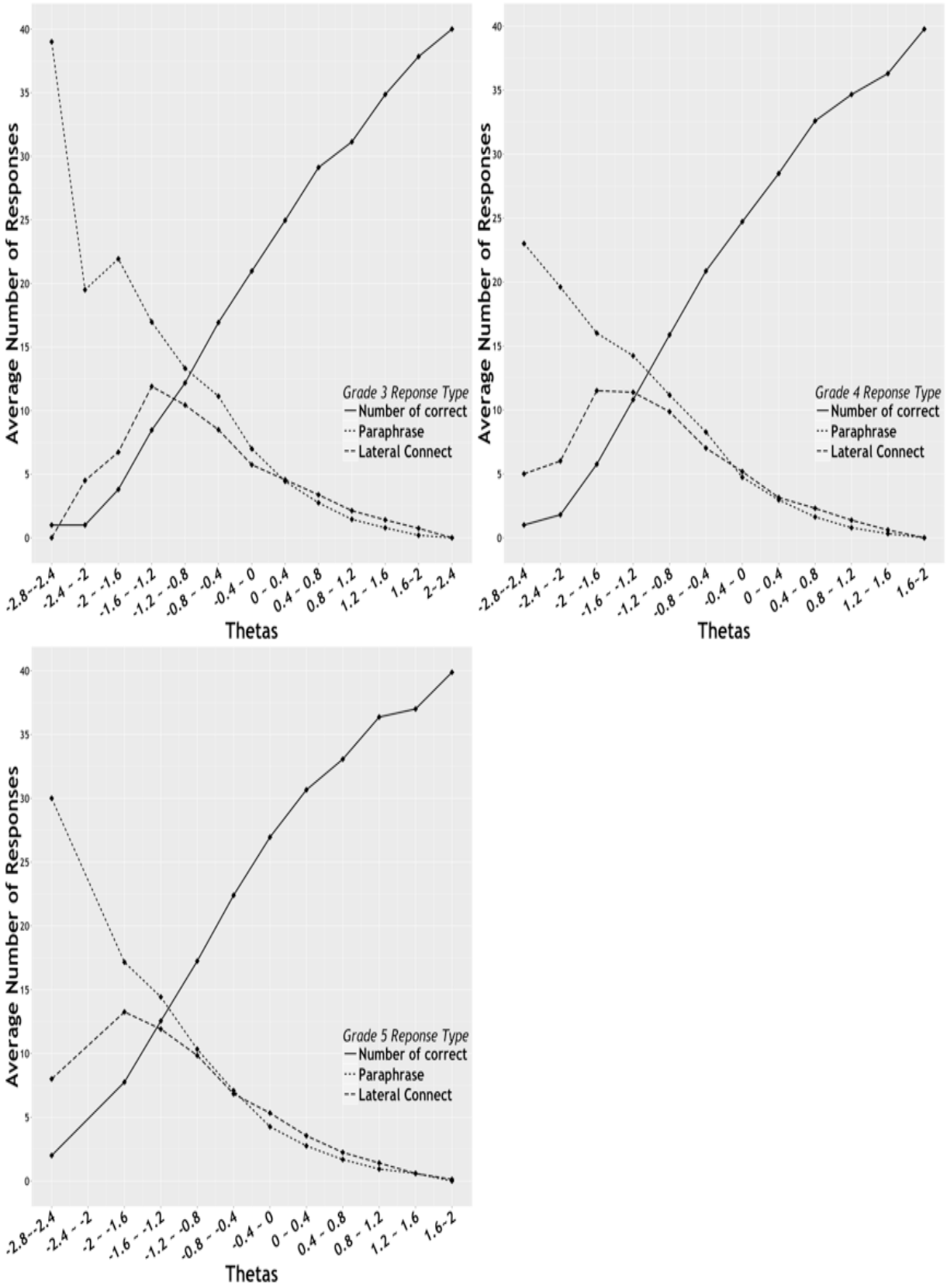

Figure 2. Average numbers of responses by theta and response type.