Technical Report # 08-06

Instrument Development Procedures for Maze Measures

Kimy Liu

Krystal Sundstrom-Hebert

Leanne R. Ketterlin-Geller

Gerald Tindal

University of Oregon

Published by

Behavioral Research and Teaching
University of Oregon • 175 Education
5262 University of Oregon • Eugene, OR 97403-5262
Phone: 541-346-3535 • Fax: 541-346-5689
http://brt.uoregon.edu

This research was supported by Project INFORM (H327B050013-07) from Office of Special Education Programs, U.S. Department of Education. BRT is affiliated with the College of Education, University of Oregon. Opinions expressed herein do not necessarily reflect those of the U.S. Department of Education or Offices within it.

Abstract

The purpose of this study was to document the instrument development of maze measures for grades 3-8. Each maze passage contained twelve omitted words that students filled in by choosing the best-fit word from among the provided options. In this technical report, we describe the process of creating, reviewing, and pilot testing the maze measures. We use three analytic approaches for estimating item difficulty of the test items sampled. The findings of content review and data analysis provide evidence supporting this instrument development process. We conclude that these maze measures are a viable reading comprehension assessment for students in grades 3-8, based on the convergence of evidence.

Instrument Development Procedures for Maze Measures

Comprehension is one of the primary goals of reading, but is likely to be multi-faceted. Certainly, comprehension is a function of background knowledge and vocabulary. A student's ability to use grammar and contextual clues to make sense of words in passages also can be an important influence on reading comprehension. The focus of this study is to develop a measure addressing these latter dimensions of comprehension.

A maze measure is a passage in which several words are omitted from the text and the omissions are substituted with blanks. For each omitted word, the correct answer, along with plausible alternative words, are provided. Students are instructed to select the most appropriate word from the given options to fill in the blank. Some maze measures omit words according to a pre-determined interval (e.g., deleting every seventh word); others select specific types of words to be omitted.

Maze measures are designed to assess students' comprehension by determining their understanding of contextual information, knowledge of syntactical rules and procedures, integration of prior learning, and application of reading skills (Fuchs & Fuchs, 1992). These skills typically are measured within the context of a narrative story, but the length of the story can vary from one to many sentences (Howell & Nolet, 2000). Advanced applications of maze techniques have also been applied to expository text to measure students' content knowledge (Ketterlin-Geller, McCoy, Twyman, & Tindal, 2006; Tindal & Marston, 1990).

Studies on the technical adequacy of maze measures have repeatedly resulted in high reliability coefficients and strong evidence for criterion-related validity. Parker, Hasbrouck, and Tindal (1992) reviewed research on the maze measures and found internal consistency reliability coefficents ranging from r = .84 to .97 for students with disabilities, students in general

education, and English-language learners. Shin, Deno, and Espin (2000) found moderately high alternate form reliability coefficients of r = .80 for testing intervals of 1 to 3 months. Fuchs and Fuchs (1992) conducted a review of research on criterion-related validity evidence and found high correlations (r = .80 to .89) with measures of oral reading fluency and moderate correlations (r = .60) with standardized tests of reading achievement.

The purpose of this technical report is to document the instrument development procedures for developing 18 maze measures for students in grades 3-8. In this report, we describe the test specifications of the maze measures, the process and results of content reviews of the measures, and pilot study results. The internal and external content reviews provide evidence on the measures' appropriateness for students in the targeted grade levels. Finally, we compare the results of three statistical approaches for estimating item difficulty of the maze measures to determine the most appropriate approach.

Methods

In this section, we describe the process used in developing and reviewing the Maze measures, the setting in which our study took place, and the analytic methods used to estimate item difficulty.

Development of the Maze Measures

The maze measures were designed to assess students' comprehension in reading grade-level narrative passages. The process of developing grade-level maze measures included: (a) writing grade-level narrative passages, (b) selecting words to be omitted and creating distracters, (c) developing the technical specifications for online delivery of the maze measures, and (d) creating answer keys and scoring algorithms.

Writing grade-level narrative passages. The item writers created three narrative passages for each grade using basic story grammar. Each grade-level passage contained a main character, setting, and story events and was of comparable length and readability. The passages written for grades 3-5 contained approximately 150 words each; the passages written for grades 6-8 contained approximately 250 words each. The Flesch-Kincaid Grade Level Formula was used to estimate the readability, with each passage targeting the mid-range of the grade level. The guidelines for composing the maze passages explicitly stated that the content, concepts, and vocabulary of the passages must be appropriate for the intended grade-level. The guidelines reminded item writers to be cognizant about the potential biases against certain subgroups of students and to avoid content that might introduce such biases. To reduce potential biases, the maze passages were required to meet two additional criteria: (a) no pre-requisite content knowledge should be necessary to comprehend the text, and (b) students' ethnic background and socio-economic status should not be barriers for understanding the text.

Selecting the omitted words and creating distractors. Each maze passage contained 12 omitted words. The item writers selected key words (words closely tied to understanding of the passage rather than conjunctions, prepositions, or articles) to be omitted and created distractors for each omitted word. For each blank, four options were provided: one correct answer and three syntactically and semantically appropriate distracters. Internal and external reviewers, whose qualifications are described in Table 2, evaluated the appropriateness of the distractors.

Technical specifications. The maze measures described in this technical report were designed for online computer-based delivery. Technical specifications for the measures included information about how students would: (a) gain access to the measures, (b) select their responses, and (c) indicate they had finished the test. Students accessed the tests by logging on to the

designated website. As students read the passage, they clicked on the blank, and a drop-down menu appeared that contained the answer choices. For each item, students were instructed to select the most appropriate word from the provided choices to fill in the blank; they were allowed to scroll back and forth within the same passage. To avoid bias introduced by order-effect, students were divided into seven groups of roughly equal size and were randomly assigned to take one of the four different combinations of measures. The order of measures was counter-balanced. Three of the seven student groups were assigned to take three maze measures; the other students did not take the maze measures and their "responses" were treated as missing data in the student response data files (see Tables 3-8). No data were collected on the length of time taken to finish the three maze passages.

To reduce the likelihood that students looking at each others' responses would introduce error in the test results, a computer-generated algorithm scrambled the order in which options appeared on the computer screen, so that a correct answer could appear as the first, second, third or fourth choice. Thus, when two students took the same measure, looking at the same omitted words, the four answer options appeared in a different order on their screens. When students finished selecting the best fitting words for the 12 omitted key words in the passage, they clicked on the *stop* button to indicate that they were ready to submit their final answers. If they had skipped any omitted words, they were prompted with a statement indicating the number of items completed (with 12 needed for submission) and were required to complete all 12 items prior to submitting their test for scoring.

Creating answer keys and scoring algorithms. Student responses were scored dichotomously, with no additional penalty for incorrect answers. The answer key was coded into the computer programming, allowing the computer to score each item as it was completed.

Internal and External Content Reviews of the Maze Measures

One internal reviewer and six external reviewers evaluated the appropriateness of the content of measures before they were distributed for the pilot test. The reviews focused on the (a) length and readability of the maze passage, (b) appropriateness of concepts, content, and vocabulary, (c) appropriateness of the distracters, and (d) fairness of the test. Qualification of the reviewers, review procedure and findings of the reviews are reported in following sections.

Qualifications of the internal reviewer. The internal reviewer was a third year doctoral student in the area of Special Education. She had two Master's degrees in education: one in general education and the other in special education. She was a certified general education teacher with two years of teaching experience. At the time of the review, she had finished the special education licensure program and her teaching certificate was pending approval. The internal reviewer also had experience in developing math curricula for a major publishing company and reading curricula for English language learners. Her in-depth knowledge of reading and instructional design, as well as her work experience with diverse student populations allowed her to provide constructive feedback on the instrument development, particularly on the issues related to content appropriateness, clarity of direction, and bias against students with limited English proficiency and students with disabilities.

Internal review procedure. For each measure, the internal reviewer evaluated the passages for (a) readability and length of the sentences, (b) grade-level appropriate vocabulary and concepts, (c) flow of sentences, (d) appropriateness of distractors, and (e) possible bias in the content. First, the internal reviewer reported the range of grade-level readability using the Flesch-Kincaid readability formula as well as the sentence length. Second, the internal reviewer inspected whether the wording and sentence topics were appropriate for the indicated grade

level. Third, she reviewed the passages to determine whether they followed the story grammar in a coherent manner. Fourth, the internal reviewer examined the distractors to determine whether they contained an obviously wrong answer but were otherwise syntactically and semantically appropriate. The internal reviewer identified and revised the ambiguous distractors that could be misconstrued as possible correct answers. Fifth, the internal reviewer commented on possible gender, cultural, or linguistic biases of the measures. She also made suggestions for revisions.

Qualifications of external reviewers. Six teachers working in local schools reviewed the passages for the grade level in which they were currently teaching. Five of them held Master's degrees; the other teacher was pursuing a Master's degree in educational leadership and administrative licensure program at the time of the review. Their teaching experience ranged from .5 year to 28 years (see Table 2).

External review procedures. The external reviewers examined the maze measures in four criterion areas: (a) language and vocabulary of the passages for grade-level appropriateness, (b) grade-level appropriateness of concepts described in the passages, (c) clarity of writing, and (d) potential bias in the language of the text. They rated the maze passages on a Likert scale of 1-4 for each criterion: A rating of 1 indicated that the maze was *not at all appropriate* in that criterion area, a rating of 2 indicated that the maze was *somewhat appropriate*, a rating of 3 indicated that the maze was *appropriate*, and a rating of 4 indicated that the maze was *extremely appropriate* in that criterion area. Finally, the external reviewers provided suggestions and comments for any maze receiving a rating of 1 or 2 in any of the criterion areas.

Findings of internal and external reviews. Using the Flesch-Kincaid readability formula to calculate the readability index for each passage, the internal reviewer found all grade 3-5 measures to be within the designated grade-level and most grade 6-8 measures to be slightly

easier than the designated grade-level. Of the eighteen passages, twelve passages were deemed grade-level appropriate for the concept, content and vocabulary. Six passages (Grade 4 Maze 2, Grade 5 Mazes 1 and 2, Grade 6 Mazes 1 and 3, and Grade 7 Maze 3) required further review and revisions. Details about the concerns and suggested revision regarding these six maze passages are reported in Appendix A.

External reviewers rated most maze passages as *appropriate* or *extremely appropriate* on concept, content and vocabulary for the intended grade-levels. Most of the recommended revisions addressed bias against certain groups of individuals and precluded two reasonable answers for any omitted word. All revisions of maze passages and distractors were made without altering the Flesch-Kincaid readability index outside the middle of the year range desired by the instrument developers.

Setting and Participants

The maze measures were administered by two trained research assistants to students in grades 3-8 attending public schools in two mid-sized towns in the Pacific Northwest. In all, 91 grade 3 students, 72 grade 4 students, 109 grade 5 students, 69 grade 6 students, 76 grade 7 students, and 80 grade 8 students took the maze measures (see Table 1).

Data Analyses

Item difficulty of each omitted word was estimated using (a) classical test theory analysis, (b) one-parameter logistic (1PL) Rasch Model, and (c) two-parameter logistic (2PL) item response model.

Classical test theory analyses. We calculated the percentage of valid responses that were correct, or *p*-values, as the estimated item difficulty under the classical test theory model. For example, for Item #2 of Grade 3 Maze 1, 72 of 90 students who responded answered it correctly.

Thus, the *p*-value of this test item was .8. (72 ÷ 90 = .8). The *p*-values of all maze measures for grades 3-8 are reported in Tables 3-8.

One parameter logistic (1PL) Rasch model. We also analyzed the student response data using the WINSTEPS software (Linacre, 2006) and obtained estimates of item difficulties using a 1PL Rasch Model. For each item, we have reported (a) the item number, (b) the number of students who responded (noted in the Tables 9-14 as *COUNT*), (c) number of students who answered correctly (noted as *SCORE*), and (d) the estimated item difficulty (noted as *MEASURE*). The range of *measures* was adjusted to be between 0 and 100 with a mean item difficulty of 50: The higher the value, the more difficult the item.

The 1PL Rasch Model analysis included calculations of fit statistics (Mean Square Outfit). The *Outfit* is an outlier-sensitive fit statistic that reflects unexpected observations by persons on items deemed relatively easy or difficult for them. The items were considered appropriately fitted if their Mean Square Outfit value was within the range of .5 to 1.5. Beyond this range, items were considered inappropriately fitting the model. Specifically, items were considered over-fitting if the Mean Square Outfit value was below .5; under-fitting if the Mean Square Outfit value was between 1.5 and 2.0; and poor-fitting if the Mean Square Outfit value exceeded 2.0 (Linacre, 2006).

Two-parameter (2PL) item response model. We obtained estimates of item difficulty and item discrimination for each item with the 2PL model using the software BILOG-MG (Zimowski, Muraki, Mislevy, & Bock, 2007). The findings generated by BILOG-MG included a parameter file and a score file for every maze passage. The parameter files identified the estimated item difficulties for all test items, including the intercepts, slopes, and item difficulty for each item as well as the standard errors of these indices (see Tables 17-31). The likelihood a

student can answer a specific question correctly can be calculated with the intercept, slope, and item difficulty of an item *and* a student's ability level. The scale score files of the 2PL model identified the estimated range of students' ability levels, given the number of items students attempted and answered correctly. Students with the same number of correct responses could have different scale scores based on different response patterns. These score files identified the minimum and maximum of the scale scores as well as the minimum and maximum of the standard errors of scale scores; they also identified the distribution of the scores across different score brackets (see Tables 32-46).

Results

In this section, we first present the results from the internal and external review of the maze measures, then present the results of our three analyses: classical test theory, 1PL and 2PL item response models.

Results of Review of Grade 3-8 Maze Measures

The internal and external reviewers agreed that most of the maze passages and distractors were appropriate for the intended grade-level students in their concepts, content, and vocabulary. The results of the internal reviewer's evaluation are presented in Appendix A. She deemed 12 of the 18 Maze passages grade-level appropriate. The external reviewers rated almost all passages either *appropriate* or *extremely appropriate* for the intended grade-level. Most of the recommendations made by the reviewers focused on revising the distractors. The following three scenarios illustrated the nature of the revisions.

First, the distractor was revised if it was too similar to the correct answer or another distractor. For example, some English language learners (ELLs) might not know the word *fur* as animals' hair, but might know the word *hair*. Although the word *fur* was the best answer, the

distractor *hair* could be a reasonable answer. Based on the reviewers' feedback, the item writer changed the distractor *hair* to *hand*.

Second, a revision was made if the vocabulary words used in the passage were too difficult for students in the target grades. For example, the words *stark* and *ominous* were considered too difficult for the sixth grade students. Those words were replaced with *barren* and *dark*, respectively.

Third, a passage was revised if the content was not cohesive or did not follow a logical story structure. Additionally, the passage content was changed if the reviewers thought it was biased against a certain group of students. For example, Grade 6, Maze 3 was a story about students going trick-or-treating in a presumably haunted house during Halloween. If the students did not have background knowledge about the customs related to Halloween and trick-or-treating, they might have difficulty choosing the best-fit word to describe how the character reacted to stepping on a squeaking floorboard. Within the four possible answers, students might have difficulty choosing between two very similar answers: carefully and cautiously. Based on the reviewers' comments, the item writer changed *carefully* to *carelessly*, so that the difference between the correct answer and distractor was not too subtle.

In all, the concepts, content, and vocabulary used in most maze passages were considered appropriate for the targeted grade-level students. The reviewers did not detect any significant bias against specific student populations. The minor revisions of maze passages and distractors were made without resulting in the passage exceed the targeted range of the Flesch-Kincaid readability index.

Classical Test Theory Analyses

Tables 3-8 show the estimates of item difficulties for the maze measures in grades 3-8, using the classical test theory model. The results indicate that the range of item difficulty varied across grades. Grade 4 items had the narrowest range of *p-value* of all grades (.67 - .97); Grade 5 items had the widest range of *p*-value (.31 - .97).

One Parameter Logistic (1PL) Rasch Model

Tables 9-14 display the estimates of item difficulties for grade 3-8 measures, using the 1-parameter logistic (1PL) Rasch model. The results indicate that the range of item difficulties varied widely across grades. Grade 3 measures had the narrowest range of item difficulties (36.60 -70.53), with Grade 5 measures having the second narrowest range of item difficulties (33.42 – 69.76). Grade 8 measures had the widest range of item difficulties (33.19 – 87.69), with Grade 7 measures the next widest range of item difficulties (31.75 – 80.18).

In Table 15 we report the number of non-productive items within each passage. In Table 16 we report the number of items considered productive, over-fitting, under-fitting and poor fitting respectively, by grade. Grade 6 measures had the most adequately fitting items (31 out of 36 items were deemed productive) and Grade 8 measures had the fewest (only 18 items). Grades 4, 6, and 7 measures had no poorly fitting items; Grades 3 and 5 measures had only one. Grade 5 Maze 3 and Grade 6 Mazes 1 and 3 each had one poorly fitting item. Grade 8 Maze 1 had 7 poorly fitting items, and each of Grade 8 Mazes 2 and 3 had 5 poorly fitting items (see Tables 15 and 16).

Two-parameter (2PL) Item Response Model

Tables 17-31 display estimates of item difficulties using a 2PL IRT models. The results indicate that the item characteristic curve of all items varied with the difference in intercepts,

slopes, and item difficulties. Items in Grade 3 measures had more similar slopes and intercepts than those in Grades 4-6 and 8. Item characteristics curves of items in Grades 4-6 and 8 differed visibly. Grade 7 data did not converge in the 2PL analyses. The findings reported in Tables 32-46 indicate that students who answered more questions correctly received higher scale scores. The error terms were larger when there was only one person scoring within that scoring bracket or the score was extremely high or low. The ranking of item difficulties differed between the 1PL and 2PL models (see Table 47).

Discussion

In this technical report, we described instrument development procedures used to create maze measures for students in grades 3-8. We reported the outcomes from three different approaches to estimating item difficulty. In our discussion of these three different statistical approaches of estimating item difficulty, we first examine whether our data met two IRT assumptions, and second how well our data fit the proposed models.

Examining the Assumptions of IRT Models

In this study, we compared three different analytical approaches for estimating item difficulty: a classical test theory model, the 1PL Rasch model, and a 2PL IRT model. First, we examine the two assumptions of IRT Models. The first assumption is that a test is unidimensional. The 1PL Rasch model analysis indicated that more than 85% of the items fit the model (Tables 9-14). Among the test items sampled, 147 of the 216 appropriately fit the IRT model, providing useful information about the construct we intended to assess. There were 50 test items classified as over-fitting and 12 as slightly under-fitting. Only 6 out of 216 items were considered poor-fitting items that might introduce construct-irrelevant variance in test scores. Given the small percentage, we assumed that the negative impact was very limited. We

concluded, therefore, that the data fit the model fairly well. If the data fit the constrained 1PL Rasch model, they would also likely fit the 2PL model. Therefore, the first assumption was met.

The second assumption in an item response model is local independence: item and person parameters fully account for interrelationships between items and persons, with no other factors influencing this interrelationship. Violation of local independence occur when (a) the respondents' speed is a factor of their performance, (b) different respondents have differential exposure to the test items (e.g. unfamiliar vocabulary for English learners) or (c) the test items are dependent (e.g. answering one item influences the answers of other items) (Yen, 1993).

The maze format, in theory, might violate the assumption of local independence because the likelihood that students will correctly respond to an item can be influenced by whether the students correctly filled in preceding words. However, in inspecting the students' response patterns, we did not find evidence of violation of the second assumptions as suggested in Yen's examples. The maze measures were not timed, so speed was not a determining factor of how many words students filled in correctly. We did not gather information about the participants' disability, SES status, or if they were English learners; therefore, we did not know whether these extraneous factors were influential in the interaction of persons and items. However, the results of the internal and external review suggested that the measures were not biased against English language learners, students with a disability, or students with low SES status. Finally, although the test items in the maze measures appeared to be dependent, we did not find direct evidence that supported the conjecture that students' incorrect responses on one item led to their making consecutive errors. In the absence of direct evidence indicating violation of local independence, we concluded that the second assumption of IRT was tentatively met. With these two IRT

assumptions met, person invariance and item invariance of the 1PL Rasch model and 2PL models can be assumed (Embretson & Reise, 2000).

Interpreting the Findings in 1PL Rasch Models

The estimated item difficulty for each test item is expressed as its *measure* in the 1PL Rasch model. By comparing the distribution of item difficulty of different passages within the same grade, we can determine which passages are more or less challenging. The scaling of items in one grade is independent of the scaling of items in other grades; therefore, no inferences can be made across-grades.

Fit statistics in a 1PL Rasch model can be a powerful indicator of how well the test items function in providing unique information about the examinees' ability on the intended construct. Linacre (2006) defines the productive items as the items with Mean Square Outfits within the range of .5 and 1.5. If we consider the passages that have more non-productive items than productive items as problematic passages, three of the 18 passages (Grade 7 Mazes 1 and 2, and Grade 8 Maze 1) would be labeled as problematic passages (Table 16). However, upon close inspection of those non-productive items, most of them were either over-fitting or slightly under-fitting. These over-fitting items might mislead test users to over-estimate the quality of the measures, but do not necessarily degrade the measures. The slightly under-fitting items might assess construct-irrelevant variance along with the intended construct (Linacre, 2006). The over-fitting items might be related to the fact that these passages are too easy for Grade 7 and 8 students, as evidenced by a majority of students filling in most words correctly. We recommend revising these passages by making the passages and words more challenging for students in Grade 7 and 8.

Compare Item Difficulties obtained in CTT, 1PL and 2PL analyses

We noted the similarity of the estimates of item difficulty among these three analytic approaches. Comparing the ranking of the estimated item difficulties of measures in grades 3-6 and 8, the results between 1PL and CTT are more similar than the results of the 1PL and 2PL model analyses (Table 47). Comparison of 1PL and 2PL models for Grade 7 mazes is not possible because the Grade 7 data did not converge in the 2PL analyses.

Although the results of using 1PL and CTT are similar, we cannot overlook the fact that the *p-values* under the CTT model are population dependent. As such, for a very skilled group, the *p*-values of the measures can be significant higher, which would suggest the items are easy. Conversely, if the group is less skillful, these same test items would yield lower *p-values*, which would suggest that the items are difficult. The *p-values* of the measures would not remain invariant between these two samples. Comparing *p-values* between students from two samples may not be meaningful if one or neither of the samples is representative of the population to which one is intending to generalize (Embretson & Reise, 2000).

By contrast, "item invariance" and "person invariance" under the chosen IRT model can be assumed when the two required assumptions of IRT models are met and items are calibrated appropriately following the IRT calibration procedures (Embretson & Reise, 2000). "Item invariance" means that when items are calibrated appropriately, the person's trait level can remain stable, regardless of which items are taken. "Person invariance" means the item difficulty remains unchanged independent of the respondents' ability levels. Because of "person invariance," the comparison across persons with different skill levels in a non-representative sample group is meaningful when they are anchored by common items. Because of "item invariance," the comparison of the results of different tests is meaningful when the tests are

anchored by common persons. For this reason, it is more advantageous to use IRT models to estimate item difficulties than to use the CTT model.

In 1PL and 2PL IRT models, the values of intercept, slope, and item difficulty are used to describe the unique item characteristics curve for each item. The difference between 1PL and 2PL models is that 1PL Rasch Model constrains the slope of all items to be unified, while the 2PL model allows the slopes of the items to vary (i.e. not all items are equally related to the latent abilities). In 2PL models, the additional parameter, item discrimination, is expressed by the steepness of the slope. The items that discriminate well are items with steep slopes. When all items have very similar slopes, it makes sense to use the 1PL model based on the principle of parsimony. However, in this study, the items had a wide range of intercepts and slopes; therefore, the 2PL model makes more sense because it allows the data to fit the model, not the other way around.

Comparing the outcomes of 1PL and 2PL analyses for each grade, we found differences in overall ranking of item difficulty among the items sampled. However, in many incidences, the relative positions of two test items were the same in these two approaches. For example, the 1PL and 2PL analyses yielded the same conclusion that items # 5, 10, 23, 25, and 35 were the most challenging items in Grade 3, but the actual rankings of these five items varied (see Table 47). We also noted that within the same passage, items with identical item difficulty in the 1PL model (e.g., Grade 3 Maze 1 Items 3, 7, 11 and 12) had different estimates of item difficulty using the 2PL model (see Table 47), primarily because these four items had different item discriminations.

The estimated scale scores and standard errors of the scale scores under the 2PL model varied depending on students' response patterns. In general, students who filled more words in correctly had higher values of scale scores than students who filled in fewer words correctly.

Two students with an equal number of correct answers could have different estimated ability levels. Students who succeeded in highly discriminating items and failed on poorly discriminating items had higher trait level estimates than students who succeeded on poorly discriminating items and failed on highly discriminating items. In IRT models, an item provides a better estimate of the respondents' ability level when the distance on the scale between the person's estimated ability level and item difficulty of the selected item is relatively small. An item bearing such a characteristic was described as an "on target" item. This finding supported test design decisions to include items with a wide range of difficulties, because it increased the likelihood of having items that were on target for the intended student population.

Cautions should be applied in making inferences about item difficulty estimates of the maze measures under the IRT models because the design of mazes, in theory, violates the assumption of local independence. The testlet model may be more appropriate to analyze students' response patterns on the maze measures; this model should be further tested. Another limitation to our study relates to the number of students who received perfect scores in our sample. Larger standard errors of scale scores are expected when respondents have all correct responses. The significant number of perfect scores indicates an insufficient number of difficult items to challenge respondents with high reading comprehension skills. However, this limitation might not be important if the purpose of the measures is to identify students with low reading skills.

Conclusion

Inspecting the values of item difficulty across three different statistical analyses, we found that the estimated difficulty of the test items varied depending on approach used, but the classification of easy versus challenging items was relatively stable. Our preliminary evidence

supported using the 2PL model as a better option than CTT and 1PL model analyses to estimate item difficulty. There was no direct evidence supporting the violation of the assumption of local independence in our data.

In all, most items on these grade 3-8 maze measures functioned appropriately. The evidence suggests that these grade 3-8 maze measures are a viable screening measure to assess students' reading comprehension. In addition, using web-based tests provides other benefits such as efficiency in scoring and reducing human scoring errors.

References

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum.

Howell, K.W., & Nolet, V. (2000). *Curriculum-based evaluation: Teaching and decision making*. Belmont, CA: Wadsworth/Thomson Learning.

Ketterlin-Geller, L.R., McCoy, J.D., Twyman, T., & Tindal, G. (2006). Using a concept maze to assess student understanding of secondary content. *Assessment for Effective Intervention, 31* (2), 39-50.

Linacre, J. B. (2006). Winsteps 3.61.1. Rasch-model computer programs [computer software]. Chicago, IL: MESA.

Parker, R., Hasbrouck, J.E., & Tindal, G. (1992). The maze as a classroom-based reading measure: construction methods, reliability, and validity. *Journal of Special Education, 26*, 195-218.

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 Education, 34*, 164-172.

Tindal, G.A., & Marston, D.B. (1990). *Classroom-based assessment: Evaluating instructional outcomes*. Columbus, OH: Merrill.

Yen, W. N. (1993). Scaling performance assessments: Strategies for managing local item dependences. *Journal of Educational Measurement, 30*, 187-213.

Zimowski, M., Muraki, E., Mislevy, R. J., & Bock, R. D. (2007). Bilog-MG-3 [computer software]. Chicago, IL: Assessment System Cooperation.

Table 1.
Number of Participants in Grades 3-8.

Grades	**Number of Participants**
Grade 3	91
Grade 4	72
Grade 5	109
Grade 6	69
Grade 7	76
Grade 8	80

Table 2.
External Reviewers' Backgrounds and Qualifications.

Teacher	**Current teaching position**	**Education**	**Teaching experience**
Teacher 1	Special Education, K-5	M. Ed	17 years
Teacher 2	5th grade	M. Ed	2 years
Teacher 3	Reading Specialist	M. Ed	4 years
Teacher 4	6th and7th grade	M. Ed	0.5 year
Teacher 5	8th grade	M. Ed	3 years
Teacher 6	7th grade	M. A.	17 years

Tables 3-8:

The descriptive statistics of the Grades 3-8 Maze Measures under Classical Test Theory (CTT) model.

Table 3A.
Grade 3 First Maze Passage.

Item	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
Correct	73	72	79	81	59	76	79	81	76	58	79	79
Incorrect	17	18	11	9	31	14	11	9	14	32	11	11
Valid	90	90	90	90	90	90	90	90	90	90	90	90
Missing*	149	149	149	149	149	149	149	149	149	149	149	149
p*-value*	.81	.80	.88	.90	.66	.84	.88	.90	.84	.64	.88	.88

* Maze measures are three subtests of the entire reading-math battery screening measures. The *Missing* is used to indicate the number of students who *either* were not assigned to take the test *or* did not take the test even they were assigned to take the tests.

** *P*-Value is often referred to the percent of participants who responded the question and answered it correctly.

Table 3B.
Grade 3 Second Maze Passage.

Item	1	2	3	4	5	6	7	8	9	10	11	12
Correct	73	83	82	69	67	80	77	76	71	79	47	67
Incorrect	18	8	9	22	24	11	14	15	20	12	44	24
Valid	91	91	91	91	91	91	91	91	91	91	91	91
Missing	148	148	148	148	148	148	148	148	148	148	148	148
***p*-value**	.80	.91	.90	.76	.74	.88	.85	.84	.78	.87	.52	.74

Table 3C.
Grade 3 Third Maze Passage.

Item	1	2	3	4	5	6	7	8	9	10	11	12
Correct	61	62	82	77	76	73	66	74	68	75	55	71
Incorrect	28	27	7	12	13	16	23	15	21	14	34	18
Valid	89	89	89	89	89	89	89	89	89	89	89	89
Missing	150	150	150	150	150	150	150	150	150	150	150	150
***p*-value**	.69	.70	.92	.87	.85	.83	.74	.83	.76	.84	.62	.80

Table 4A.
Grade 4 First Maze Passage.

Item	1	2	3	4	5	6	7	8	9	10	11	12
Correct	63	63	68	62	66	52	62	68	66	69	65	63
Incorrect	9	9	4	10	6	20	10	4	6	3	7	9
Valid	72	72	72	72	72	72	72	72	72	72	72	72
Missing	113	113	113	113	113	113	113	113	113	113	113	113
P-value	.88	.88	.94	.86	.92	.72	.86	.94	.92	.96	.90	.88

Table 4B.
Grade 4 Second Maze Passage.

Item	1	2	3	4	5	6	7	8	9	10	11	12
Correct	67	59	65	57	63	63	61	50	51	62	56	38
Incorrect	2	10	4	12	6	6	8	18	17	6	12	30
Valid	69	69	69	69	69	69	68	68	68	68	68	68
Missing	116	116	116	116	116	116	117	117	117	117	117	117
p-value	.97	.86	.94	.83	.91	.91	.88	.74	.75	.91	.82	.56

Table 4C.
Grade 4 Third Maze Passage.

Item	1	2	3	4	5	6	7	8	9	10	11	12
Correct	58	63	53	47	50	65	50	64	63	56	60	64
Incorrect	12	7	17	23	20	5	20	6	7	14	10	6
Valid	70	70	70	70	70	70	70	70	70	70	70	70
Missing	115	115	115	115	115	115	115	115	115	115	115	115
P-value	.83	.90	.76	.67	.71	.93	.71	.91	.90	.80	.86	.91

Table 5A.
Grade 5 First Maze Passage.

Item	1	2	3	4	5	6	7	8	9	10	11	12
Correct	104	104	86	77	90	101	99	91	100	97	100	103
Incorrect	3	3	21	30	17	6	8	16	7	9	6	3
Valid	107	107	107	107	107	107	107	107	107	106	106	106
Missing	163	163	163	163	163	163	163	163	163	164	164	164
***p*-value**	.97	.97	.80	.72	.84	.94	.93	.85	.93	.92	.94	.97

Table 5B.
Grade 5 Second Maze Measures.

Item	1	2	3	4	5	6	7	8	9	10	11	12
Correct	106	99	105	95	100	105	101	101	102	104	104	101
Incorrect	2	9	3	13	8	3	7	7	6	4	4	7
Valid	108	108	108	108	108	108	108	108	108	108	108	108
Missing	162	162	162	162	162	162	162	162	162	162	162	162
***p*-value**	.98	.92	.97	.88	.93	.97	.94	.94	.94	.96	.96	.94

Table 5C.
Grade 5 Third Maze Passage.

Item	1	2	3	4	5	6	7	8	9	10	11	12
Correct	100	105	97	102	101	88	100	88	100	103	68	69
Incorrect	9	4	11	6	7	20	8	20	8	5	40	39
Valid	109	109	108	108	108	108	108	108	108	108	108	108
Missing	161	161	162	162	162	162	162	162	162	162	162	162
***p*-value**	.92	.96	.90	.94	.94	.81	.93	.81	.93	.95	.63	.64

Table 6A.
Grade 6 First Maze Passage.

Item	1	2	3	4	5	6	7	8	9	10	11	12
Correct	63	63	29	41	61	59	62	64	63	64	58	41
Incorrect	5	5	39	27	7	9	6	4	5	4	10	27
Valid	68	68	68	68	68	68	68	68	68	68	68	68
Missing	134	134	134	134	134	134	134	134	134	134	134	134
***p*-value**	.93	.93	.43	.60	.90	.87	.91	.94	.93	.94	.85	.60

Table 6B.
Grade 6 Second Maze Passage.

Item	1	2	3	4	5	6	7	8	9	10	11	12
Correct	65	64	51	64	59	30	37	66	64	62	64	54
Incorrect	3	4	17	4	9	38	31	2	4	6	3	13
Valid	68	68	68	68	68	68	68	68	68	68	67	67
Missing	134	134	134	134	134	134	134	134	134	134	135	135
***p*-value**	.96	.94	.75	.94	.87	.44	.54	.97	.94	.91	.96	.81

Table 6C.
Grade 6 Third Maze Passage.

Item	1	2	3	4	5	6	7	8	9	10	11	12
Correct	64	48	41	64	21	60	36	43	64	60	63	52
Incorrect	5	20	27	4	47	8	32	25	4	8	5	16
Valid	69	68	68	68	68	68	68	68	68	68	68	68
Missing	133	134	134	134	134	134	134	134	134	134	134	134
P-value	.93	.71	.60	.94	.31	.88	.53	.63	.94	.88	.93	.76

Table 7A.
Grade 7 First Maze Passage.

Item	1	2	3	4	5	6	7	8	9	10	11	12
Correct	61	65	68	69	65	69	72	72	69	62	71	71
Incorrect	15	11	8	7	10	6	3	3	6	13	4	4
Valid	76	76	76	76	75	75	75	75	75	75	75	75
Missing	128	128	128	128	129	129	129	129	129	129	129	129
***p*-value**	.80	.86	.89	.91	.87	.92	.96	.96	.92	.83	.95	.95

Table 7B.
Grade 7 Second Maze Passage.

Item	1	2	3	4	5	6	7	8	9	10	11	12
Correct	70	68	68	74	74	66	74	72	73	71	74	65
Incorrect	6	8	8	2	2	10	2	4	3	5	2	11
Valid	76	76	76	76	76	76	76	76	76	76	76	76
Missing	128	128	128	128	128	128	128	128	128	128	128	128
P-value	.92	.89	.89	.97	.97	.87	.97	.95	.96	.93	.97	.86

Table 7C.
Grade 7 Third Maze Passage.

Item	1	2	3	4	5	6	7	8	9	10	11	12
Correct	72	64	74	75	73	53	72	60	74	69	59	63
Incorrect	4	12	2	1	3	23	4	16	2	7	17	13
Valid	76	76	76	76	76	76	76	76	76	76	76	76
Missing	128	128	128	128	128	128	128	128	128	128	128	128
***p*-value**	.95	.84	.97	.99	.96	.70	.95	.79	.97	.91	.78	.83

Table 8A.
Grade 8 First Maze Passage.

Item	1	2	3	4	5	6	7	8	9	10	11	12
Correct	72	75	74	74	71	68	71	73	73	75	64	43
Incorrect	7	4	4	4	7	10	7	5	5	3	4	35
Valid	79	79	78	78	78	78	78	78	78	78	78	78
Missing	130	130	131	131	131	131	131	131	131	131	131	131
P-value	.91	.95	.95	.95	.91	.87	.91	.94	.94	.96	.82	.55

Table 8B.
Grade 8 Second Maze Passage.

Item	1	2	3	4	5	6	7	8	9	10	11	12
Correct	66	65	74	72	70	69	70	64	60	72	69	66
Incorrect	10	11	2	4	6	7	6	12	16	3	6	9
Valid	76	76	76	76	76	76	76	76	76	75	75	75
Missing	133	133	133	133	133	133	133	133	133	134	134	134
***p*-value**	.87	.86	.97	.95	.92	.91	.92	.84	.79	.96	.92	.88

Table 8C.
Grade 8 Third Maze Passage.

Item	1	2	3	4	5	6	7	8	9	10	11	12
Correct	73	76	74	74	73	76	73	71	71	71	78	78
Incorrect	7	4	6	6	7	4	7	9	9	9	2	2
Valid	80	80	80	80	80	80	80	80	80	80	80	80
Missing	129	129	129	129	129	129	129	129	129	129	129	129
P-value	.91	.95	.93	.93	.91	.95	.91	.89	.89	.89	.98	.98

Tables 9-14:

The fit statistics of the Grades 3-8 Maze Measures under 1PL Rasch model.

Table 9.
Grade 3 Mazes.

Item	Measure	Count	Score	Outfit	Outfit	Obs.	Exp.
1	50.37	86	69	0.83	-0.39	80.20	85.40
2	51.39	86	68	1.38	1.12	73.30	84.70
3	43.05	86	75	0.88	-0.05	87.20	89.40
4	39.94	86	77	0.83	-0.08	91.90	90.70
5	62.13	86	55	1.01	0.13	66.30	76.20
6	47.00	86	72	2.17	2.19	83.70	87.50
7	43.05	86	75	0.44	-1.16	91.90	89.40
8	39.94	86	77	0.30	-1.37	94.20	90.70
9	47.00	86	72	0.50	-1.27	90.70	87.50
10	62.84	86	54	1.23	1.12	66.30	75.70
11	43.05	86	75	0.57	-0.77	89.50	89.40
12	43.05	86	75	1.60	1.13	89.50	89.40
13	51.36	87	69	1.20	0.65	82.80	84.80
14	38.17	87	79	1.15	0.44	93.10	91.60
15	39.93	87	78	0.20	-1.75	94.30	90.80
16	55.09	87	65	1.01	0.13	80.50	82.10
17	56.78	87	63	1.14	0.60	77.00	80.70
18	43.03	87	76	0.95	0.09	92.00	89.50
19	46.98	87	73	0.67	-0.71	93.10	87.60
20	48.15	87	72	0.73	-0.61	86.20	86.90
21	53.29	87	67	0.88	-0.31	88.50	83.50
22	44.43	87	75	0.52	-1.01	90.80	88.90
23	70.53	87	43	1.54	2.31	64.40	71.50
24	56.78	87	63	1.17	0.71	81.60	80.70
25	60.64	85	57	1.37	1.58	72.90	77.90
26	59.88	85	58	1.33	1.39	74.10	78.50
27	36.60	85	78	0.49	-0.57	91.80	92.30
28	44.91	85	73	0.42	-1.36	92.90	88.90
29	46.24	85	72	0.66	-0.69	84.70	88.20
30	49.83	85	69	0.43	-1.82	90.60	86.20
31	56.62	85	62	1.38	1.37	76.50	81.30
32	48.70	85	70	0.37	-1.97	92.90	86.90
33	54.85	85	64	1.18	0.68	76.50	82.80
34	47.50	85	71	0.43	-1.57	91.80	87.50
35	64.94	85	51	1.44	2.00	76.50	74.90
36	51.96	85	67	0.81	-0.47	84.70	84.80

Table 10.
Grade 4 Maze Measure.

Item	Measure	Count	Score	Outfit	Outfit	Obs.	Exp.
1	49.40	67	58	0.75	-0.40	89.60	88.30
2	49.40	67	58	0.52	-1.01	92.50	88.30
3	38.41	67	63	0.43	-0.53	94.00	94.00
4	50.97	67	57	1.36	0.87	91.00	87.20
5	43.68	67	61	0.36	-1.05	94.00	91.40
6	62.47	67	47	0.94	-0.18	71.60	77.00
7	50.97	67	57	0.55	-1.05	91.00	87.20
8	38.41	67	63	0.24	-0.98	94.00	94.00
9	43.68	67	61	1.77	1.16	91.00	91.40
10	34.88	67	64	0.34	-0.50	95.50	95.50
11	45.80	67	60	0.82	-0.13	91.00	90.30
12	49.40	67	58	0.52	-1.02	92.50	88.30
13	30.60	64	62	1.56	0.80	96.90	96.90
14	51.78	64	54	1.00	0.16	82.80	86.90
15	38.97	64	60	1.71	0.96	92.20	93.70
16	54.66	64	52	1.12	0.42	78.10	84.90
17	44.34	64	58	0.54	-0.61	90.60	91.10
18	44.34	64	58	1.71	1.09	90.60	91.10
19	48.43	64	56	0.31	-1.60	93.80	89.10
20	61.85	63	45	1.73	2.40	68.30	78.30
21	60.80	63	46	1.02	0.17	79.40	79.30
22	44.45	63	57	0.71	-0.24	90.50	91.00
23	54.83	63	51	0.66	-0.89	87.30	84.70
24	72.58	63	33	1.30	1.50	66.70	70.20
25	54.12	66	54	0.88	-0.20	84.80	85.10
26	46.07	66	59	0.43	-1.03	92.40	90.20
27	59.95	66	49	0.91	-0.26	80.30	79.90
28	65.60	66	43	1.50	2.15	66.70	74.50
29	62.90	66	46	1.35	1.41	71.20	76.90
30	41.50	66	61	0.88	0.09	93.90	92.60
31	62.90	66	46	1.21	0.91	65.20	76.90
32	43.94	66	60	0.74	-0.20	90.90	91.30
33	46.07	66	59	0.34	-1.30	92.40	90.20
34	56.63	66	52	1.03	0.21	77.30	83.00
35	51.29	66	56	0.61	-0.87	86.40	87.10
36	43.94	66	60	0.90	0.08	90.90	91.30

Table 11.
Grade 5 Maze Measures.

Item	Measure	Count	Score	Outfit MSQ	Outfit ZSTD	Obs. Match	Exp. Match
1	37.96	88	85	0.43	-0.45	96.60	96.60
2	37.96	88	85	0.31	-0.71	96.60	96.60
3	63.60	88	67	0.82	-0.72	83.00	79.20
4	69.76	88	58	1.18	1.09	75.00	72.80
5	60.32	88	71	1.05	0.28	79.50	82.90
6	46.22	88	82	0.34	-1.26	93.20	93.40
7	49.86	88	80	0.54	-0.91	90.90	91.50
8	59.41	88	72	0.92	-0.19	83.00	83.80
9	48.15	88	81	0.62	-0.60	92.00	92.50
10	51.58	87	78	0.51	-1.12	89.70	90.40
11	46.38	87	81	0.31	-1.35	93.10	93.30
12	38.09	87	84	0.19	-1.02	96.60	96.50
13	33.42	88	86	0.08	-1.03	97.70	97.70
14	51.43	88	79	1.11	0.39	89.80	90.50
15	37.98	88	85	0.28	-0.78	96.60	96.60
16	56.44	88	75	0.99	0.10	86.40	86.70
17	49.89	88	80	1.21	0.55	93.20	91.50
18	37.98	88	85	1.14	0.46	96.60	96.60
19	48.18	88	81	1.35	0.75	92.00	92.50
20	48.18	88	81	1.60	1.10	94.30	92.50
21	46.25	88	82	0.64	-0.47	95.50	93.40
22	41.33	88	84	0.97	0.23	95.50	95.40
23	41.33	88	84	0.97	0.23	95.50	95.40
24	48.18	88	81	1.45	0.90	94.30	92.50
25	50.42	89	80	0.67	-0.60	93.30	91.10
26	39.90	89	85	2.09	1.26	95.50	95.50
27	53.21	89	78	1.02	0.18	85.40	89.30
28	45.01	89	83	1.19	0.49	95.50	93.60
29	47.03	89	82	1.38	0.78	93.30	92.80
30	62.21	89	69	1.30	1.16	80.90	81.10
31	48.81	89	81	0.62	-0.64	92.10	91.90
32	62.21	89	69	0.92	-0.24	76.40	81.10
33	48.81	89	81	0.98	0.15	92.10	91.90
34	42.68	89	84	1.12	0.40	95.50	94.50
35	75.18	89	49	1.39	2.39	62.90	68.90
36	74.61	89	50	1.26	1.68	62.90	69.10

Table 12.
Grade 6 Maze Measures.

Item	Measure	Count	Score	Outfit	Outfit	Obs.	Exp.
1	40.94	68	63	0.73	-0.25	92.60	92.60
2	40.94	68	63	1.06	0.30	92.60	92.60
3	73.75	68	29	0.89	-0.73	75.00	68.20
4	65.11	68	41	1.33	1.94	64.70	69.60
5	44.94	68	61	1.12	0.39	89.70	89.70
6	48.08	68	59	1.56	1.23	85.30	86.90
7	43.08	68	62	0.54	-0.76	91.20	91.20
8	38.39	68	64	0.99	0.24	94.10	94.10
9	40.94	68	63	0.84	-0.04	92.60	92.60
10	38.39	68	64	0.71	-0.18	94.10	94.10
11	49.45	68	58	1.33	0.87	86.80	85.50
12	65.11	68	41	1.25	1.54	70.60	69.60
13	35.53	68	65	0.29	-0.86	95.60	95.60
14	38.74	68	64	0.67	-0.24	94.10	94.10
15	57.43	68	51	1.45	1.58	72.10	77.10
16	38.74	68	64	0.34	-0.93	94.10	94.10
17	48.46	68	59	1.07	0.30	88.20	86.90
18	73.51	68	30	0.98	-0.08	66.20	68.10
19	68.48	68	37	1.23	1.54	63.20	68.00
20	31.15	68	66	0.23	-0.77	97.10	97.10
21	38.74	68	64	0.56	-0.44	94.10	94.10
22	43.44	68	62	0.64	-0.50	91.20	91.20
23	35.67	67	64	0.87	0.14	95.50	95.50
24	53.64	67	54	0.94	-0.06	83.60	81.20
25	40.89	69	64	0.70	-0.29	92.80	92.70
26	59.87	68	48	0.84	-0.67	77.90	74.20
27	65.35	68	41	1.01	0.14	66.20	69.60
28	38.59	68	64	0.42	-0.74	94.10	94.10
29	80.18	68	21	1.31	1.42	76.50	73.70
30	46.79	68	60	1.22	0.58	89.70	88.30
31	68.99	68	36	1.06	0.48	69.10	67.60
32	63.85	68	43	0.69	-1.91	80.90	70.70
33	38.59	68	64	0.63	-0.31	94.10	94.10
34	46.79	68	60	0.75	-0.40	86.80	88.30
35	41.14	68	63	0.37	-1.07	92.60	92.60
36	56.3	68	52	1.05	0.28	76.50	78.00

Table 13.
Grade 7 Maze Measures.

Item	Measure	Count	Score	Outfit MSQ	Outfit ZSTD	Obs. Match	Exp. Match
1	65.69	56	41	0.93	-0.15	78.6	79.1
2	60.36	56	45	1.11	0.4	85.7	84.8
3	55.24	56	48	0.42	-1.33	89.3	88.5
4	53.16	56	49	0.33	-1.47	92.9	89.8
5	59.32	55	45	1.37	0.96	90.9	86
6	51.25	55	49	0.4	-1.05	94.5	91.1
7	41.05	55	52	0.11	-1.1	96.4	95.4
8	41.05	55	52	0.11	-1.1	96.4	95.4
9	51.25	55	49	1.01	0.23	90.9	91.1
10	63.74	55	42	1.19	0.66	74.5	81.9
11	45.2	55	51	0.66	-0.16	94.5	93.8
12	45.2	55	51	0.42	-0.59	94.5	93.8
13	50.8	56	50	1.27	0.6	87.5	91.1
14	55.23	56	48	1.38	0.85	85.7	88.5
15	55.23	56	48	1.08	0.33	89.3	88.5
16	35.14	56	54	1.53	0.79	94.6	96.8
17	35.14	56	54	0.86	0.36	94.6	96.8
18	58.79	56	46	1.4	1.01	80.4	86.1
19	35.14	56	54	0.1	-0.66	98.2	96.8
20	44.81	56	52	0.79	0.02	94.6	93.9
21	40.71	56	53	0.16	-0.94	96.4	95.4
22	48.08	56	51	0.44	-0.75	94.6	92.4
23	35.14	56	54	0.09	-0.69	98.2	96.8
24	60.35	56	45	1.56	1.4	82.1	84.8
25	44.77	57	53	0.41	-0.61	94.7	94
26	61.72	57	45	0.81	-0.45	87.7	83.7
27	35.12	57	55	1.2	0.6	94.7	96.9
28	26.22	57	56	0.04	-0.94	98.2	98.2
29	40.68	57	54	1.4	0.69	96.5	95.5
30	73.82	57	34	0.99	0.02	68.4	71.2
31	44.77	57	53	0.67	-0.14	94.7	94
32	66.72	57	41	1.9	2.69	75.4	78
33	35.12	57	55	0.16	-0.51	98.2	96.9
34	53.1	57	50	1.28	0.65	86	90
35	67.83	57	40	0.99	0.03	78.9	76.7
36	63.07	57	44	1.17	0.62	75.4	82.3

Table 14.
Grade 8 Maze Measures.

Item	Measure	Count	Score	Outfit MSQ	Outfit ZSTD	Obs. Match	Exp. Match
1	50.97	52	45	0.3	-1.24	94.2	89.2
2	42.34	52	48	0.81	0.17	88.5	92.7
3	42.34	52	48	0.48	-0.25	92.3	92.7
4	42.34	52	48	0.3	-0.57	92.3	92.7
5	50.97	52	45	0.48	-0.72	90.4	89.2
6	57.20	52	42	1.06	0.28	82.7	85.8
7	50.97	52	45	0.88	0.04	90.4	89.2
8	45.63	52	47	0.68	-0.09	88.5	91.7
9	45.63	52	47	0.19	-1.1	96.2	91.7
10	38.36	52	49	0.29	-0.34	94.2	94.2
11	63.73	52	38	0.81	-0.46	82.7	82.2
12	87.69	52	17	1.97	1.88	71.2	72.7
13	57.77	50	40	0.54	-0.98	94	85.6
14	59.59	50	39	1.33	0.82	80	84.7
15	33.38	50	48	3.15	1.46	96	96
16	42.64	50	46	3.02	1.6	92	92.5
17	48.84	50	44	0.19	-1.37	94	90.2
18	51.40	50	43	0.25	-1.4	94	88.9
19	48.84	50	44	1.4	0.72	90	90.2
20	61.30	50	38	1.98	2.03	70	83.9
21	67.32	50	34	1.31	1.04	66	79.8
22	38.67	49	46	4.46	1.9	95.9	93.9
23	48.93	49	43	0.79	-0.02	93.9	90
24	55.94	49	40	1.68	1.22	87.8	86.4
25	51.01	51	44	0.62	-0.41	82.4	89
26	42.37	51	47	0.36	-0.45	92.2	92.6
27	48.50	51	45	0.36	-0.85	90.2	90.3
28	48.50	51	45	0.6	-0.33	82.4	90.3
29	51.01	51	44	0.27	-1.31	90.2	89
30	42.37	51	47	0.33	-0.5	92.2	92.6
31	51.01	51	44	0.7	-0.26	86.3	89
32	55.35	51	42	1.19	0.52	84.3	86.7
33	55.35	51	42	0.4	-1.27	84.3	86.7
34	55.35	51	42	0.79	-0.24	80.4	86.7
35	33.19	51	49	0.62	0.16	96.1	96.1
36	33.19	51	49	0.51	0.05	96.1	96.1

Table 15.
Number of poorly fitting items in each maze passage.

	Gr. 3	Gr. 4	Gr. 5	Gr. 6	Gr. 7	Gr. 8
Passage 1*	4	5	5	1	6	7
Passage 2*	2	4	3	2	6	5
Passage 3*	5	2	1	1	3	5

* Each passage has 12 multiple-choice questions

Table 16.
Categorization of Items by Grades and the Mean Square Outfit Values.

	Number of Items					
	Gr. 3	**Gr. 4**	**Gr. 5**	**Gr. 6**	**Gr. 7**	**Gr. 8**
Productive Items (.5 ≤ Mean Square Outfit ≤ 1.5)	25	25	27	31	21	18
Over-Fit Items (Mean Square Outfit < .5)	8	5	7	4	13	13
Under-fit Items (1.5 <Mean Square Outfit ≤2.0)	2	6	1	1	2	1
Poor-fitting Items (Mean Square Outfit < 2.0)	1	0	1	0	0	4

Tables 17-32:
Parameter files for the Grades 3-8 Maze Measures under the 2PL Model

Table 17.
Grade 3 Maze 1.

	Intercept	Intercept SE	Slope	Slope SE	Difficulty	Difficulty SE
Item 01	1.34	0.33	0.94	0.30	-1.43	0.35
Item 02	1.03	0.21	0.53	0.16	-1.96	0.64
Item 03	1.41	0.33	0.94	0.31	-1.51	0.48
Item 04	1.88	0.45	1.11	0.34	-1.69	0.39
Item 05	0.42	0.20	0.81	0.23	-0.52	0.26
Item 06	1.05	0.23	0.63	0.19	-1.67	0.55
Item 07	2.42	0.64	1.45	0.46	-1.66	0.30
Item 08	8.69	5.18	6.69	3.48	-1.30	0.20
Item 09	1.89	0.59	1.38	0.53	-1.37	0.30
Item 10	0.48	0.18	0.71	0.21	-0.67	0.30
Item 11	2.17	0.41	0.94	0.31	-2.30	0.57
Item 12	1.75	0.43	1.09	0.34	-1.61	0.36

Table 18.
Grade 3 Maze 2.

	Intercept	Intercept SE	Slope	Slope SE	Difficulty	Difficulty SE
Item 01	0.81	0.33	1.08	0.31	-0.75	0.28
Item 02	5.08	0.61	1.53	0.31	-3.33	0.50
Item 03	7.76	6.18	12.42	5.10	-0.63	0.23
Item 04	1.32	0.27	0.74	0.22	-1.78	0.55
Item 05	1.09	0.24	0.61	0.16	-1.78	0.53
Item 06	2.67	0.49	1.17	0.32	-2.29	0.56
Item 07	1.96	0.72	1.69	0.72	-1.16	0.38
Item 08	2.29	0.64	1.49	0.53	-1.54	0.41
Item 09	1.55	0.47	1.26	0.49	-1.23	0.42
Item 10	2.16	0.50	1.14	0.41	-1.89	0.59
Item 11	0.06	0.19	0.61	0.18	-0.10	0.30
Item 12	1.08	0.24	0.67	0.22	-1.62	0.62

Table 19.
Grade 3 Maze 3.

	Intercept	Intercept SE	Slope	Slope SE	Difficulty	Difficulty SE
Item 01	0.51	0.21	0.93	0.23	-0.55	0.25
Item 02	0.68	0.16	0.52	0.14	-1.31	0.44
Item 03	1.50	0.48	0.72	0.26	-2.08	0.65
Item 04	2.69	1.04	1.49	0.61	-1.80	0.28
Item 05	1.78	0.51	1.14	0.34	-1.56	0.35
Item 06	1.78	1.05	2.23	0.90	-0.80	0.24
Item 07	1.02	0.26	0.98	0.24	-1.04	0.28
Item 08	2.72	1.09	2.14	0.78	-1.27	0.20
Item 09	0.94	0.16	0.48	0.13	-1.95	0.61
Item 10	1.88	0.45	1.23	0.33	-1.53	0.34
Item 11	0.53	0.29	1.52	0.31	-0.35	0.15
Item 12	1.41	0.46	1.30	0.37	-1.08	0.22

Table 20.
Grade 4 Maze 1.

	Intercept	Intercept SE	Slope	Slope SE	Difficulty	Difficulty SE
Item 01	2.81	0.26	0.76	0.24	-3.68	1.16
Item 02	3.00	0.22	0.85	0.28	-3.55	1.20
Item 03	4.57	0.49	1.20	0.51	-3.82	1.58
Item 04	1.66	0.22	0.71	0.24	-2.34	0.88
Item 05	12.98	2.86	3.84	1.33	-3.38	0.64
Item 06	0.65	0.14	0.56	0.19	-1.16	0.52
Item 07	2.31	0.39	1.06	0.41	-2.19	0.86
Item 08	15.33	3.48	5.86	1.38	-2.61	0.22
Item 09	3.34	1.44	4.52	1.08	-0.74	0.14
Item 10	3.15	2.04	2.13	0.90	-1.48	0.41
Item 11	7.86	2.28	6.28	1.33	-1.25	0.12
Item 12	2.89	2.58	4.33	2.68	-0.67	0.17

Table 21.
Grade 4 Maze 2.

	Intercept	Intercept SE	Slope	Slope SE	Difficulty	Difficulty SE
Item 01	0.00	0.00	0.00	0.00	0.00	0.00
Item 02	1.43	0.30	0.96	0.25	-1.49	0.37
Item 03	1.88	0.42	0.83	0.31	-2.27	0.76
Item 04	1.63	0.23	0.58	0.17	-2.80	0.79
Item 05	2.67	1.28	1.72	0.78	-1.56	0.31
Item 06	1.66	0.27	0.66	0.19	-2.53	0.76
Item 07	3.55	3.05	2.23	1.73	-1.59	0.29
Item 08	0.55	0.13	0.39	0.12	-1.39	0.60
Item 09	0.85	0.20	0.73	0.20	-1.17	0.36
Item 10	1.61	0.38	0.84	0.27	-1.92	0.48
Item 11	2.12	0.63	1.58	0.43	-1.34	0.21
Item 12	0.25	0.14	0.59	0.15	-0.42	0.29

Table 22.
Grade 4 Maze 3.

	Intercept	Intercept SE	Slope	Slope SE	Difficulty	Difficulty SE
Item 01	1.11	0.24	0.91	0.27	-1.22	0.36
Item 02	4.97	5.51	3.39	3.39	-1.47	0.26
Item 03	0.98	0.22	0.83	0.22	-1.18	0.30
Item 04	0.35	0.13	0.50	0.15	-0.69	0.38
Item 05	0.52	0.13	0.37	0.11	-1.41	0.57
Item 06	3.20	1.08	2.07	1.00	-1.54	0.49
Item 07	0.36	0.15	0.69	0.20	-0.53	0.28
Item 08	2.32	0.55	1.28	0.45	-1.82	0.51
Item 09	4.54	3.10	3.40	2.01	-1.34	0.18
Item 10	1.19	0.21	0.74	0.21	-1.61	0.48
Item 11	4.39	2.67	3.35	1.82	-1.31	0.17
Item 12	1.56	0.24	0.59	0.18	-2.66	0.83

Table 23.
Grade 5 Maze 1.

	Intercept	Intercept SE	Slope	Slope SE	Difficulty	Difficulty SE
Item 01	3.17	0.67	1.15	0.40	-2.75	0.68
Item 02	5.03	1.90	1.49	0.70	-3.38	0.66
Item 03	0.83	0.13	0.68	0.20	-1.22	0.40
Item 04	0.82	0.18	1.05	0.26	-0.79	0.19
Item 05	1.24	0.18	0.73	0.21	-1.70	0.46
Item 06	6.62	2.46	2.11	0.96	-3.14	0.56
Item 07	2.52	0.42	1.08	0.45	-2.33	0.86
Item 08	1.35	0.24	1.22	0.28	-1.10	0.19
Item 09	2.52	0.86	1.70	0.82	-1.48	0.39
Item 10	2.36	0.54	1.32	0.49	-1.79	0.42
Item 11	6.26	3.02	2.67	1.60	-2.35	0.50
Item 12	6.59	6.96	2.91	2.65	-2.26	0.63

Table 24.
Grade 5 Maze 2.

	Intercept	Intercept SE	Slope	Slope SE	Difficulty	Difficulty SE
Item 01	6.87	9.39	4.33	3.99	-1.58	1.13
Item 02	1.43	0.18	0.53	0.16	-2.70	0.88
Item 03	9.15	3.26	3.35	1.39	-2.73	0.46
Item 04	2.73	0.51	1.32	0.52	-2.08	0.72
Item 05	2.00	0.48	1.17	0.47	-1.71	0.58
Item 06	3.33	1.28	1.66	1.02	-2.00	0.96
Item 07	1.70	0.27	0.80	0.24	-2.12	0.62
Item 08	4.74	2.11	2.35	1.44	-2.01	0.73
Item 09	5.70	2.55	2.52	1.66	-2.26	0.89
Item 10	2.42	0.36	0.84	0.28	-2.87	0.91
Item 11	2.19	0.30	0.76	0.25	-2.90	0.95
Item 12	2.65	0.99	1.71	0.82	-1.55	0.49

Table 25.
Grade 5 Maze 3.

	Intercept	Intercept SE	Slope	Slope SE	Difficulty	Difficulty SE
Item 01	2.43	0.29	0.41	0.13	-5.86	1.87
Item 02	1.47	0.16	0.48	0.15	-3.03	0.91
Item 03	2.03	0.31	0.70	0.26	-2.89	0.93
Item 04	1.78	0.25	0.49	0.17	-3.65	1.19
Item 05	1.00	0.12	0.40	0.12	-2.49	0.78
Item 06	1.63	0.18	0.54	0.17	-3.03	0.87
Item 07	1.12	0.23	0.84	0.26	-1.34	0.30
Item 08	2.23	0.38	0.95	0.41	-2.35	0.79
Item 09	8.08	4.39	5.25	2.84	-1.54	0.15
Item 10	0.50	0.10	0.55	0.14	-0.91	0.27
Item 11	0.30	0.12	0.80	0.19	-0.37	0.16
Item 12	2.43	0.29	0.41	0.13	-5.86	1.87

Table 26.
Grade 6 Maze 1.

	Intercept	Intercept SE	Slope	Slope SE	Difficulty	Difficulty SE
Item 01	4.58	0.29	0.42	0.16	-10.89	4.12
Item 02	5.63	0.51	0.67	0.25	-8.35	3.11
Item 03	1.24	0.16	0.32	0.10	-3.84	1.40
Item 04	1.80	0.15	0.27	0.07	-6.65	1.96
Item 05	4.36	0.97	1.01	0.48	-4.33	1.94
Item 06	2.58	0.29	0.43	0.13	-6.01	1.98
Item 07	8.40	0.41	0.45	0.17	-18.52	6.98
Item 08	10.55	0.63	0.51	0.20	-20.52	7.37
Item 09	8.95	0.74	0.88	0.30	-10.21	3.29
Item 10	2.30	4.47	3.23	1.81	-0.71	0.39
Item 11	0.00	0.00	0.00	0.00	0.00	0.00
Item 12	1.52	0.16	0.30	0.09	-5.12	1.77

Table 27.
Grade 6 Maze 2.

	Intercept	Intercept SE	Slope	Slope SE	Difficulty	Difficulty SE
Item 01	4.98	0.97	1.17	0.48	-4.26	1.33
Item 02	2.15	0.46	1.15	0.48	-1.87	0.61
Item 03	0.58	0.11	0.31	0.09	-1.87	0.72
Item 04	4.30	1.58	1.61	0.93	-2.67	0.83
Item 05	1.36	0.20	0.37	0.12	-3.72	1.37
Item 06	-0.35	0.17	1.24	0.45	0.28	0.12
Item 07	0.05	0.10	0.22	0.06	-0.21	0.49
Item 08	8.06	7.78	2.83	2.88	-2.85	0.88
Item 09	2.80	0.93	1.39	0.71	-2.01	0.55
Item 10	2.06	0.31	0.87	0.32	-2.38	0.74
Item 11	3.26	1.17	1.73	1.06	-1.89	0.84
Item 12	0.88	0.19	0.85	0.28	-1.04	0.37

Table 28.
Grade 6 Maze 3.

	Intercept	Intercept SE	Slope	Slope SE	Difficulty	Difficulty SE
Item 01	2.07	0.32	0.85	0.29	-2.43	0.78
Item 02	0.63	0.26	1.27	0.29	-0.50	0.15
Item 03	0.15	0.12	0.39	0.11	-0.37	0.33
Item 04	2.41	0.77	0.90	0.36	-2.67	0.87
Item 05	-0.37	0.12	0.29	0.08	1.30	0.47
Item 06	2.34	0.39	1.00	0.27	-2.34	0.55
Item 07	0.35	0.12	0.39	0.11	-0.90	0.45
Item 08	0.45	0.27	1.41	0.38	-0.32	0.14
Item 09	3.41	2.39	2.35	1.84	-1.45	0.47
Item 10	3.22	0.32	0.94	0.28	-3.43	0.93
Item 11	5.00	5.57	3.18	3.33	-1.57	0.35
Item 12	0.82	0.14	0.43	0.13	-1.88	0.69

N. B. Grade 7 data would not converge; therefore, there were no parameter files for Grade 7.

Table 29.
Grade 8 Maze 1.

	Intercept	Intercept SE	Slope	Slope SE	Difficulty	Difficulty SE
Item 01	4.46	2.85	2.81	1.99	-1.59	0.47
Item 02	3.09	0.50	0.82	0.29	-3.78	1.32
Item 03	4.47	1.36	1.87	0.71	-2.39	0.43
Item 04	3.13	0.61	1.12	0.36	-2.79	0.88
Item 05	2.39	0.56	1.12	0.42	-2.14	0.72
Item 06	2.08	0.55	1.53	0.62	-1.36	0.45
Item 07	3.15	1.07	1.64	0.70	-1.92	0.43
Item 08	2.19	0.58	1.63	0.43	-1.35	0.32
Item 09	6.16	3.57	2.48	1.41	-2.48	0.29
Item 10	5.71	4.69	2.74	1.87	-2.08	0.25
Item 11	1.35	0.24	0.93	0.32	-1.45	0.52
Item 12	0.00	0.13	0.81	0.24	0.00	0.16

Table 30.
Grade 8 Maze 2.

	Intercept	Intercept SE	Slope	Slope SE	Difficulty	Difficulty SE
Item 01	2.91	2.84	8.08	4.40	-0.36	0.18
Item 02	6.58	0.50	2.48	0.43	-2.65	0.43
Item 03	3.19	0.42	0.84	0.28	-3.79	1.24
Item 04	2.47	0.36	1.30	0.33	-1.90	0.50
Item 05	4.45	4.81	8.25	4.55	-0.54	0.23
Item 06	4.03	3.78	8.37	4.53	-0.48	0.28
Item 07	4.03	0.28	1.09	0.32	-3.69	1.08
Item 08	2.21	0.19	0.75	0.24	-2.94	1.00
Item 09	13.92	1.41	5.72	0.95	-2.43	0.27
Item 10	5.22	0.58	1.07	0.21	-4.90	0.56
Item 11	17.24	1.45	6.36	0.64	-2.71	0.16
Item 12	15.84	1.70	7.09	1.06	-2.23	0.22

Table 31.
Grade 8 Maze 3.

	Intercept	Intercept SE	Slope	Slope SE	Difficulty	Difficulty SE
Item 01	2.36	0.59	1.50	0.53	-1.57	0.44
Item 02	6.43	4.79	2.84	2.16	-2.26	0.62
Item 03	2.95	1.28	1.63	0.55	-1.81	0.23
Item 04	3.53	0.61	1.11	0.39	-3.17	0.98
Item 05	3.67	0.73	1.57	0.44	-2.34	0.45
Item 06	6.47	5.13	2.98	2.32	-2.17	0.49
Item 07	2.65	0.56	1.46	0.59	-1.81	0.58
Item 08	2.33	0.77	2.11	1.01	-1.10	0.35
Item 09	3.48	1.68	2.26	1.15	-1.54	0.27
Item 10	1.65	0.50	1.49	0.52	-1.11	0.33
Item 11	4.09	1.83	1.66	0.83	-2.46	0.88
Item 12	7.19	3.97	2.29	1.48	-3.14	0.74

Table 32-46:

The Score files of the Mazes Measures under the 2PL Model

Table 32.
Grade 3 Maze 1 Score Files.

Total Items	No. of Correct Items	Proportion Correct	No. of Cases	Scale Scores		Standard Errors of Scale Scores	
				Min.	Max.	Min.	Max.
12	12	100	24	0.91	0.91	0.74	0.74
12	11	91.67	27	0.08	0.47	0.57	0.66
12	10	83.33	12	-0.34	-0.09	0.4	0.51
12	9	75	11	-1.33	-0.3	0.03	0.44
12	8	66.67	3	-0.82	-0.58	0.41	0.47
12	7	53.33	5	-1.34	-0.87	0.08	0.47
12	6	50	1	-1.29	-1.29	0.19	0.19
12	5	41.67	1	-1.48	-1.48	0.34	0.34
12	4	33.33	3	-1.9	-1.76	0.45	0.46
12	3	25	1	-2.13	-2.13	0.35	0.35
12	2	16.67	1	-2.27	-2.27	0.31	0.31
12	1	8.33	1	-2.29	-2.29	0.32	0.32
12	0	0	0	--	--	--	--

* Scale scores and standard errors of scale scores vary depending on which questions are answered correctly. To show the range, the authors reported the maximum and minimum of the scale scores and standard errors of scale scores.

Table 33.
Grade 3 Maze 2 Score Files.

Total Items	No. of Correct Items	Proportion Correct	No. of Cases	Scale Scores		Standard Errors of Scale Scores	
				Min.	Max.	Min.	Max.
12	12	100	22	0.93	0.93	0.75	0.75
12	11	91.67	22	-0.13	0.42	0.42	0.65
12	10	83.33	20	-0.31	0.05	0.21	0.53
12	9	75	6	-0.35	-0.2	0.1	0.36
12	8	66.67	9	-0.35	-0.3	0.06	0.23
12	7	53.33	0	--	--	--	--
12	6	50	3	-1.37	-0.36	0.02	0.29
12	5	41.67	0	--	--	--	--
12	4	33.33	3	-1.48	-0.36	0.04	0.38
12	3	25	0	--	--	--	--
12	2	16.67	4	-2.47	-2.21	0.34	0.45
12	1	8.33	1	-2.89	-2.89	0.48	0.48
12	0	0	1	-3.11	999	0.46	0.46
*9	8	88.89	0	--	--	--	--

Table 34.
Grade 3 Maze 3 Score Files.

Total Items	No. of Correct Items	Proportion Correct	No. of Cases	Scale Scores		Standard Errors of Scale Scores	
				Minimum	Maximum	Minimum	Maximum
12	12	100	12	0.93	0.93	0.67	0.67
12	11	91.67	24	0.08	0.61	0.5	0.59
12	10	83.33	14	-0.4	0.32	0.24	0.54
12	9	75	6	-0.41	-0.34	0.24	0.26
12	8	66.67	4	-0.71	-0.4	0.24	0.42
12	7	53.33	2	-0.79	-0.45	0.26	0.43
12	6	50	2	-1.25	-1.05	0.16	0.38
12	5	41.67	3	-1.32	-1.27	0.14	0.22
12	4	33.33	5	-1.56	-1.36	0.28	0.41
12	3	25	5	-1.73	-1.68	0.44	0.44
12	2	16.67	2	-2.11	-2.02	0.34	0.36
12	1	8.33	3	-2.1	-2.1	0.34	0.34
12	0	0	1	--	--	--	--

Table 35.
Grade 4 Maze 1 Score Files.

Total Items	No. of Correct Items	Proportion Correct	No. of Cases	Scale Scores		Standard Errors of Scale Scores	
				Minimal	Maximum	Minimal	Maximum
12	12	100	37	.53	.53	.69	.69
12	11	91.67	18	-.87	.16	.30	.55
12	10	83.33	7	-.80	-.10	.29	.38
12	9	75	2	-1.09	-.25	.03	.21
12	8	66.67	1	-1.08	-1.08	.04	.04
12	7	53.33	0	-1.09	-1.09	.07	.07
12	6	50	2	-2.70	-1.08	.03	.14
12	5	41.67	0	--	--	--	--
12	4	33.33	1	-2.49	-1.89	.23	.34
12	3	25	0	--	--	--	--
12	2	16.67	1	-2.29	-2.29	.40	.40
12	1	8.33	0	--	--	--	--
12	0	0	1	-3.47	999.00	.07	.07

* Scale scores and standard errors of scale scores vary depending on which questions are answered correctly. To show the range, the authors reported the maximum and minimum of the scale scores and standard errors of scale scores.

Table 36.
Grade 4 Maze 2 Score Files.

Total Items	No. of Correct Items	Proportion Correct	No. of Cases	Scale Scores		Standard Errors of Scale Scores	
				Min.	Max.	Min.	Max.
12	11	91.67	17	.89	.89	.78	.78
12	10	83.33	20	-.28	.51	.54	.72
12	9	75	14	-1.00	.06	.41	.62
12	8	66.67	3	-.54	-.27	.52	.54
12	7	53.33	6	-1.32	-.58	.23	.52
12	6	50	3	-1.64	.49	.28	.84
12	5	41.67	5	-1.93	-1.27	.23	.41
12	4	33.33	0	--	--	--	--
12	3	25	1	-2.13	-2.13	.30	.30
12	2	16.67	0	--	--	--	--
12	1	8.33	0	--	--	--	--
12	0	0	0	--	--	--	--

Table 37.
Grade 4 Maze 3 Score Files.

Total Items	No. of Correct Items	Proportion Correct	No. of Cases	Scale Scores		Standard Errors of Scale Scores	
				Min.	Max.	Min.	Max.
12	12	100	18	.95	.95	.77	.77
12	11	91.67	14	.20	.61	.62	.71
12	10	83.33	15	-.57	.23	.42	.63
12	9	75	14	-1.29	-.14	.05	.49
12	8	66.67	0	--	--	--	--
12	7	53.33	3	-1.29	-1.28	.03	.10
12	6	50	2	-1.29	-1.29	.03	.05
12	5	41.67	1	-1.29	-1.29	.03	.03
12	4	33.33	1	-2.07	-2.07	.33	.33
12	3	25	1	-2.20	-2.20	.23	.23
12	2	16.67	1	-2.30	-2.30	.31	.31
12	1	8.33	0	--	--	--	--
12	0	0	0	--	--	--	--

Table 37.
Grade 5 Maze 1 Score Files.

Total Items	No. of Correct Items	Proportion Correct	No. of Cases	Scale Scores		Standard Errors of Scale Scores	
				Min.	Max.	Min.	Max.
12	12	100	54	.59	.59	.70	.70
12	11	91.67	26	-.20	.10	.55	.61
12	10	83.33	9	-.84	-.30	.44	.53
12	9	75	7	-.1.15	-.90	.34	.42
12	8	66.67	1	-1.21	-1.21	.35	.35
12	7	58.33	2	-1.90	-.90	.20	.45
12	6	50	0	--	--	--	--
12	5	41.67	1	-1.93	-1.93	.15	.15
12	4	33.33	1	-2.15	-2.15	.35	.35
12	3	25	2	-2.77	-2.75	.16	.17
12	2	16.67	0	--	--	--	--
12	1	8.33	1	-3.06	-3.06	.39	.39
12	0	0	0	--	--	--	--
*9	7	77.78	1	-.90	-.90	.45	.45

* One student only answered 9 questions, instead of 12 questions.

Table 39.
Grade 5 Maze 2 Score Files.

Total Items	No. of Correct Items	Proportion Correct	No. of Cases	Scale Scores		Standard Errors of Scale Scores	
				Minimum	Maximum	Minimum	Maximum
12	12	100	67	.50	.50	.82	.82
12	11	91.67	26	-1.02	-.04	.32	.72
12	10	83.33	12	-1.14	-.75	.16	.52
12	9	75	1	-1.11	-1.11	.20	.20
12	8	66..67	0	--	--	--	--
12	7	58.33	0	--	--	--	--
12	6	50	0	--	--	--	--
12	5	41.67	0	--	--	--	--
12	4	33.33	0	--	--	--	--
12	3	25	1	-2.90	-2.90	.14	.14
12	2	16.67	0	--	--	--	--
12	1	8.33	1	-2.98	-2.98	.25	.25
12	0	0	0	--	--	--	--

Table 40.
Grade 5 Maze 3 Score Files.

Total Items	No. of Correct Items	Proportion Correct	No. of Cases	Scale Scores		Standard Errors of Scale Scores	
				Minimum	Maximum	Minimum	Maximum
12	12	100	31	.87	.87	.79	.79
12	11	91.67	25	-1.34	.47	.16	.74
12	10	83.33	26	-.57	.05	.59	.68
12	9	75	18	-.84	-.52	.53	.60
12	8	66..67	4	-1.70	-.82	.43	.54
12	7	58.33	2	-1.15	-1,15	.37	.38
12	6	50	0	--	--	--	--
12	5	41.67	0	--	--	--	--
12	4	33.33	0	--	--	--	--
12	3	25	1	-2.55	-2.55	.48	.48
12	2	16.67	0	--	--	--	--
12	1	8.33	1	-3.15	-3.15	.55	.55
12	0	0	0	--	--	--	--
*2	2	100	1	.14	.14	.99	.99

* One student only answered two questions.

Table 41.
Grade 6 Maze 1 Score Files.

Total Items	No. of Correct Items	Proportion Correct	No. of Cases	Scale Scores		Standard Errors of Scale Scores	
				Minimal	Maximum	Minimal	Maximum
11	11	100	8	0.83	0.83	0.92	0.92
11	10	90.91	19	-1.24	0.48	0.52	0.81
11	9	81.82	24	-0.33	0.2	0.56	0.69
11	8	72.73	5	-0.36	-0.02	0.56	0.6
11	7	63.64	7	-2.49	-0.26	0.56	0.86
11	6	54.55	3	-2.57	-0.97	0.52	0.87
11	5	45.46	1	-2.92	-2.92	0.86	0.86
11	4	36.37	1	-1.22	-1.22	0.51	0.51
11	3	27.28	0	--	--	--	--
11	2	18.19	0	--	--	--	--
11	1	9.10	0	--	--	--	--
11	0	0	0	--	--	--	--

* Scale scores and standard errors of scale scores vary depending on which questions are answered correctly. To show the range, the authors reported the maximum and minimum of the scale scores and standard errors of scale scores.

Table 42.
Grade 6 Maze 2 Score Files.

Total Items	No. of Correct Items	Proportion Correct	No. of Cases	Scale Scores		Standard Errors of Scale Scores	
				Min.	Max.	Min.	Max.
12	12	100	9	0.95	0.95	0.7	0.7
12	11	91.67	19	0.07	0.77	0.61	0.68
12	10	83.33	25	-0.85	0.54	0.5	0.65
12	9	75	8	-0.97	-0.25	0.48	0.58
12	8	66.67	1	-0.7	-0.7	0.53	0.53
12	7	53.33	2	-1.5	-1.18	0.45	0.45
12	6	50	2	-1.45	-1.23	0.45	0.54
12	5	41.67	1	-2.55	-2.55	0.41	0.41
12	4	33.33	0	--	--	--	--
12	3	25	0	--	--	--	--
12	2	16.67	1	-3.21	-3.21	0.41	0.41
12	1	8.33	0	--	--	--	--
12	0	0	0	--	--	--	--

Table 43.
Grade 6 Maze 3 Score Files.

Total Items	No. of Correct Items	Proportion Correct	No. of Cases	Scale Scores		Standard Errors of Scale Scores	
				Min.	Max.	Min.	Max.
12	12	100	5	1.17	1.17	0.76	0.76
12	11	91.67	14	0.37	0.91	0.62	0.72
12	10	83.33	11	-0.03	0.6	0.53	0.66
12	9	75	17	-0.39	0.33	0.44	0.61
12	8	66.67	8	-1.32	-0.29	0.16	0.47
12	7	53.33	7	-1.28	-0.44	0.19	0.48
12	6	50	2	-1.34	-1.22	0.19	0.28
12	5	41.67	1	-1.22	-1.22	0.28	0.28
12	4	33.33	1	-1.54	-1.54	0.4	0.4
12	3	25	1	-2.20	-2.2	0.32	0.32
12	2	16.67	0	--	--	--	--
12	1	8.33	1	-2.67	0.16	0.51	0.99
12	0	0	0	--	--	--	--

Table 44.
Grade 8 Maze 1 Score Files.

Total Items	No. of Correct Items	Proportion Correct	No. of Cases	Scale Scores		Standard Errors of Scale Scores	
				Minimal	Maximum	Minimal	Maximum
12	12	100	34	.64	.64	.71	.71
12	11	91.67	27	.40	.04	.53	.61
12	10	83.33	7	-.93	-.40	.35	.53
12	9	75	2	-1.11	-1.06	.24	.26
12	8	66.67	3	-1.38	-1.10	.24	.38
12	7	53.33	0	--		--	
12	6	50	0	--		--	
12	5	41.67	0	--		--	
12	4	33.33	3	-2.30	-1.94	.18	.40
12	3	25	0	--		--	
12	2	16.67	1	.17	.17	.82	.82
12	1	8.33	2	-2.80	-2.66	.19	.29
12	0	0	0	--		--	

* Scale scores and standard errors of scale scores vary depending on which questions are answered correctly. To show the range, the authors reported the maximum and minimum of the scale scores and standard errors of scale scores.

Table 45.
Grade 8 Maze 2 Score Files.

Total Items	No. of Correct Items	Proportion Correct	No. of Cases	Scale Scores		Standard Errors of Scale Scores	
				Min.	Max.	Min.	Max.
12	12	100	36	.66	.66	.58	.58
12	11	91.67	24	-.45	.31	.01	.48
12	10	83.33	7	-.45	-.08	.00	.43
12	9	75	2	-.88	-.45	.00	.42
12	8	66.67	0	--	--	--	--
12	7	53.33	2	-2.09	-2.09	.01	.02
12	6	50	0	--	--	--	--
12	5	41.67	0	--	--	--	--
12	4	33.33	1	-2.92	-2.92	.09	.09
12	3	25	1	-2.95	-2.95	.19	.19
12	2	16.67	2	-2.98	-2.93	.12	.13
12	1	8.33	0	--	--	--	--
12	0	0	0	--	--	--	--
*9	8	88.89	1	-0.45	-0.45	0.05	0.05

* One student only answered nine questions.

Table 46.
Grade 8 Maze 3 Score Files.

Total Items	No. of Correct Items	Proportion Correct	No. of Cases	Scale Scores		Standard Errors of Scale Scores	
				Minimum	Maximum	Minimum	Maximum
12	12	100	61	.39	.39	.70	.70
12	11	91.67	8	-.69	-.30	.45	.52
12	10	83.33	2	-1.07	-.94	.25	.36
12	9	75	1	-1.17	-1.17	.20	.20
12	8	66.67	2	-1.41	-1.39	.37	.39
12	7	53.33	0	--	--	--	--
12	6	50	1	-1.92	-1.92	.12	.12
12	5	41.67	2	-1.94	-1.94	.06	.08
12	4	33.33	1	-2.36	-2.36	.39	.39
12	3	25	0	--	--	--	--
12	2	16.67	1	-2.73	-2.73	.13	.13
12	1	8.33	1	-2.82	-2.82	.25	.25
12	0	0	0	--	--	--	--

Table 47.
Comparison of Item Difficulty among CTT, 1PL and 2PL Models: Grade 3 Measures.

	Estimated Item Difficulty			Difficulty Ranking		
Item	CTT	1PL	2PL	CTT	1PL	2PL
1	0.81	50.37	-1.43	27	27	14
2	0.8	51.39	-1.96	14	14	11
3	0.88	43.05	-1.51	4	15	18
4	0.9	39.94	-1.69	8	4	27
5	0.66	62.13	-0.52	3	8	2
6	0.84	47	-1.67	7	18	33
7	0.88	43.05	-1.66	11	3	22
8	0.9	39.94	-1.3	12	7	28
9	0.85	47	-1.37	18	11	16
10	0.64	62.84	-0.67	22	12	17
11	0.88	43.05	-2.3	28	22	4
12	0.88	43.05	-1.61	9	28	6
13	0.8	51.36	-0.75	19	29	7
14	0.91	38.17	-3.33	29	19	24
15	0.76	39.93	-0.63	6	6	12
16	0.76	55.09	-1.78	20	9	29
17	0.74	56.78	-1.78	34	34	20
18	0.88	43.03	-2.29	030	20	34
19	0.85	46.98	-1.16	32	32	3
20	0.84	48.15	-1.54	1	30	1
21	0.78	53.29	-1.23	2	1	8
22	0.87	44.43	-1.89	13	13	26
23	0.52	70.53	-0.1	36	2	9
24	0.74	56.78	-1.62	21	36	32
25	0.69	60.64	-0.55	15	21	21
26	0.7	59.88	-1.31	16	33	19
27	0.92	36.6	-2.08	33	16	36
28	0.87	44.91	-1.8	17	31	31
29	0.85	46.24	-1.56	24	17	30
30	0.83	49.83	-0.8	31	24	13
31	0.74	56.62	-1.04	26	26	10
32	0.83	48.7	-1.27	25	25	15
33	0.76	54.85	-1.95	5	5	25
34	0.84	47.5	-1.53	10	10	5
35	0.62	64.94	-0.35	35	35	35
36	0.8	51.96	-1.08	23	23	23

Appendix A

GOMs Maze Measure Internal Review

Review Rubric:

- *Readability***:** Determine the reading level using the Flesch-Kincaid Readability Index; note this directly on the passage. For example, the range of readability index for an appropriate third grade passage is between 3.0 and 3.9. The targeted readability index for GOM Maze passage is in the mid-range. For example, the readability index for an appropriate grade three passage is between 3.4 and 3.7.
- *Appropriateness of language*: Are the questions and response options written so that students in the assigned grade can understand the meaning of the problem? Is the vocabulary written at the appropriate grade level?
- *Appropriateness of concepts*: Can students in the assigned grade complete the task? Is this information taught within the normal curriculum of the grade?
- *Alignment to Test Specifications*: Does the measure accurately reflect the test specifications that are identified in the review document?
- *Bias in language or graphics*: Does the item require background knowledge unrelated to the concept being tested that would differ for students with different backgrounds? Is the language sensitive to students from diverse backgrounds?

Table A1.
Grade 3 Measures.

	Passage 1	**Passage 2**	**Passage 3**
Readability Index	3.1	3.2	3.7
Appropriateness of language	Grade-level Appropriate	Grade-level Appropriate	Grade-level Appropriate
Appropriateness of concepts	Grade-level Appropriate	Grade-level Appropriate	Grade-level Appropriate
Words per sentence	10.5	10.7	10.6
Biases in language	None	None	None
Suggestions for revisions	None	Children of certain subgroups might not know what "sand toys" are.	None

Table A2.
Grade 4 Measures.

	Passage 1	**Passage 2**	**Passage 3**
Readability Index	4.6	4.5	4.7
Appropriateness of language	Grade-level Appropriate	The content is within the school settings. Students should know what "time table means" if they are in third and fourth grade.	Grade-level Appropriate
Appropriateness of concepts	Grade-level Appropriate	Grade-level Appropriate	Grade-level Appropriate
Words per sentence	12.5	11.6	11.4
Biases in language	None	None	It assumed students know the belt system in karate schools.
Suggestions for revisions	None	None	For the last omission, ELL students might choose "won" instead of "completed."

Table A3.
Grade 5 Measures.

	Passage 1	**Passage 2**	**Passage 3**
Readability Index	5.5	5.1	5.7
Appropriateness of language	Some students might not know ***regular*** can be used as a noun.	***Just knowing that he wasn't the only newcomer made Mark feel better.*** This sentence appears fragmented. Consider revise.	Grade-level Appropriate
Appropriateness of concepts	Grade-level Appropriate	Grade-level Appropriate	Grade-level Appropriate
Words per sentence	10.8	13.4	12.6
Biases in language	The writer assumed students know "mysteries" as a literary genre and "young adult" referred to upper grade students.	None	None
Suggestions for revisions	Change "mystery" to fiction. Change the omitted word from ***young*** to ***adults*** (with ***painters, athletes, books*** as distractors).	It made him feel better knowing that he wasn't the only newcomer. Add a comma after the phrase ***by recess*** and ***by lunchtime***. It will improve the clarity of the sentence.	None

Table A4.
Grade 6 Measures.

	Passage 1	**Passage 2**	**Passage 3**
Readability Index	5.6	5.2	6.5
Appropriateness of language	*"Tyrone drew closer to the door and he could see that it was wide-open, leading into a narrow hallway that ended at the base of a tall staircase."* The sentence is wordy, consider revise.	Grade-level Appropriate	The passage is challenging for Grade 6 students.
Appropriateness of concepts	Grade-level Appropriate	Grade-level Appropriate	Grade-level Appropriate
Words per sentence	12.4	12.1	14.9
Biases in language	Halloween is an American holidays, Some new immigrant families or other ethnic groups might not be familiar with it. The lack of background on Halloween is a barrier of comprehension.	Children of certain subgroups might not know what "sand toys" are.	This is a challenging passage for ELLs and the concept of "stark" can be biased against students with visual impairment.
Suggestions for revisions	Tyrone [carefully, cautiously, hastily or recklessly] approached the front door, ... Carefully and cautiously both work in the sentence. Consider revision. Suggestion: change *carefully* to *carelessly*.	While Heather really liked to play, she only felt comfortable [performing, playing, dancing, relaxing] in front of her parents in the comfort of her own home. In this case, both playing and performing work in the sentence, consider revising the distractors. After the band was done [N.B. change was done to finished], they gathered their [meal, homework, books, music] and got inline to walk out on the stage. (Students might consider it is weird to bring books on the stage. Consider revising the distractors.	Revise the following sentences: She thought the wide-open plains and distant rolling hill were stark, yet [beautiful, average, interesting or frightful]. In this sentence, beautiful and interesting are both reasonable answers. One especially [frigid, crisp, sweltering and boiling] day, they decided to go swimming in the river. (Change boiling to blistering). They finally arrived at their house, soaking wet and [snuck, dashed, waltzed and scurried] inside. (Cases can be made for the answers dashed, and scurried.)

Table A5.
Grade 7 Measures.

	Passage 1	**Passage 2**	**Passage 3**
Readability Index	6.2	6.8	6.6
Appropriateness of language	Grade-level Appropriate	Grade-level Appropriate	If students do not know what choreography is, the passage can be difficult to understand.
Appropriateness of concepts	Grade-level Appropriate	Grade-level Appropriate	Grade-level Appropriate
Words per sentence	14.1	18.5	14.2
Biases in language	None	None	Some students might not know what "homecoming rally" is.
Suggestions for revisions	None	None	Calligraphy can be a great distractor for the answer, choreography.

Table A6.
Grade 8 Measures.

	Passage 1	**Passage 2**	**Passage 3**
Readability Index	7.5	7.9	7.5
Appropriateness of language	Grade-level Appropriate	Grade-level Appropriate	Grade-level Appropriate
Appropriateness of concepts	Grade-level Appropriate	Grade-level Appropriate	Grade-level Appropriate
Words per sentence	19.3	17.0	16.4
Biases in language	None	None	None
Suggestions for revisions	None	None	None

Appendix B: Directions for Administration

Project INFORM

Directions for Administration

My name is _____. Today we are going to work with you in math and reading. You will work on the computer to complete the tasks.

Please your very best job on the reading problems and math problems. Sometimes you will be asked to read some sentences or some paragraphs. Please make sure you read every word on the sentences or paragraphs. I know it might be tempting to NOT read the sentences or paragraphs, but we really want to make sure you read all of the words. Sometimes you will be asked questions about the reading that you do, so please do your best reading.

To solve the math problems, you will have scratch paper and a pencil to use.

Each of you will have a different set of tasks so it is very important to FOLLOW DIRECTIONS.
Some of you might have some more stories and fewer math problems or more math problems and fewer stories. So it doesn't matter who finishes first.
When you finish, please raise your hand and we will excuse you.
Is everyone ready? Does anyone have any questions? [Wait for questions.]
When we excuse you to the computers, please find your teacher's name on the list. Then find your name on the list.
DO NOT GO ON UNTIL WE HAVE CHECKED TO MAKE SURE YOU HAVE ALL OF THE CORRECT INFORMATION.

Are you ready? [Excuse students one at a time; helper will help them get set up on the computer.]
What do you do first? *Find your teacher's name, then your name.*
What do you do after you have selected your name? *Wait for the teacher.*
What do you do when you're finished? *Raise your hand.*