7-11-2016

Getting More From Your Maze: Examining Differences in Distractors

Sarah J. Conoyer

Erica S. Lembke

John L. Hosp

Christine A. Espin

Michelle K. Hosp

See next page for additional authors

Authors

Sarah J. Conoyer, Erica S. Lembke, John L. Hosp, Christine A. Espin, Michelle K. Hosp, and Apryl L. Poch

Getting More From Your Maze: Examining Differences in Distractors

Sarah J. Conoyer,[1] Erica S. Lembke,[2] John L. Hosp,[3] Christine A. Espin,[4] Michelle K. Hosp,[3] and Apryl L. Poch[2]

[1]Texas A&M University–Commerce, Commerce, Texas, USA;

[2]University of Missouri, Columbia, Missouri, USA;

[3]University of Massachusetts Amherst, Amherst, Massachusetts, USA;

[4]Leiden University, Leiden, The Netherlands

To cite this article: Sarah J. Conoyer, Erica S. Lembke, John L. Hosp, Christine A. Espin, Michelle K. Hosp & Apryl L. Poch (2017) Getting More From Your Maze: Examining Differences in Distractors, Reading & Writing Quarterly, 33:2, 141-154, DOI: https://doi.org/10.1080/10573569.2016.1142913

ABSTRACT

The present study examined the technical adequacy of maze-selection tasks constructed in 2 different ways: typical versus novel. We selected distractors for each measure systematically based on rules related to the content of the passage and the part of speech of the correct choice. Participants included 262 middle school students who were randomly assigned to 1 of the 2 maze formats. Scoring of the maze included both correct and correct-minus-incorrect scores. Students completed 3 criterion-reading tests: the Scholastic Reading Inventory, the AIMSweb R-Maze, and a high-stakes state assessment (the Missouri Assessment Program). Alternate-forms reliability was similar across maze formats; however, with regard to scoring procedure, reliability coefficients were consistently higher for correct than for correct-minus- incorrect scores. Validity coefficients were also similar across format with 1 exception: The coefficients for typical maze scores were stronger when compared with the Missouri Assessment Program scores than the coefficients for novel maze scores.

At the secondary school level, curriculum priorities shift from *learning to read* to *reading to learn* (Alley & Deshler, 1979), yet many students enter secondary school without the reading skills necessary to acquire information from texts. These students experience difficulties across multiple areas of reading, including fluency, vocabulary, and comprehension (Busch & Espin, 2003). It is not just students with diagnosed reading disabilities who struggle with reading. On the 2013 National Assessment of Educational Progress, 64% of eighth graders and 62% of 12th graders scored at or below a basic level in reading (National Center for Education Statistics, 2014). Reading difficulties affect not only academic success but also personal, professional, economic, and familial success. Such dismal outcomes have led some to refer to the reading problems of secondary school students as a "public health crisis" (Fuchs, Fuchs, & Compton, 2010, p. 26) and to call for different and/or more intensive instruction for older struggling readers (Espin, Wallace, Lembke, Campbell, & Long, 2010; Fuchs et al., 2010). Although reading instruction has traditionally fallen under the purview of elementary schools, for struggling readers it may be important to extend instruction into the secondary school years (Espin et al., 2010; Fuchs et al., 2010).

One framework currently gaining popularity as a way of organizing instructional programs for struggling readers is multitiered system of support (MTSS)/response to intervention (RTI; Castillo & Batsche, 2012). An MTSS/RTI framework involves the implementation of tiered, high-quality instruction and the use of screening and progress monitoring data for instructional decision making (Grosche & Volpe, 2013). MTSS/RTI approaches have been implemented primarily at the elementary school level. At the secondary school level, implementation of MTSS/RTI presents unique challenges, not the least of which is the selection of appropriate measures for screening, progress monitoring, and instructional decision making (Fuchs et al., 2010; Prewett et al., 2012). A measurement system often used for screening and progress monitoring within MTSS/RTI at the elementary school level is curriculum-based measurement (CBM; Deno, 1985). The scores produced by CBM measures have been shown to be valid and reliable indicators of performance and of progress for elementary school students[1] (see reviews by Reschly, Busch, Betts, Deno, & Long, 2009;

[1]See Christ, Zopluoglu, Monaghen, and Van Norman (2013) for a discussion of the number of data points needed to produce stable rates of growth with reading-aloud measures.

Wayman, Wallace, Wiley, Tichá, & Espin, 2007). Only recently has research on CBM reading measures been extended to the secondary school level.

Measuring reading performance and progress at the secondary school level

CBM reading research at the secondary school level has focused on two types of measures: reading aloud and maze selection. For reading aloud, students read aloud from text for 1 to 3 min, and the number of words read correctly is scored. For maze selection, students read silently for 1 to 3 min from a passage in which every seventh word is deleted and replaced with a multiple-choice item that includes the correct word and two distractors. Students slash or circle the word that correctly restores meaning to the sentence, and the number of correct or the number of correct-minus-incorrect choices is scored.

Research on the technical adequacy of the scores produced by reading-aloud and maze-selection measures for secondary school students has in general supported the use of both measures as indicators of reading *performance* (Espin & Deno, 1993; Espin & Foegen, 1996; Espin et al., 2010; Hale et al., 2011; Jenkins & Jewell, 1993; Johnson, Semmelroth, Allison, & Fritsch, 2013; Silberglitt, Burns, Madyun, & Lail, 2006; Tichá, Espin, & Wayman, 2009; Tolar et al., 2012). For measuring reading *progress,* stronger support has been found for the use of the maze-selection measure (Espin et al., 2010; Tichá et al., 2009; Tolar et al., 2012).

Although results of the research have been *generally* positive, Tolar et al. (2012) pointed out that results have varied across studies. They highlighted potential reasons for this variability in outcomes, such as differences in the construction and administration of the maze (Tolar et al., 2012). In this research, we examine the effects different approaches to the construction and administration of the maze on the technical adequacy of maze scores. We also examine different approaches to scoring the maze. With regard to maze construction, we examine the effects of different approaches for selecting distractor items. With regard to administration and scoring, we examine the effects of different time frames and the use of correct versus correct-minus-incorrect scoring.

Differences in methods used to select distractor items

Various methods have been used to select the distractor items used in the maze task.

One of the initial uses of the maze as a CBM measure was as a part of a screening instrument called the Basic Academic Skills Samples (BASS; Deno, Maruyama, Espin, & Cohen, 1989; http://www.progressmonitoring.org). In the BASS, the maze was constructed to be a proxy for a reading-aloud measure. Because reading aloud had to be administered one on one, it was time consuming and impractical when large numbers of students had to be tested. Maze selection could be group administered and thus was more efficient for screening and for assessing large numbers of students. The most important rule in the construction of the BASS mazes was that the distractors had to be *clearly* wrong. In this way, students could progress through the text in the same way they would if they were to read the passage aloud, and the scores on the maze would relate to reading-aloud scores. The purpose of the multiple-choice items in the BASS mazes, then, was to estimate how far the students had progressed in 1 min when reading a text silently (S. Deno, personal communication, October 31, 2015).

In the early 1990s, Fuchs and Fuchs included the maze as part of a computer-based system designed to aid teachers in the administration, scoring, and graphing of CBM measures (see Fuchs & Fuchs, 2002). Fairly specific rules were used to create the maze passages for this system. For example, (a) distractors had to be no more than one letter shorter or longer than the correct response; (b) if the word to be deleted was an article or proper noun, then the next appropriate word was selected for deletion; (c) distractors were chosen that did not contextually make sense, rhyme with the correct response, or sound like or look like the correct response; (d) the distractor could not be a nonsense word or be so high in vocabulary level that the student would mistake it for a nonsense word; and (e) the distractor could not require the student to read more than 1.5 lines ahead in the text in order to eliminate it as a correct choice. Several studies at the secondary school level adopted these or similar rules to create their maze passages (e.g., Espin et al., 2010; Johnson et al., 2013; Tichá et al., 2009; Torgesen, Nettle, Howard, & Winterbottom, 2005).

In 2002, Shinn and Shinn proposed a slightly different approach for creating maze distractors. Specifically, they recommended the use of near and far distractors. A *near distractor* was defined as a word that was the same part of speech as the deleted word (e.g., noun, verb, adjective) but that did not preserve meaning or make sense in the sentence. A *far distractor* was defined as a word that was selected randomly from the story but did not make sense in the context of the reading and was not the same part of

speech as the correct choice. Studies by Silberglitt et al. (2006) and Tolar et al. (2012) used the procedures outlined by Shinn and Shinn (2002) to construct mazes.

Few studies have directly compared the effects of different approaches to creating distractor items on the technical adequacy of maze scores. In 1992, Parker, Hasbrouck, and Tindal reviewed research conducted between 1970 and 1990 on the maze as a reading measure. They located 14 published and five unpublished studies that examined the maze as a reading measure. Only one of the studies directly compared different approaches to selecting distractor items (McKenna & Miller, 1980), and this study focused on the effects of distractor selection on item difficulty, not on reliability and validity. However, of importance for the current study is that Parker et al. created a classification scheme that organized the various methods used to select distractor items. The classification scheme took into account three components: (a) whether the distractors were designed to be meaningful or not meaningful in the sentence, (b) whether the distractors were selected from the same or different part of speech as the original word, and (c) whether the distractors were related or unrelated to the content of the passage.

By combining these three factors, Parker et al. (1992) identified six different subtypes of distractors, which they then ordered in terms of difficulty based on how many and which types of discriminations were required and how much of the passage the reader needed to understand to successfully make these discriminations. The most difficult distractors, thus, were those that were meaningful in the sentence, were the same part of speech as the original word, and were related to the content. The easiest distractors were those that were not meaningful in the sentence, were a different part of speech than the original word, and were not content related.

The assumption underlying Parker et al.'s (1992) classification scheme was that the more difficult the distractors, the more likely it would be that the maze would assess *deep* comprehension—that is, comprehension that extended beyond the sentence level. They further reasoned that the use of more difficult distractors would result in a maze that better reflected cognitive conceptions of reading and thus would produce scores with greater construct validity. Ketterlin-Geller, McCoy, Twyman, and Tindal (2006) echoed this logic and argued that to adequately measure passage-level reading comprehension, distractors had to be "grammatically correct, syntactically possible, and reasonably and meaningfully related to the

context of the passage" (p. 42).

Although the logic used by Parker et al. (1992) and Ketterlin-Geller et al. (2006) is reasonable, it is possible to construct other arguments that might lead to different conclusions about the extent to which the maze reflects cognitive conceptualizations of reading. For example, it is possible to argue that difficult distractors get in the way of reading comprehension because they interfere with the process of constructing a coherent mental representation of the text. Underlying this argument are results of a study conducted by van den Broek, Risden, Tzeng, Trabasso, and Basch (2001), who examined the effects of inferential questioning during reading on comprehension. Results revealed that questioning *during* reading interfered with the reading processes of younger and poorer readers because the questions placed additional demands on the readers' working memory capacity. These additional demands interfered with the process of constructing a coherent mental representation of text. If one were to apply this logic to the maze task, one might argue that a maze with more difficult distractors might lead to a less valid maze score—especially for poorer readers—because difficult distractors require the reader to search for and think about the correct answer. This search-and-think process might tax the reader's working memory capacity and thus interfere with the construction of a coherent mental representation of the text.

These two competing explanations illustrate the need to examine multiple sources of evidence for construct validity. Messick (1989a, 1989b) depicted validity as a unitary concept, with construct validity as the whole of validity. Within this framework, multiple sources of evidence are examined to determine the extent to which scores from a measure support construct validity (see Espin & Deno, 2016, for a discussion of this point as it relates to CBM research). The arguments presented previously relate to one source of evidence—the extent to which scores from a measure fit with theoretical conceptualizations of the construct being measured. It is important to examine other sources of evidence as well, such as criterion-related and predictive sources of evidence. In the current study, we examine criterion-related and predictive evidence for the validity of maze-selection scores and examine the differences in these sources of evidence as they relate to the different approaches used to select maze distractor items. We focus primarily on the use of maze scores as *general indicators of broad reading performance,* which is how the scores are used within a CBM system. However, in the discussion, we reflect on the extent to which the maze represents various aspects of reading

performance, including reading comprehension.

Differences in timing and scoring

It is not only the methods of maze construction that have differed across secondary school studies but also methods of administration and scoring. Generally speaking, administration times for the maze have varied from 2 min to 10 min, and scores have included both correct and correct minus incorrect, with various scoring rules used to account for guessing (see Pierce, McMaster, & Deno, 2010; Wayman et al., 2007). A small number of studies have directly compared differences in administration time and scoring procedures on the technical adequacy of maze scores. With regard to time, Espin et al. (2010) and Tichá et al. (2009) compared the reliability, validity, and sensitivity to growth for scores produced from 2-, 3-, and 4-min mazes and found few differences related to time frame, with the exception that alternate-forms reliability tended to increase with administration time. Espin et al. (2010), Pierce et al. (2010), and Wayman et al. (2009) compared reliability, validity, and/or sensitivity to growth for scores produced with various scoring procedures and found few differences related to the use of different scoring procedures.

Although results from the studies to date suggest that timing and scoring procedures will have little effect on the technical adequacy of maze scores, we include them in the current study because the two factors might interact with the method used to select distractors. For example, if distractors are made to be more difficult, then it might be necessary to provide a longer time frame to obtain reliable scores from students. Likewise, with more difficult distractors, scoring correct-minus-incorrect answers might lead to more reliable and valid scores than scoring correct answers only. Thus, in the current study, we examine not only the effects of various methods for distractor selection but also the effects of time frame and scoring procedures.

Research questions

In sum, in the current study, we examine two approaches to distractor selection for the maze task. Based on the classification system created by Parker et al. (1992), we create a more and a less difficult form of the maze measure. In addition, we examine the effects of time and scoring procedures, specifically, differences between 1-, 2-, and 3-min time frames and between scoring the number of correct versus the number of correct-minus-incorrect answers.

Our general research question is this: Are there differences in the technical adequacy of CBM maze-selection scores based on the methods used to construct, administer, and score maze passages? Our specific research questions are as follows:

1. Are there differences in the alternate-forms reliability of maze-selection scores related to distractor selection, time, and scoring procedures?
2. Are there differences in the validity of maze-selection scores related to distractor selection and scoring procedures?
3. Are there differences in the accuracy of the prediction of scores on a state standards test related to distractor selection and scoring procedures?

Methods

Participants

The study was conducted in eight eighth-grade communication arts classrooms in a suburban school district in the Midwest. The district had a student enrollment of 17,882. A total of 14% of students received special education, and 57% were eligible for free or reduced lunch. Within the district, 72% of the students were Black, 24% were White, and 4% were "other."

The study sample consisted of 262 students (145 of whom were male). Of the students, 84% were African American, 12% White, 2% Hispanic, and 3% Asian. Finally, 11% of students qualified for special education services, and 48% of students qualified for free or reduced lunch.

Measures

The measures administered included the researcher-created maze passages and three criterion measures. The criterion measures, which were already being administered by the school in response to district or state mandates, included the Scholastic Reading Inventory (SRI), AIMSweb maze passages, and the Missouri Assessment Program (MAP) in Communication Arts, which is the Missouri state high-stakes assessment. Although the measures were chosen because of convenience, they were also strongly aligned with the research questions and would have been chosen by us to administer if they were not already being given by the school.

Researcher-created maze passages

Maze passages of approximately 900 words in length were created from read-aloud passages that had been written for the Iowa Department of Education to use to monitor secondary school students (Flansberg, 2012). Passages were written in the form of newspaper articles and covered topics that were thought to be of general interest to adolescents and that did not require extensive background knowledge to understand.

A total of 29 reading-aloud potential passages were available for use. All passages were examined for readability to ensure an appropriate reading level for the students and to aid in passage selection for the study. Ten different readability estimate indices (FORCAST, Spache, Dale-Chall, Flesch- Kincaid Grade Level, Coleman-Liau, Automated Readability Index, Flesch Reading Ease, Fog Index, Lix Formula, SMOG-Grading).were used to determine the text complexity of the passages. Each pass- age was then classified as typical, intermediate, or difficult based on each index score. Passages with seven or more indices falling within the easy or difficult range were excluded, and thus 10 of the 29 passages were eliminated.

The Flesch-Kincaid Grade Level Index (Flesch, 1948) was then calculated for each passage using Microsoft Word. The Flesch-Kincaid measures word and sentence length to determine the complexity of a reading passage at a specific grade level (Flesch, 1948). Only those passages that obtained a seventh-grade level according to Flesch-Kincaid were considered for probe development, which resulted in a set of seven passages.

The seven remaining passages were then examined for content. Passages whose content appeared too technical or required specific background knowledge (i.e., related directly to Iowa) were excluded. As a result, two passages emerged as appropriate in content for probes. The passage length for both texts was limited to 900 words. The two reading-aloud passages were converted into maze probes using the Maze Passage Generator (www.interventioncentral.org). The first sentence in the passage was left intact, and a multiple-choice response item was created at every seventh word. The response item included the correct word and two distractor items. The procedures used to select the distractor items are described next.

Typical maze probes. Distractors for the typical maze probes were selected via the Maze Passage Gen- erator. The Maze Passage Generator offers three different options for selecting distractors: select from a list of common English words, select from other words in the

passage, or a select from a word list provided by the creator. For the purposes of this study, distractors were selected from the list of com- mon English words. After initial maze passages were generated, the distractors were checked to ensure that they met the following criteria:

1. The distractors did not begin with the same letter as the correct answer.
2. The distractors were within one letter of the correct answer in length.
3. The distractors were a different part of speech than the correct answer.

Because some distractor items created by the Maze Passage Generator on Intervention Central violated some of the aforementioned criteria, each passage was reviewed manually and random selections of distractor words that met criteria were obtained from a word list generator (listofrandomwords. com).

In the typical maze probes, the correct choice was clearly distinguishable from the distractors. That is, the distractors were selected so that if a student was fluently reading and comprehending the pass- age, the correct choice would be immediately obvious and typical to select. Once distractors were selected, probes were formatted so that all three word choices were on one line of text. For an example of the typical probe, please see Figure 1.

Novel maze probes. Distractors for the novel maze probes were developed based on methods described by Parker et al. (1992). For each multiple-choice item, one distractor was content related and the other was not content related. Both distractors were the same part of speech (e.g., noun, verb, adjective) as the correct word. Neither distractor preserved the meaning or made sense in the sentence. The two distractors differed in terms of how they related to the story. Content-related distractors were words that appeared in the story, whereas non-content-related distractors were words that did not appear in the story. To illustrate

Typical

Ex. Shopping for a car may seem like an easy task. But a car expert says that **(in, be, she)** order to ensure you're pleased with **(need, your, water)** purchase, you need to do homework **(before, modern, reading)** you part with your hard-earned money.

Novel

Ex. Shopping for a car may seem like an easy task. But a car expert says that **(for, past, in)** order to ensure you're pleased with **(most, few, your)** purchase, you need to do homework **(before, after, form)** you part with your hard-earned money.

Figure 1. Typical and novel maze probe examples.

the different distractors, we provide this sentence taken from a novel probe titled *Car Shopping:* "Mike is not a fan of (**growing, marching, trading**) cars on a regular basis." *Trading* is the correct choice; *growing* is a content-related distractor because it appears earlier in the story; and *marching* is a non-content-related distractor because it does not appear in the story, nor is it related to the content.

In novel maze probes, the correct choice is not as clearly distinguishable from the distractors as in typical maze probes. That is, the selection of the correct word in a novel probe requires semantic understanding at the sentence level. The assumption is that content-related distractors create more challenging maze-selection tasks because of the additional content similarity (Parker et al., 1992).

1. Distractors for the novel probes met the following criteria:
2. The distractors did not begin with the same letter as the correct answer.
3. The distractors were within two letters of the correct answer in length.
4. The distractors were the same part of speech, same tense, and same plurality but not meaningful.
5. One distractor was related to the content and one was unrelated to the content.

Again, distractor words were selected from within the story or from the word list generator (http:// www.wordlistgenerator.net/). For non-content-related distractors the passage was searched to ensure that the word was not found anywhere in the story. Once distractors were selected, probes were for- matted so that all three word choices were on one line of text. For an example of the novel probe, please see Figure 1.

Criterion measures

Three different reading criterion measures were used to examine the criterion validity of scores from the typical and novel maze probes.

SRI. The SRI is a computer-adaptive reading assessment developed by Scholastic (2007). The SRI is used to measure reading comprehension of literary and expository texts of varying degrees of difficulty for kindergarten through Grade 12. The SRI focuses on identifying details in a passage, identifying cause-and-effect relations and the sequence of events, drawing conclusions, and making comparisons and generalizations (Scholastic, 2007). The SRI generates criterion- and norm- referenced scores for each student,

including a percentile rank, stanine, normal curve equivalent, grade-level standard, performance standard, and Lexile® score. For the purposes of this study, the percentile rank was used, as it was the only score available. A student's reading level is represented by a Lexile score, ranging from 100 L for beginning readers to 1500 L for advanced readers (Scholastic, 2007). The Lexile score is determined by the difficulty of the items the student answers correctly and incorrectly. The test–retest reliability (r = .89) and the criterion validity (r = .70–.83) are adequate (Scholastic, 2007).

Aimsweb CBM reading R-Maze. AIMSweb provides assessment tools for schools to assist with screening, progress monitoring, and data-based decision making. The AIMSweb R-Maze task uses passages between 150 and 400 words as indicators of overall reading proficiency through a silent reading, context-based task. The first sentence is left intact, and then every seventh word is replaced with three choices in parentheses. The three word choices consist of the correct answer, a distractor selected randomly that is unrelated to the passage, and a distractor selected randomly from the passage (Shinn & Shinn, 2002). Students have 3 min to make as many replacements as they can. In the maze task every seventh word is replaced with a choice of three words (the correct word and two distractors). Students are directed to read the passage and select, by circling, the missing words from the choices. Students' scores on the maze are calculated by totaling the number of words circled correctly during the 3-min administration. Maze instruments have been found to be valid and reliable measures of students' reading skills. The maze has strong criterion-related validity with Curriculum- Based Measures in Reading (CBM-R), with coefficients ranging from .77 to .86 (Espin, Deno, Maruyama, & Cohen, 1989). The concurrent validity of the maze has been established with other group-administered tests of reading achievement (Jenkins & Jewell, 1993). The AIMSweb R-Maze was administered by the school district in the spring, 2 weeks prior to the administration of the maze tasks for this study. The median score on three AIMSweb passages for each student was used as a criterion measure. Although this is a similar measure to the researcher-created maze, the distractors for the maze passages in the current study were selected using a different methodology, and given that AIMSweb is a standard measure given to thousands of students across the country, it was important to see how our measures performed against it.

MAP: Communication arts. The MAP is a standardized test that assesses

Missouri students' knowledge, skill, and competencies in Grades 3–8 in the areas of communication arts, mathematics, and science (Grades 5 and 8). The Missouri Department of Elementary and Secondary Education (2012) uses scores from this assessment to monitor the educational progress of Missouri students toward meeting state standards and to identify students falling below academic proficiency in certain areas. It assesses a range of language arts skills, including reading comprehension, reading and evaluating fiction and nonfiction text, and language skills such as vocabulary and grammar. The MAP is the group-administered high-stakes assessment used in all school districts in the state of Missouri. Two types of scores are reported for the MAP: a scale score and its associated level of achievement (below basic, basic, proficient, and advanced). An independent contractor evaluated the technical adequacy of the MAP test. Data published in the technical report of the 2012 version of the MAP test reported a Cronbach's alpha of .91 for the third-grade communication arts test. Moreover, the third-grade communication arts and mathematics tests were correlated at .69. For this study, standard scores were used as the criterion (Missouri Department of Elementary and Secondary Education, 2012).

Procedure

Administration

Students were randomly assigned to a probe condition (typical vs. novel; $n = 131$ per condition), and each student received a packet containing a cover page with a sample item and two probes. Within each condition, the order in which the two probes were administered was counterbalanced. Thus, four different packets were randomly distributed to participants: typical (Probe 1 then 2), typical (Probe 2 then 1), novel (Probe 1 then 2), and novel (Probe 2 then 1).

All measures were group administered to students during their communication arts class. The first and second authors completed the administration of probes. Prior to completing the maze tasks, students completed a practice maze item. Directions and sample items were adapted from Shinn and Shinn (2002). Students were instructed that they would have 3 min to silently read the passage. At 1 and 2 min students were asked to put a slash through the word they were reading at that moment. This was done so that the effects of probe duration on reliability and validity could be examined.

Scoring

The research team scored measures using a scoring template that was created by cutting out the correct word from a blank probe and laying the scoring template over the student copy. Probes were scored for each minute mark (1, 2, 3 min). However, it should be noted that students in one entire class did not mark their place at the 1- and 2-min marks because of administrator error in providing instructions. Also, there were a few more students who randomly did not mark minutes, but after further analysis we found that this group of students was less than 9% of the total population and their scores were not significantly different than those of students who did make minute marks. Because of these differences in marking the passages, results and tables reflect various sample sizes for 1 min and 2 min. Last, probes were scored in two ways: number of correct and number of correct-minus-incorrect choices. Correct choices are easier to score; however, one might assume that using correct minus incorrect adds a control for guessing and might result in more valid scores.

Data analysis

To compare performance across the two types of maze (typical and novel), we conducted a series of *t* tests. To correct for the inflated probability of Type I error due to the number of tests conducted, we used a Bonferroni adjusted *p* value of .005 for interpretation; however, specific *p* values are reported.

To calculate alternate-forms reliability, we examined correlations between the two maze forms within conditions. Correlation coefficients were transformed using a Fisher's *z* transformation and the significance of differences in correlations tested using a *z* test for independent samples (Rosenthal & Rosnow, 1991). To calculate validity, we calculated the mean of the two maze probes, and correlations between this mean score and scores on each of the criterion measures (AIMSweb, MAP, SRI) were calculated. Differences were again tested using a *z* test for independent samples.

Receiver-operating characteristic (ROC) analysis was used to examine how well each of the maze passages served as a predictor of student performance on each of the criterion measures. The ROC analysis is important if measures are to be used as screening measures. Maze passage types were then compared on area under the curve (AUC). AUC is a metric

ranging from .5 to 1 that provides the probability of a predictor correctly classifying a pair of students from two different categories (e.g., proficient, nonproficient) and can be used as an effect size statistic (Swets, 1988). AUC values greater than .70 are considered adequate, those between .80 and .90 good, and those above .90 excellent (Swets, 1988).

Results

Descriptive analysis

Means and standard deviations of student performance on the typical and novel maze passages are reported in Table 1. Performance on the three criterion measures was remarkably similar for students in the typical and novel groups, confirming the equivalence of the groups formed via the random assignment procedures. Examination of performance on the maze formats, however, revealed differences in mean scores. In general, scores were somewhat higher and standard deviations some- what larger for the typical than the novel maze. However, differences reached a significance level of $p < .05$ for only two comparisons—2-min and 3-min correct minus incorrect ($d = .29$)—although these were not significant when we used the values adjusted to account for family-wise error. All measures and metrics were normally distributed with the exception of the MAP, which appeared to have a somewhat restricted range of scores as evidenced by kurtosis values greater than 2.

When we examined mean scores across each of the 3 min, it appeared that in the typical sample students answered at an approximately equal rate for the 3 min, with about 8.5 items per minute in both correct and correct-minus-incorrect scoring. When we scored correct choices for the novel sample, the rate of answers was similar to the typical; however, when correct-minus-incorrect scoring procedures were applied, students answered at an approximately equal but lower rate (7.6, 7.5, and 7.3, respectively).

When we reviewed the number of errors in the two conditions (differences between the means of number correct and correct minus incorrect), it appeared that in the typical sample correct- minus-incorrect scores were very similar to the number correct, which implies few errors. In contrast, the number of errors appeared slightly larger in the novel condition. *T* tests revealed that at each minute, correct scores were slightly higher ($p < .05$) than correct-minus-incorrect scores (1 min, t 2.05, p .04; 2 min, t 2.08, p .04; 3 min, t 2.50, p .01), although not this was statistically significant when we used the adjusted

criterion (.005). The direction of this trend may be expected because the novel probe is intended to be a more rigorous measure to increase variance in scores.

Table 1. Descriptive statistics for the means of two maze probes by minute and criterion measure.

	Typical					Novel						
Minute	***n***	***M***	***SD***	**Skewness**	**Kurtosis**	***n***	***M***	***SD***	**Skewness**	**Kurtosis**	***t***	***p***
Minute 1												
Correct	106	8.71	3.15	0.39	-0.17	110	8.43	2.87	0.31	0.13	0.684	.495
Correct - Incorrect	106	8.20	3.47	0.09	0.50	110	7.60	3.13	0.25	0.10	1.332	.184
Minute 2												
Correct	121	17.88	6.27	0.14	-0.38	123	16.70	5.42	0.23	-0.41	1.563	.119
Correct- Incorrect	121	17.06	6.83	-0.12	0.29	123	15.19	5.95	0.18	-0.43	2.277	.024
Minute 3												
Correct	131	26.69	9.96	-0.04	-0.23	131	24.77	7.98	0.07	0.09	1.816	0.71
Correct - Incorrect	131	24.90	10.00	-0.20	0.25	131	22.12	9.13	-0.26	0.61	2.332	.020
Measure												
AIMSweb	53	22.21	7.62	0.58	0.05	54	22.46	7.30	0.75	0.19	-0.177	.860
MAP	126	686.09	34.41	-1.52	4.44	127	682.12	36.43	-1.76	5.53	0.891	.374
SRI	114	51.94	32.68	-0.17	-1.43	112	50.99	31.66	0.02	-1.34	0.221	.825

Note. MAP = Missouri Assessment Program; SRI = Scholastic Reading Inventory.

Table 2. Alternate-forms reliability and confidence intervals for the maze passages.

	Traditional probes			**Subtype probes**		
Maze choice	**Min 0-1**	**Min 0-2**	**Min 0-3**	**Min 0-1**	**Min 0-2**	**Min 0-3**
Correct	.64 [.44, .75]	.71 [.56, .82]	.81 [.73, .95]	.69 [.60, .95]	.71 [.58, .84]	.80 [.66, .86]
Correct - Incorrect	.50 [.44, .72]	.67 [.57, .83]	.76 [.67, .90]	.53 [.38, .67]	.61 [.45, .71]	.79 [.65, .85]

Note. n= 131.

All correlations are significant at $p < .001$.

Research question 1: Alternate-forms reliability

Alternate-forms reliability coefficients increased with probe duration for both correct and correct- minus-incorrect scoring, with 3 min yielding the largest coefficients, which included the entire sample (see Table 2). Because of inconsistent student marking at 1 min and 2 min during administration, only the 3-min correlations are discussed in this section. Alternate-forms reliability using correct restorations as the metric was similar for the typical (r [131] = .81) and novel (r [131] = .80) passages (z = .23, p = .820). For correct-minus-incorrect restorations, reliability was also similar for typical (r [131] = .76) and novel (r [131] = .79) passages (z = .60, p = .547). Coefficients were similar across both metrics and at the low end to be considered sufficient for making individual decisions (Salvia, Ysseldyke, & Bolt, 2012).

Research question 2: Criterion-related validity

Criterion-related validity (see Table 3) was calculated for 3-min probes only because they produced the highest reliability coefficients. Coefficients were similar across typical (r = .61–.83) and novel (r = .63–.82) maze passages. All correlations were statistically significant. The largest validity coefficients were found between maze scores and AIMSweb scores, a not unexpected finding given that both are maze tasks. Correlations with the MAP and SRI were moderate to strong, ranging from r = .55 to .69. No statistically significant differences in correlations were found between the two maze formats regardless of criterion measure (AIMSweb, MAP, SRI) or scoring metric (correct restorations, correct-minus-incorrect restorations).

Table 3. Criterion-related validity for traditional and subtype 3-min maze passages with criterion measures.

Measure	Metric	Traditional *r [CI]*	Subtype *r [CI]*	z	p
AIMSweb (n=53/54)	Correct	.83 [.65, .96]	.82 [.75, .99]	0.16	.874
	Correct - incorrect	.81 [.64, .97]	.78 [.68, .99]	0.54	.590
MAP (n=126/127	Correct	.70 [.52, .76]	.63 [.54, .85]	0.84	.403
	Correct – incorrect	.73 [.57, .79]	.65 [.56, .85]	1.21	.228
SRI (n=114/112)	Correct	.55 [.41, .74]	.64 [.57, .91]	1.04	.300
	Correct - incorrect	.61 [.51, .84]	.66 [.59, .92]	0.62	.534

Note. CI = 95% confidence interval; MAP = Missouri Assessment Program; SRI = Scholastic Reading Inventory.

Table 4. Receiver-operating characteristic analysis results (AUC) for typical and novel maze passages.

		AUC		
Criterion measure	Metric	Typical	Novel	*p*
MAP—Prof	Correct	.834	.807	.576
	Correct – incorrect	.835	.817	.704
MAP—Adv	Correct	.892	.757	.005
	Correct – incorrect	.906	.791	.012
SRI—40th	Correct	.798	.826	.589
	Correct – incorrect	.833	.845	.803
SRI—50th	Correct	.792	.829	.478
	Correct – incorrect	.825	.836	.826
SRI—75th	Correct	.777	.827	.347
	Correct – incorrect	.805	.837	.529
Maze 50th	Correct	.910	.860	.418
	Correct – incorrect	.915	.858	.358
Maze 75th	Correct	1.000	.951	.107
	Correct – incorrect	1.000	.958	.136

Note. AUC = area under the curve; MAP = Missouri Assessment Program; Prof = proficient score on the MAP; Adv = advanced score on the MAP; SRI = Scholastic Reading Inventory; Maze = AIMSweb maze probes.

Research question 3: ROC analysis

AUCs for typical and novel maze passages were good to very good for all proficiency levels for the MAP and SRI (range = .757–.906) across both metrics (see Table 4). All AUCs exceed the minimum criterion as adopted by the Center for Response to Intervention (n.d.), yet only the typical type predicting advanced status on the MAP and the AIMSweb maze exceeded the optimal criterion of .85. AUCs for prediction to the AIMSweb maze were all high, which is not surprising given that both are maze tasks. AUCs were similar for prediction to all criterion measures at each level of proficiency using each metric except for MAP at the advanced level of performance. For this outcome, both metrics (correct, correct minus incorrect) were stronger when the typical maze passages rather than the novel maze passages were used.

Discussion

In the present study we examined the technical adequacy of scores from maze passages created using two different sets of rules for selecting distractors: one meant to create more difficult distractors and the other easier distractors. First, a comparison of means and standard deviations revealed no significant differences between typical and novel probes when we adjusted for the number of comparisons. Second, regarding alternate-forms reliability, results revealed that coefficients were similar across probe types. However, as with earlier research (Espin et al.,

2010; Tichá et al., 2009), for both types of probes, reliability increased with probe duration, with the strongest coefficients found for 3 min. In addition, for both types of probes, reliability coefficients were consistently larger for correct than for correct-minus-incorrect scores. Thus, similar to earlier studies, depending on the decision to be made, our results suggest the need to use a 3-min time frame.

To reduce the overall number of analysis, we examined only scores from 3-min maze probes in the validity of analyses. Results revealed that scores on both typical and novel maze probes were positively and moderately strongly correlated with all three criterion measures, with correlations ranging from $r = .55$ to .73 for the two general reading proficiency measures (MAP and SRI) and from .77 to .83 for the other maze measures (AIMSweb). There were no statistically significant differences related to probe type, suggesting that probe types reflected general reading skills equally well. However, it is worthwhile to note that although differences were not statistically significant, correlations between the maze and MAP scores were slightly higher for the typical than for the novel passages, whereas correlations between the maze and SRI scores were slightly higher for the novel than for the typical pas- sages. One explanation for this obtained result might relate to the fact that that SRI assesses reading comprehension only, whereas the MAP assesses a range of language arts skills, including reading com- prehension but also reading and evaluating fiction and nonfiction text and language skills such as vocabulary and grammar. The pattern of results suggests that scores from typical maze probes might better reflect broad reading outcomes (which is the type of score desired for CBM), whereas scores from novel maze probes might better reflect reading comprehension skills. However, keeping in mind the fact that differences were not significant, it would be important to directly test this assumption in a future study to see whether the pattern replicates across studies. Findings from such a study would contribute to the question raised at the beginning of the article regarding the extent to which the maze measure reflects the theoretical underpinning of reading in general and reading comprehension in particular.

The fact that correlations were strongest between the maze scores and the AIMSweb scores is not surprising given the similarity between the two measures. Correlations with the other criterion measures were reasonably strong and within the range of correlations found in other research (Tichá et al., 2009; Tolar et al., 2012; Torgesen et al., 2005).

With regard to scoring procedures, similar to the findings of Pierce et al. (2010),

there were no statistically significant differences between correct versus correct-minus-incorrect scores, although validity coefficients for correct-minus-incorrect scores tended to be somewhat higher than for correct scores only for the MAP and SRI. Our results suggest, however, that for screening purposes, one could use either correct only or correct minus incorrect. If mazes need to be scored by hand, correct only is much simpler than correct minus incorrect. It will be important in future research to examine whether our results also hold true when CBM measures are used as ongoing progress measures. One scoring system may be more sensitive to growth than the other. In addition, it would be important to examine whether the effects of scoring system differ for subgroups of students, such as struggling readers or students who are English language learners. It may be that these students make more errors than students in general, and the errors may contribute to better discrimination within these groups of students.

With regard to the use of scores from two types of maze-selection passages to predict performance on a state standardized assessment, the ROC analyses speak to the extent to which the typical and novel measures accurately predict student performance on each of the criterion measures. It appears that compared to the novel measures, the typical probes demonstrated a stronger prediction; however, because of the slightly skewed and kurtotic MAP results (MAP—advanced had 8% and 6% base rates), the AUC may have been highly influenced by one or two students. However, overall AUC coefficients were high, which provides support that scores from both types of maze measures appear to be a suitable indicator of proficiency or nonproficiency on the remaining criterion measures.

Limitations and future directions

There are some limitations to the present study that may have affected the findings. First, the measures were given to a single sample from one school. A larger sample size from different schools and/ or districts would increase the generalizability of the findings. Second, the content of the passages may not have been so interesting to adolescents. It is also possible that students in the study sample may not have had extensive background knowledge related to the content of the passages. In the future, it may be beneficial to work more closely with school personnel to develop passages related to the curriculum or current events to ensure that students have the background knowledge to assist in their reading and comprehension.

Third, the use of the measures at a single point in time might have resulted in narrow results based on the conditions on that day. In the future, averaging performance on a passage each day for three different days would help control for this potential effect. Further investigation of novel probes with a reading comprehension criterion measure, equating maze passages for difficulty of items, and examining reliability and validity for maze probes at the secondary level, is warranted.

Conclusions

In conclusion, our results suggest that for CBM screening purposes, educators can use a typical approach to maze construction, that mazes should be administered for 3 min, and that mazes can be scored for either number correct or correct minus incorrect. However, as mentioned earlier, it will be important in future research to examine whether these suggestions hold true for various subgroups of students and when the maze is used as an ongoing progress measure. With the availability of technologies such as the Maze Passage Generator (www.interventioncentral.org), educators and practitioners have access to an efficient and effective means of probe development for typical maze probes at the secondary school level. With regard to scoring procedures, scoring the number correct is more efficient than scoring correct minus incorrect; however, if probes are scored via an electronic progress monitoring program, differences in efficiency disappear.

References

Alley, G. R., & Deshler, D. D. (1979). *Teaching the learning disabled adolescent: Strategies and methods*. Denver, CO: Love.

Busch, T., & Espin, C. A. (2003). Using curriculum-based measurement to prevent failure and assess learning in content areas. *Assessment for Effective Intervention*, *28*, 49–58. doi:10.1177/073724770302800306

Castillo, J. M., & Batsche, G. M. (2012). Scaling up response to intervention: The influence of policy and research and the role of program evaluation. *Communiqué*, *40*(8), 14–16.

Center for Response to Intervention at American Institutes for Research. (n.d.) *Screening Tools Chart Rating System*. Retrieved from http://www.rti4success.org/resources/tools-

charts/screening-tools-chart/screening-tools-chart-rating- system

Christ, T. J., Zopluoglu, C., Monaghen, B. D., & Van Norman, E. R. (2013). Curriculum-based measurement of oral reading: Multi-study evaluation of schedule, duration, and dataset quality on progress monitoring outcomes. *Journal of School Psychology*, *51*, 19–57. doi:10.1016/j.jsp.2012.11.001

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

Deno, S. L., Maruyama, G., Espin, C. A., & Cohen, C. (1989). *The Basic Academic Skills Samples (BASS)*. Retrieved from http://www.progressmonitoring.org/RIPMProducts2.html#bass

Espin, C. A., & Deno, S. L. (1993). Performance in reading from content-area text as an indicator of achievement. *Remedial and Special Education*, *14*(6), 47–59. doi:10.1177/074193259301400610

Espin, C. A., & Deno, S. L. (2016). Oral reading fluency or reading aloud from text: An analysis through a unified view of construct validity. In K. D. Cummings & Y. Petscher (Eds.), *The fluency construct: Curriculum-Based Measurement concepts and applications*. New York, NY: Springer.

Espin, C. A., Deno, S. L., Maruyama, G., & Cohen, C. (1989, March). *The Basic Academic Skills Samples (BASS): An instrument for the screening and identification of children at risk for failure in regular education classrooms*. Paper presented at the National Convention of the American Educational Research Association, San Francisco, CA.

Espin, C. A., & Foegen, A. (1996). Validity of three general outcome measures for predicting secondary students' performance on content-area tasks. *Exceptional Children*, *62*, 497–514. doi:10.1177/001440299606200602

Espin, C., Wallace, T., Lembke, E., Campbell, H., & Long, J. D. (2010). Creating a progress-monitoring system in reading for middle-school students: Tracking progress toward meeting high-stakes standards. *Learning Disability Research & Practice*, *25*(2), 60–75. doi:10.1111/j.1540–5826.2010.00304.x

Flansberg. (2012). *Maze reading passages*. Des Moines, IA.

Flesch, R. F. (1948). A new readability yardstick. *Journal of Applied Psychology*, *32*(3),

221–233. doi:10.1037/h0057532 Fuchs, L. S., & Fuchs, D. (2002). Computer applications to curriculum-based measurement. *Special Services in the Schools*, *17*(1–2), 1–14. doi:10.1300/j008v17n01_01

Fuchs, L. S., Fuchs, D., & Compton, D. L. (2010). Rethinking response to intervention at middle and high school. *School Psychology Review*, *39*(1), 22–28.

Grosche, M., & Volpe, R. J. (2013). Response-to-intervention (RTI) as a model to facilitate inclusion for students with learning and behaviour problems. *European Journal of Special Needs Education*, *28*, 254–269. doi:10.1080/ 08856257.2013.768452

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

Jenkins, J. R., & Jewell, M. (1993). Examining the validity of two measures for formative teaching: Reading aloud and maze. *Exceptional Children*, *59*, 421–432.

Johnson, E., Semmelroth, C., Allison, J., & Fritsch, T. (2013). The technical properties of science content maze passages for middle school students. *Assessment for Effective Intervention*, *38*(4), 214–223. doi:10.1177/1534508413489337

Ketterlin-Geller, L. R., McCoy, J. D., Twyman, T., & Tindal, G. (2006). Using a concept maze to assess student understanding of secondary-level content. *Assessment for Effective Intervention*, *31*(2), 39–50. doi:10.1177/ 073724770603100204

McKenna, M. C., & Miller, J. W. (1980). The effects of age and distractor type on maze performance. In M. L. Kamil (Ed.), *Perspectives on reading research and instruction: 29th yearbook of the National Reading Conference* (pp. 288– 292). Washington, DC: National Reading Conference.

Messick, S. (1989a). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher*, *18*, 5–11.

Messick, S. (1989b). Validity. In R. L. Linn (Ed.), *Educational measurement (*3rd ed., pp. 13–103). New York, NY: Macmillan.

Missouri Department of Elementary & Secondary Education. (2012). *Missouri Assessment Program grade-level assessment: Technical report*. Retrieved from http://dese.mo.gov/sites/default/files/asmt-gl-2012-tech-report.pdf

National Center for Education Statistics. (2014). *NAEP data explorer*. Retrieved from http://nces.ed.gov/nationsreportcard/ naepdata/

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. doi:10.1177/002246699202600205

Pierce, R. L., McMaster, K. L., & Deno, S. L. (2010). The effects of using different procedures to score maze measures. *Learning Disabilities Research & Practice*, *25*(3), 151–160. doi:10.1111/j.1540–5826.2010.00313.x

Prewett, S., Mellard, D., Deshler, D., Allen, J., Alexander, R., & Stern, A. (2012). Response to intervention in middle schools: Practices and outcomes. *Learning Disabilities Research & Practice*, *27*(3), 136–147. doi:10.1111/j.1540– 5826.2012.00359.x

Reschly, A. L., Busch, T. W., Betts, J., Deno, S. L., & Long, J. D. (2009). Curriculum-based measurement oral reading as an indicator of reading achievement: A meta-analysis of the correlational evidence. *Journal of School Psychology*, *47*, 427–469.

Rosenthal, R., & Rosnow, R. L. (1991). *Essentials of behavioral research: Methods and data analysis (*2nd ed.). New York, NY: McGraw-Hill.

Salvia, J., Ysseldyke, J., & Bolt, S. (2012). *Assessment: In special and inclusive education*. New York, NY: Cengage. Scholastic. (2007). *Scholastic Reading Inventory technical guide*. New York, NY: Scholastic Reading.

Shin, J., Deno, S. L., & Espin, C. (2000). Technical adequacy of the maze task for curriculum-based measurement of reading growth. *Journal of Special Education*, *34*(3), 164–172. doi:10.1177/002246690003400305

Shinn, M. R., & Shinn, M. M. (2002). *Administration and scoring of curriculum-based measurement maze for use in general outcome measurement*. Eden Prairie, MN: Edformation.

Silberglitt, B., Burns, M. K., Madyun, N., & Lail, K. E. (2006). Relationship of reading fluency assessment data with state accountability test scores: A longitudinal comparison of grade levels. *Psychology in Schools*, *43*, 527–535. doi:10.1002/ pits.20175

Swets, J. (1988, June 3). Measuring the accuracy of diagnostic systems. *Science*, *240*(4857), 1285–1293.

Tichá, R., Espin, C. A., & Wayman, M. M. (2009). Reading progress monitoring for

secondary-school students: Reliability, validity, and sensitivity to growth of reading-aloud and maze-selection measures. *Learning Disabilities Research & Practice*, *24*(3), 132–142. doi:10.1111/j.1540–5826.2009.00287.x

Tolar, T., Barth, A., Francis, D., Fletcher, J., Stuebing, K., & Vaughn, S. (2012). Psychometric properties of maze tasks in middle school students. *Assessment for Effective Intervention*, *37*(3), 131–146. doi:10.1177/1534508411413913

Torgesen, J. K., Nettle, S., Howard, P., & Winterbottom, R. (2005). *Brief report of a study to investigate the relationship between several brief measures of reading fluency and performance on the Florida Comprehensive Assessment Test— Reading in 4th, 6th, 8th, and 10th grades* (FCRR Tech. Rep. No. 6). Retrieved from http://www.fcrr.org/TechnicalRe- ports/Progress_monitoring_report.pdf

van den Broek, P., Risden, K., Tzeng, Y., Trabasso, T., & Basch, P. (2001). Inferential questioning: Effects on compre- hension of narrative texts as a function of grade and timing. *Journal of Educational Psychology*, *93*, 521–529. doi:10.1037//0022–0663.93.3.521

Wayman, M., Tichá, R., Espin, C. A., Wallace, T., Ives-Wiley, H., Du, X., & Long, J. (2009). *Comparison of different scoring procedures for the CBM maze selection measures* (Tech. Rep. No. 10). Retrieved from the Research Institute on Progress Monitoring website: http://progressmonitoring.org/pdf/TR10scringmath.pdf

Wayman, M., Wallace, T., Wiley, H. I., Tichá, R., & Espin, C. A. (2007). Literature synthesis on curriculum-based measurement in reading. *Journal of Special Education*, *41*, 85–120. doi:10.1177/00224669070410020 40