Curriculum-based measurement of oral reading: Passage equivalence and probe-set development

Theodore J. Christ,⁎ Scott P. Ardoin

University of Minnesota, United States
University of South Carolina, United States

Received 30 May 2007; received in revised form 30 July 2008; accepted 30 September 2008

Abstract

Curriculum-based measurement of reading (CBM-R) is used to estimate oral reading fluency. Unlike many traditional published tests, CBM-R materials are often comprised of 20 to 30 alternate forms/passages. Historically, CBM-R assessment materials were sampled from curricular materials. Recent research has documented the potentially deleterious effects of poorly controlled alternate forms on CBM-R outcomes. The purpose of this study was to examine alternate procedures for the selection of passages that comprise CBM-R passage-sets. The study examined four procedures for the evaluation and selection of passages, including random sampling, Spache readability formula, mean level of performance evaluation, and Euclidean Distance evaluation. The latter two procedures relied on field testing and evaluation of student performance. Each of eighty-eight students in second- and third-grade were administered 50 CBM-R passages. Generalizability and dependability studies were used to examine students’ performance on these passages and evaluate CBM-R passage selection procedures. Results provide support for the use of field testing methods (i.e., calculating performance means and Euclidean Distances) for passage selection. Implications are discussed for future research and practice.

Keywords: Curriculum-based measurement of reading; Curriculum-based measurement; Assessment

⁎ Corresponding author. University of Minnesota, School Psychology, 343 Elliott Hall, 75 E. River Road, Minneapolis, MN 55455, United States.
E-mail address: tchrist@umn.edu (T.J. Christ).
ACTION EDITOR: Tanya Eckert.

0022-4405/$ - see front matter © 2008 Society for the Study of School Psychology. Published by Elsevier Ltd. All rights reserved.
Curriculum-based measurement of reading: passage equivalence and selection

Thirty years after its inception, Curriculum-based measurement (CBM) is often used as a primary data source to guide both low stakes and high stakes decisions within problem solving and response-to-intervention decision-making models (Fuchs, Fuchs, McMaster, & Al Otaiba, 2003; Fuchs, 2003; Speece & Case, 2001; Speece, Case, & Molloy, 2003). The use of CBM for such purposes was not initially foreseen (Shinn, Deno, Fuchs, Germann, & Marston, 2004). Historically, CBM data were used primarily to guide low-stakes decisions (e.g., responsiveness to classroom instruction, pre-referral intervention effectiveness), which are routine, frequent, and easily reversible if an error is discovered. However, with the advent of response-to-intervention, CBM data are likely to guide high-stakes decisions that have potentially irreversible effects, including diagnosis and eligibility determination. Of the various CBM measures (e.g., basic skills math, writing, spelling, maze), CBM oral reading fluency has gained the most attention. CBM-R results are used to evaluate the level and rate of students’ oral reading fluency based on words read correctly per minute (WRC/min) during a CBM-R administration (Shinn, 1989; Shinn & Shinn, 2004).

A CBM-R score is generally derived from the median of three 1-min readings. It is typical to use alternate CBM-R passages for each administration. Student performance fluctuates across administrations; and, therefore, the convention was established to define the median level of performance as the best estimate of a student’s true oral reading fluency at a given point in time (Shinn, 2002). As previously suggested in the literature, the information about the precision of CBM-R level estimates can be gleaned from either the distribution of scores across multiple CBM-R administrations or the standard error of measurement (SEM) (Christ & Coolong-Chaffin, 2007; Christ & Hintze, 2007; Christ & Silberglitt, 2007). Results of published research suggest that the SEM for CBM-R is in the range of 4 to 15 WRC/min; however, SEM is most likely to be within 5 to 9 WRC/min if measurement conditions are controlled to mitigate extraneous influences (e.g., variability due to passage inconsistencies; Christ & Silberglitt, 2007; Poncy, Skinner, & Axtell, 2005). The best estimate and actual value for SEM is context specific. That is, the magnitude of SEM will depend in part on the quality of the administration conditions (e.g., noise, delivery of directions, characteristics of administrator), manner in which the data will be interpreted (e.g., relative or absolute), and the quality of instrumentation (e.g., parallel passages). Those factors have a similar influence on the quality and stability of growth estimates as derived from progress monitoring data (Christ, 2006).

Student performance on CBM-R, quantified in units of WRC/min, can be interpreted as either a relative or absolute score/value. A relative score interpretation is consistent with norm-referenced interpretations. Such interpretations are used to evaluate the relative standing of a score within some norm or referent group. In contrast, when making absolute interpretations we are interested in the exact level of performance (i.e., WRC/min) achieved by the student, not how a student’s performance compares to others. Absolute score interpretation is consistent with criterion-referenced interpretations (e.g., use of benchmarks) when an observed WRC/min is used to evaluate or predict future achievement. Similarly, when interpreting CBM-R progress monitoring data, the data are evaluated to determine whether the absolute change in student performance across the span of intervention suggests that the student made adequate gain in reading achievement. Such
interpretations rely on the absolute value of the measurement outcome in WRC/min (Brennan, 2003; Christ, Johnson-Gros, & Hintze, 2005; Christ & Vining, 2006).

CBM-R research indicates that reliability-like coefficients and magnitudes of SEM tend to be more robust for relative score interpretations as compared to those associated with absolute score interpretations (Poncy, Skinner & Axtell, 2005). Poncy et al. estimated the magnitudes of both reliability-like coefficients and SEMs across a variety of scenarios. They observed that student performance on any 1 randomly selected passage from a collection of 20 corresponded with reliability-like coefficients of .90 (SEM = 12 WRC/min) and .81 (SEM = 18 WRC/min) for relative and absolute interpretations, respectively. When 3 measurements were taken as the basis for the score, the estimates were .96 (SEM = 7 WRC/min) and .93 (SEM = 10 WRC/min). When 5 measurements were taken, the estimates were .98 (SEM = 5 WRC/min) and .95 (SEM = 8 WRC/min). Their results illustrate that both the number of measurements and manner of interpretation influence the precision of measurement outcomes.

Further analyses by Poncy et al. (2005) illustrate that reliability-like coefficients and SEMs are improved for absolute interpretations as the quality of instrumentation is improved. When they limited their sets of passages to include alternate forms/passages that were similar (i.e., equivalent) in resultant WRC/min (M₁ ≃ M₂ ≃ M₃ ≃ Mₙ), reliability-like coefficients and SEMs were improved. For example, the reliability-like coefficient was .81 (SEM = 18 WRC/min) when 1 among 20 alternate forms was used and improved to .89 (SEM = 12 WRC/min) when the passage set was restricted to 7 alternate forms that were within 5 WRC/min of mean levels of performance. That is, the restricted CBM-R passage set was comprised of alternate passages with similar resultant mean oral reading fluency (i.e., WRC/min) among the student group. In their conclusions, Poncy et al. recommended the use of field testing and performance analysis to select passages that deviate by no more than ±5 WRC/min from mean performance. It should be noted that the improvements were observed only for absolute interpretations and not relative interpretations. Nevertheless, the results are important precisely because absolute interpretations are the dominant interpretive approach for CBM-R (i.e., benchmarking and progress monitoring). The findings of Poncy et al. provide a foundation and direction for future research to improve and optimize CBM-R instrumentation as past and existing procedures have failed to produce CBM-R passages sets consisting of equivalent passages.

Historically, CBM-R passages were drawn directly from student-specific curriculum materials (Deno, 1985; Deno, Marston, & Tindal, 1985; Fuchs & Deno, 1994; Good & Jefferson, 1998). The sampling of materials from within the students’ curriculum helped to ensure high content overlap between assessment and instructional materials. However, random sampling of curriculum materials did not ensure any substantial control of difficulty level across alternate forms or passages within sets of CBM-R passages used to monitor progress (i.e., CBM-R passage sets). These passage sets resulted in inconsistent student performance (Fuchs & Deno, 1992, 1994) as students’ WRC/min varied as a function of the variability in the difficulty of the passages within passage sets. Recognizing the deleterious effect of inconsistencies in passage characteristics across alternate passages on the reliability, accuracy, and stability of CBM-R outcomes, the field has since shifted to the use of curriculum-like, rather than curriculum-based CBM-R passage sets (Christ, 2003; Hintze, Daly, & Shapiro, 1998; Hintze, Owen, Shapiro, & Daly, 2000; Hintze & Shapiro, 1997; Hintze, Shapiro, & Lutz,
1994). Although the characteristics of a reading passage can vary along multiple dimensions, what is critical is that they are equivalent in resultant WRC/min. Careful analysis and procedures to equate alternate forms are necessary in light of probable non-equivalence between passages (Betts, Pickard, & Heistad, 2009-this issue; Francis, Santi, Barr, Fletcher, Varisco, & Foorman, 2008). Such cautions have long been overlooked within the context of CBM interpretations.

In an attempt to develop passage sets composed of equivalent level CBM-R passages, readability formulas have been employed to select and assess the difficulty of passages. These formulas, however, integrate syntactic and semantic information to yield quantitative estimates of comprehension difficulty and not reading fluency (Bailin & Grafstein, 2001; Klare, 1988). That is, readability formulas are not good predictors of passage-specific reading rates because they were not devised for that purpose. This limits their utility for CBM-R passage set development. Nevertheless, developers of CBM-R passage sets (e.g., Good & Kaminski, 2002; Howe & Shinn, 2002) and researchers use readability formulas to constrain passage difficulty across passages. Ardoin, Suldo, Witt, Aldrich, and McDonald (2005) identified 34 published studies where readability formulas were used to estimate the difficulty of CBM-R passages. Among the identified studies, only a select few empirically examined whether readability formulas functioned to improve CBM-R outcomes. Researchers who studied the use of readability formulas all concluded that readability formulas were poor predictors of students’ oral reading fluency on CBM-R passages (Ardoin et al., 2005; Compton, Appleton, & Hosp, 2004; Hintze & Christ, 2004). These findings suggest that other avenues to optimize the development of CBM-R passage sets consisting of equivalent level passages must be explored.

Poncy et al. (2005) is one of only two studies in the extant literature to evaluate an alternative procedure for selecting equivalent level passages. As stated previously, the results of Poncy et al. provide initial evidence to support the use of field testing and analysis to identify sets of passages with similar mean levels of performances across students. Another possible field testing approach, which has yet to be examined, is to select passages based upon their Euclidean Distance (ED). ED is the square root of the sum of squared differences between repeated measurements. This procedure, and other similar procedures (e.g., Mahalanobis distance, Cook’s distance, Hotelling’s T), are commonly used in classification studies (e.g., cluster analysis) and to detect outlier values for modeling (e.g., linear regression) and multivariate analysis. Distance measures are used to group (similar) observations/passages and identify outliers (dissimilar) observations/passages. The $M(ED)$ provides an estimate of the relative (dis)similarity of student performances on specific passages. For example, a set of 20 CBM-R passages with $M(ED) < 120$ would yield more consistent performances than those with $M(ED) > 120$. Because $M(ED)$ is an estimate of (dis)similar passage-specific performance within and across students, the selection of passages with lower magnitudes of $M(ED)$ is likely to improve the consistency of performances across CBM-R administrations that depend on alternate forms. Researchers hypothesized that passage selection using $M(ED)$ would improve reliability-like coefficients as compared to those observed for random passage-selection or selection of passages with similar mean levels of performance (viz., Poncy et al.).

G-theory provides a good framework for evaluating sources of variance and comparing methods with respect to the generalizability and dependability of results. It is a modern measurement theory built upon applications of both Classical Test Theory (CTT) and ANOVA
The context of CTT is commonly used to examine issues of score consistency. However, G-theory might provide a more flexible and refined approach. That is, error variance is treated as one undifferentiated mass whose component influences are left unexamined in CTT (Cronbach & Gleser, 1972). In contrast, G-theory is used to deconstruct and examine the component influences of random variance. Prior to analysis, the potential influences are identified and then defined as facets of measurement. Each facet is a characteristic of the measurement conditions with multiple levels. For example, when Poncy et al. (2005) examined the influence of individual CBM-R passages, the facet included 20 levels (i.e., alternate passage forms). Other facets that appear within the CBM literature include the duration of assessment (Christ et al., 2005), arrangement of stimulus content (Christ & Vining, 2006), type/uniformity of stimulus content (Hintze, Christ, & Keller, 2002), and number of measurements used to estimate growth (Hintze et al., 2000).

G-theory is used to estimate the relative influence of one or more facets. Moreover, the universe of admissible measurements, which defines the parameters of acceptable measurement conditions, is specified by the researcher so that measurement conditions can be as loosely or tightly controlled as the purpose and context of assessment require. In general, the universe of admissible CBM-R conditions is broadly defined to include any number of variations across administrators/raters, passages, and settings. G-theory might be used to extend the findings of previous research and contribute to developing an improved understanding of how CBM-R passages and passage sets influence outcomes.

The purpose of this study was to examine four procedures that might be used to assist in the selection of CBM-R passages: (1) random selection, (2) selection based on readability results, (3) selection based on mean levels of performance from field testing, and (4) use of ED procedures. Methods of random selection and readability represent the historical approaches to CBM-R passage set development. The method to select passages with proximate mean levels of performance during field testing was derived from procedures used by Poncy et al. (2005). Finally, the method of selecting passages with minimum ED levels constitutes a novel procedure. Because ED is an estimate of inconsistencies in performance within and across both students and passages, it was hypothesized ED would be the optimal method to identify CBM-R passages in comparison to the other three methods.

**Method**

**Participants and setting**

Letters of consent were sent to the parents of all general education second and third grade students enrolled in an elementary school serving approximately 500 students in grades kindergarten through fifth. Demographic information provided by the school indicated that 76% of the school population was Caucasian, 17% African American, and 6% were identified as Other. Approximately 16.6% of students enrolled at the school received free or reduced lunch.

Participants included 46 second (20 male) and 42 third (18 male) grade students, representing approximately 50% of the second grade students and 40% of the third grade
students enrolled in each grade. Second grade students ranged in age from 7 years, 6 months to 8 years, 8 months (M = 8 years, 1 month) and third grade students ranged in age from 8 years, 7 months to 9 years 7 months (M = 9 years). Most of the participants were Caucasian (77%) or African American (19%). All sessions took place either in a quiet hallway near the students’ classrooms or in teacher workrooms. Audio recorders were present for the majority of sessions in order to assess inter-rater agreement and procedural integrity.

Materials

Curriculum-based measurement passages

Passages were selected from nine third grade reading textbooks using methods described by Shinn (1989) and Hintze et al. (1998) (Appendix A). A random page of each story served as the beginning point for each CBM-R passage. The first full paragraph on the page began the passage. Only a single passage was sampled from each story, even if the story was published in more than one text. Only stories that were narrative in text were used. Passages that were poetic or written in dramatic style were excluded. The proper nouns within all stories were reviewed and proper nouns that were considered difficult to pronounce by two undergraduates were changed to more common proper nouns. Passages were pasted into a web-based program (OKAPI; available on www.interventioncentral.com, Wright, 2005), which formats text into reading passages for the examiner (i.e., passage plus cumulative word count) and the student (i.e., passage). A total of 50 third grade CBM-R passages were developed and administered to both second and third grade students.

Four procedures were used to develop passage sets from the original 50 third grade CBM-R passages. Those procedures are described as random, Spache, mean, and Euclidean Distance (ED) selection methods; and they are described in detail below. A total of 20 passages were selected to comprise each passage set. The procedures yielded 4 sets of 20 passages with some passages being common across passage sets. Although all second and third grade students read the same 50 third grade passages, 4 sets of 20 passages were developed for second grade students and 4 sets of 20 passages were developed for third grade students. On average, there was an overlap of 7 passages selected across the passage sets (Table 1). Overlap in passages across passage sets ranged from 4 passages (i.e., 2nd grade mean and 2nd grade ED passage sets) to 10 passages (i.e., 2nd grade random and 2nd grade ED passage sets; 3rd grade Spache and 3rd grade ED passage sets).

The random passage sets were comprised of 20 passages selected with a Microsoft Excel-based random number generator. Each of the 50 third grade passages was assigned a number. A spreadsheet was developed to randomly select 20 passages from the initial collection of 50 passages. The procedure was first conducted to develop a random set of passage for the analyses to be conducted with the second grade data and then repeated to develop a set of random passages to be conducted in the analyses of third grade data (Table 1). The procedure was generally consistent with and representative of random curriculum sampling. Although curriculum sampling is typically confined to the specific curriculum materials of the student, research has established that the reliability, criterion validity, and developmental growth rates of CBM-R generalize across curricula (Fuchs & Deno, 1992).
Table 1
Descriptive statistics for student performance on 50 CBM-R passages within and across grades

<table>
<thead>
<tr>
<th>Passage</th>
<th>Spache&lt;sup&gt;a&lt;/sup&gt;</th>
<th>Performance (WRC/min)&lt;sup&gt;b&lt;/sup&gt;</th>
<th>Sp</th>
<th>Passage set composition&lt;sup&gt;c&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2nd and 3rd</td>
<td>2nd 3rd Combined</td>
<td></td>
<td>R M ED R M ED</td>
</tr>
<tr>
<td></td>
<td>GE M (SD) M (SD) M (SD)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>4.93 109 (35) 138 (47) 123 (44)</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>4.39 100 (39) 126 (45) 112 (44)</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>3.13 93 (34) 115 (36) 104 (36)</td>
<td>X X X X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>4.00 100 (36) 132 (45) 116 (43)</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>3.45 113 (40) 141 (46) 127 (45)</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>3.38 99 (39) 127 (48) 112 (46)</td>
<td>X X X X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>3.68 105 (36) 131 (41) 118 (40)</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>4.33 108 (41) 144 (48) 125 (48)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>3.09 75 (37) 103 (43) 88 (42)</td>
<td>X X X X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>3.68 88 (35) 114 (41) 101 (40)</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>4.11 104 (38) 134 (47) 118 (45)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>3.17 94 (36) 126 (44) 110 (43)</td>
<td>X X X X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>3.68 96 (35) 127 (44) 111 (42)</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>4.81 78 (38) 108 (45) 93 (44)</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>3.43 95 (31) 124 (38) 109 (37)</td>
<td>X X X X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>3.33 109 (35) 138 (45) 123 (43)</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>3.31 95 (39) 124 (48) 109 (46)</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>4.36 76 (35) 114 (45) 94 (44)</td>
<td>X X X X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>4.06 104 (39) 130 (44) 117 (43)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>3.73 96 (42) 128 (48) 112 (48)</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>4.24 88 (36) 116 (41) 102 (41)</td>
<td>X X X X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>4.21 70 (32) 103 (42) 86 (41)</td>
<td>X X X X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>23</td>
<td>3.80 101 (35) 124 (42) 112 (40)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>24</td>
<td>3.92 81 (33) 112 (46) 96 (42)</td>
<td>X X X X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>3.24 90 (36) 116 (41) 102 (41)</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>26</td>
<td>3.76 95 (35) 122 (40) 108 (40)</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>27</td>
<td>3.52 101 (39) 130 (47) 115 (46)</td>
<td>X X X X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>28</td>
<td>4.07 94 (34) 118 (42) 106 (40)</td>
<td>X X X X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>29</td>
<td>4.10 113 (34) 134 (40) 123 (38)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>30</td>
<td>3.64 84 (32) 110 (39) 97 (38)</td>
<td>X X X X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>31</td>
<td>4.02 95 (37) 117 (42) 106 (41)</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>32</td>
<td>4.59 90 (36) 112 (44) 101 (41)</td>
<td>X X X X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>33</td>
<td>4.60 103 (29) 121 (33) 112 (32)</td>
<td>X X X X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>34</td>
<td>3.81 83 (35) 112 (43) 97 (42)</td>
<td>X X X X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>35</td>
<td>3.47 116 (43) 141 (53) 128 (49)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>36</td>
<td>3.96 76 (32) 104 (36) 90 (37)</td>
<td>X X X X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>37</td>
<td>3.45 93 (31) 111 (33) 102 (33)</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>38</td>
<td>3.60 97 (44) 129 (47) 112 (48)</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>39</td>
<td>3.66 90 (35) 123 (39) 106 (40)</td>
<td>X X X X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>40</td>
<td>3.43 97 (35) 119 (41) 108 (40)</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>41</td>
<td>4.78 71 (37) 98 (39) 84 (40)</td>
<td>X X X X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>42</td>
<td>3.58 86 (36) 117 (45) 101 (43)</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>43</td>
<td>4.25 89 (34) 117 (44) 103 (41)</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>44</td>
<td>3.73 91 (39) 116 (43) 103 (43)</td>
<td>X X X X X X X X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>45</td>
<td>3.92 78 (32) 102 (33) 89 (35)</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>46</td>
<td>3.97 93 (33) 122 (42) 107 (40)</td>
<td>X</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(continued on next page)
The Spache passage set was comprised of 20 passages that were selected based on Spache readability estimates.¹ Those estimates are presented in Table 1 (M = 3.88, SD = .46, range: 3.09–4.93). The passage samples that were used to construct each passage were derived from third grade curricula; however, the participants were comprised of second and third grade students. Therefore, passages were arranged in rank order by Spache readability estimates in grade equivalent of units. The passage closest to 3.5 was identified (i.e., Passage 27) as the center point. The passages ranked below (n = 9) and above (n = 10) the center point were selected, resulting in a total of 20 passages (Table 1). The Spache readability of these passages ranged from 3.24 to 3.73 (M = 3.54). The same Spache passage set was used across grades two and three.

Mean passage sets were comprised of 20 passages each based upon data collected through field testing. The passages were selected with reference to descriptive statistics of group-level performance on each passage. The passages were sorted by the mean performance among students within each grade and the middle 20 passages were selected (i.e., rank orders of 15–35; Table 1). The Spache readability of these passages ranged from 3.24 to 3.73 (M = 3.54). The same Spache passage set was used across grades two and three.

ED passage sets were comprised of 20 passages each that were selected based on data collected through field testing. Estimates of Euclidean Distances, which is the square root of the sum of squared differences for student performances across CBM-R passages, were used to select passages. SPSS (SPSS Inc., 2004) was used to estimate Euclidean Distances for each passage. Distances were greater when student performance was more variable within and across passages. Distances were lesser when student performance was less

¹ Spache (1953). These readability estimates are vocabulary based and place considerable emphasis on the percentage of words in a passage that are not present in the Dale list of 769 words (Dale, 1931). Words not included in the Dale list are considered difficult, but are only counted once when calculating the readability estimates regardless of the number of times they appear. Also considered in these readability estimates are the average number of sentences per 100 words. Formula = (.121 × word/sentence) + (.082 × percent unfamiliar words) + .659.
variable within and across passages. The passages were then sorted by the mean Euclidean Distances within each grade. The 20 passages for each grade with the smallest mean Euclidean Distance were selected (i.e., rank order 1 through 20; Table 1).

Procedures

For each student, a random number function in Microsoft Excel was used to sequence the order of the 50 CBM-R passages. Prior to the first session, passages were arranged in the designated order for each student and the 50 passages were divided into five groups of 10 passages. Standard CBM-R instructions were provided to students prior to the first three passages on day 1 and prior to the first passage on subsequent days. Following the reading of the first three passages on day 1 and the first passage on subsequent days, the CBM-R instructions were modified and students were told, “same instructions as before.” In addition to informing students of the instructions each day, students were also informed, prior to reading each passage, that they could earn $0.05 for exceeding a predetermined score. This contingency was established to reduce the likelihood that changes in student performance on passages within and across days was a function of changes in the difficulty of the passages and not changes in student motivation. Researchers assumed that student interest and motivation might wane over the course of 50 CBM-R administrations. A similar contingency was maintained for all students across all administrations. For the first three passages administered on day 1, students were informed that they could earn $0.05 for doing their best reading. All students were provided with a nickel for each of their first three readings. For all remaining passages on day 1 and all passages on subsequent days, students were informed of their criterion for earning a nickel prior to reading each passage. The criterion for each student was the student’s median WRC/min from the first three passages read on day 1. See Appendix B for exact instructions used across passages administered.

We attempted to administer all 50 passages within an academic week for each student in order to minimize the effect of learning on student performance. Unfortunately, due to student absences and school vacations all passages were not administered within one school week, with the exception of one case (i.e., passages administered within 4 days). The greatest number of days that elapsed between the first and last session for any student was 12 days, with the mean number of days between the first and last sessions being 6.38 days.

Passages were selected to comprise each of four passage sets within each of two grades. The Spache passage set was based on readability analysis, which yielded a consistent set of 20 passages across grades (Table 1). The random passage sets were based on the results of a random number generator, which was run independently to select 20 passages within each grade. The mean passage sets and ED passage sets were based on the analysis of within grade field testing results. Descriptive statistics were calculated for each passage within each grade level condition. A $M$(performance) and $M$(ED) score was estimated for each passage within each grade. The 20 passages with $M$(performances) that best approximated the grade mean within each grade were selected to comprise the mean passage set. The 20 passages with the smallest $M$(ED) within each grade were selected to comprise the ED passage set.
Procedural integrity and interscorer reliability

Experimenters were trained in administration and scoring procedures during two 45-min sessions. Prior to conducting assessments, examiners had to achieve 100% procedural integrity and interscorer reliability with the first author during three consecutive assessments. During each administration, examiners followed a 22-step procedural checklist to ensure that they provided students with correct instructions. The checklist is available from the first author upon request.

Procedural integrity and interscorer agreement were assessed on 40% of the sessions. Independent raters listened to tape recorded sessions, while scoring a reading passage and recording whether the examiner adhered to the administration protocol. Procedural integrity was calculated by dividing the number of steps completed by the total number of steps and multiplying by 100 to obtain a percentage. Procedural integrity ranged from 90% to 100% ($M=99.99\%$). Interscorer agreement was assessed on a word-by-word basis by comparing each word the examiner recorded as correct or incorrect to each word the independent rater scored as correct or incorrect. The number of agreements (correct and incorrect) between the examiner and rater was divided by the total number of words and multiplied by 100 to obtain a percentage. Inter-rater agreement ranged from 71% (1 session, next lowest session=91%) to 100% ($M=99.72\%$).

Data analysis

A Generalizability Study (G-study) was used to estimate the amount of variance associated with the target of measurement ($p$, person), the item facet ($i$, item), and residual error ($e$, $pi$, $e$). The students and their level of reading fluency were the targets of measurement. The passages were the items. Multiple G-studies were conducted to estimate total variance and the proportions associated with the target of measurement ($p$), passage/item ($i$), and residual error ($e$, $pi$, $e$). The VarComp procedures in SPSS (2004) were used to estimate the magnitude of variance associated with person, item, and residual error. The statistics were then entered into GENOVA (Crick & Brennan, 1983) to complete a series of Dependability Studies (D-studies), which estimated the generalizability coefficient ($E\rho^2$) and dependability index ($\Phi$). These coefficients are similar to reliability coefficients as they are conceptualized within CTT. The generalizability coefficient is a reliability-like estimate used to guide relative interpretations of test scores, which includes inter-student rank order interpretations (e.g., percentile ranks, standard scores). The dependability index is a reliability-like estimate used to guide absolute interpretations of test scores, which includes the interpretation of the actual value of untransformed test scores such as WRC/min. Both the generalizability coefficient and dependability index were used to estimate the SEM for relative ($\delta$) and absolute ($\Delta$) decisions. The equations to calculate these values are outlined in Poncy et al. (2005).

Results

Data were screened for missing values. One case from each grade-level condition was removed because it contained one or more missing values. Each distribution within and
Fig. 1. Mean performance and standard errors (WRC/min) on 50 CBM-R passages for students within second and third grades.

Fig. 2. Deviation scores (WRC/min) on 50 CBM-R passages for students within second (bottom panel) and third grades (top panel).
across passage and grade-level conditions were within reasonable limits to assume normality (skew ≤ 1.5, kurtosis ≤ 1.5). Group level performance among students in the second and third grades is summarized in Table 1 and Fig. 1. For second grade participants, the grand mean across all CBM-R passages was 94 (SD = 36) WRC/min. The lowest mean for any passage was 70 (SD = 32) and the highest was 116 (SD = 43) WRC/min. The grand mean among third grade participants was 121 (SD = 43) WRC/min. The lowest mean was 98 (SD = 39) and the highest was 144 (SD = 48) WRC/min. Deviation scores were calculated within each grade. Each deviation score was calculated as the difference between the mean level of performance across passages and performance within each passage (See Fig. 2).

Although the mean level of student performance did fluctuate substantially between alternate passages, alternate form reliability was robust. The mean of all alternate form reliability coefficients was .92 (SD = .02) for second grade and .93 (SD = .02) for third grade. The smallest coefficients were .82 and .84 for the second and third grade conditions, respectively. The largest coefficient was .97 in both grade-level conditions. These coefficients converge with the outcomes of previous research (Ardoin & Christ, 2008; Christ & Silberglitt, 2007; Howe & Shinn, 2002; Marston, 1989). Alternate form reliability remained robust despite substantial differences in the mean level of performances between passages. Further analysis was conducted to examine the proportion of variance associated with student specific skills, passage characteristics, and error.

Table 2 summarizes the results of the G-studies. In the case of the random passage sets, the variances (percentage of total variance) associated with person, item, and residual error

<table>
<thead>
<tr>
<th>Words read correct per min (WRC/min)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Observed variance</td>
</tr>
<tr>
<td>Person</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>2nd grade</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Random</td>
<td>1151.03</td>
<td>139.42</td>
</tr>
<tr>
<td>Spache</td>
<td>1246.70</td>
<td>70.60</td>
</tr>
<tr>
<td>Mean</td>
<td>1113.36</td>
<td>47.3</td>
</tr>
<tr>
<td>ED</td>
<td>1137.89</td>
<td>15.77</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>3rd grade</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Random</td>
<td>1635.42</td>
<td>109.88</td>
</tr>
<tr>
<td>Spache</td>
<td>1771.93</td>
<td>80.11</td>
</tr>
<tr>
<td>Mean</td>
<td>1674.46</td>
<td>29.98</td>
</tr>
<tr>
<td>ED</td>
<td>1671.68</td>
<td>17.56</td>
</tr>
</tbody>
</table>

---

*a Variances estimated with Type 3 sum of squares.

*b One facet model with person (i.e., student) as the target of measurement and item (i.e., passages) as the measurement facet; percentages for variance components were rounded so the sum of estimates might diverge from 100%.

*c N(person) = 45, N(item) = 20; the items for each G-study were specific to the random, mean, or Euclidean Distance (ED) passage set.

*d N(person) = 42, N(item) = 20; the items for each G-study were specific to the random, mean, or Euclidean Distance (ED) passage set.
were 1151.03 (82%), 139.42 (10%), and 121.85 (9%) in second grade and 1635.42 (86%), 109.88 (6%), and 139.79 (8%) in third grade. Person variance can be interpreted as the variance of mean scores (across all passages in the passage set) for all persons in the population. Item variance is the variance of mean scores (across all persons in the population) for all passages in the universe of admissible observations. The error variance is that which is attributed to either an interaction between person and passages or residual error. The percentage of total variance can be interpreted as the proportion of variance attributable to each variance component.

Person variance was increased and item variance was reduced below what was observed for the random passage sets in the cases of the Spache, mean, and ED passage sets. Person variance (percentage of total variance) increased to 1246.70 (86%), 1113.36 (87%), and 1137.89 (91%) for the Spache, mean, and ED passage sets within passage sets selected based upon the second grade students’ performances and they increased to 1771.93 (89%), 1674.46 (91%), and 1671.68 (93%) for those passage sets selected based upon third grade student performances. Item variance (percentage of total variance) was reduced to 70.60 (5%), 47.3 (4%), and 15.77 (1%) for the Spache, mean, and ED passage sets within the second grade condition. In third grade, they were reduced to

Fig. 3. Dependability studies (D-studies) for the second grade sample assuming a range of one to ten CBM-R are administered to estimate the universe score. The generalizability coefficients ($\hat{E}\hat{p}^2$; upper left quadrant) is a reliability-like estimate for relative interpretations, which corresponds with a standard error of measurement ($SEM$) present below it ($\delta$; lower left quadrant). The dependability index ($\Phi$; upper right quadrant) is a reliability-like estimate for absolute decisions, which corresponds with $SEMs$ presented below it ($\Delta$; lower left quadrant).
80.11 (4%), 29.98 (2%), and 17.56 (1%) for those passage sets within the third grade condition.

A series of D-studies was run to estimate generalizability coefficients (\(E\hat{\rho}^2\)) and dependability indices (\(\Phi\)) along with the estimates of \(SEM\) for relative (\(\delta\)) and absolute (\(\Delta\)) interpretations of CBM-R outcomes (Figs. 3 and 4). The results for relative interpretations converged with those of previous research. The \(E\hat{\rho}^2\) (\(\delta\)) in both second and third grade conditions were approximately .91 (11 WRC/min), .95 (8 WRC/min), and .97 (6 WRC/min) when one, two, and three CBM-R passage readings were used to estimate typical performance. The mean, Spache, and ED passage sets each performed better than the random passage set, but differences were generally trivial for relative interpretations. That is, the magnitudes of both \(E\hat{\rho}^2\) and \(\delta\) were generally consistent across passage sets with only a minor improvement for the ED passage set.

In contrast, differences were robust for absolute interpretations so that the magnitudes of both \(\Phi\) and \(\Delta\) were most optimal for the ED passage set. As depicted in Fig. 3, if one CBM administration were used to estimate the level of student performance, then the magnitude of \(\Phi\) (\(\Delta\)) was .91 (10 WRC/min). Lower estimates were obtained for the other conditions: the random, Spache and mean passages sets corresponded with estimates of .82 (16 WRC/min), .86 (14 WRC/min), and .87 (13 WRC/min), respectively. The panels on the right of Fig. 3 illustrate that the magnitudes of both \(\Phi\) and \(\Delta\) for absolute decisions.
are consistently less desirable for the random passage set. In rank order, the most optimal magnitudes were for ED, mean and Spache passage sets. The same pattern emerged for absolute decisions the third graders, as is illustrated in Fig. 4. It seems that ED passage sets might be preferred in the case of absolute interpretations, which include benchmarking and progress monitoring applications.

Discussion

The purpose of this study was to examine four procedures to guide CBM-R passage-set selection: (1) random selection, (2) selection based on readability (Spache) results, (3) selection based on mean levels of performance from field testing, and (4) use of ED procedures. The procedures used to construct the random passage sets, the Spache passage set (see Hintze & Christ, 2004; Hintze et al., 1998; Hintze & Shapiro, 1997), and the mean passage set (see Poncy et al., 2005) were each consistent with previously published recommendations. The ED passage set was examined as a novel alternative to previously published methods for passage selection. Descriptive statistics, psychometric analysis and generalizability theory were used to evaluate student performance on an initial set of 50 passages and their performance on each of four subsets of 20 passages, which were established using each of the four passage selection procedures identified above. The quality of the passage sets were initially evaluated in a series of G-studies, which yielded estimates for the distribution of variances associated with the target of measurement (student/person), passages, and residual error. Those were followed by a series of D-studies, which yielded estimates of reliability-like coefficients (\( E\rho^2 \), \( \Phi \)) and SEM (\( \delta, \Delta \)).

The results suggest that readability-based passage selection is slightly better than random selection; however, neither approach is likely to yield optimal passage sets. The consistency of performance across passages is poorly controlled by random or readability-based passage selection. Instead, the process of field testing and performance analysis is likely to result in a more optimal set of alternate forms. In this study, improved outcomes were observed in both the mean and ED conditions, with the most improvement associated with ED passage sets. That observation was replicated across two grade level conditions.

The grand means of student performances across 50 passages were 94 (SD=36) and 121 (SD=43) WRC/min for students in second and third grades, respectively. As expected, third grade students read more WRC/min than students in second grade. There was also substantial passage-specific variability within each grade (Fig. 1) and deviation scores of robust magnitudes (Fig. 2). The difference between the average performance on the easiest and hardest passages was 46 WRC/min. That difference was observed in both second and third grade. Passage-specific differences of those magnitudes would certainly translate into substantial variability during screening, benchmarking or progress monitoring. The variability due to inconsistent passage characteristics could certainly obscure interpretation and introduce a further source of error in data-based decision making.

Despite the observed variability, there were robust levels of alternate form reliability (rank order, \( E\rho^2 \)). The mean for alternate form reliabilities across CBM-R passages was .92 (SD=.02) and .93 (SD=.02) within the second and third grade conditions, respectively. Those outcomes are consistent with previously published studies (Ardoin & Christ, 2008;
Christ & Silberglitt, 2007; Howe & Shinn, 2002; Marston, 1989). However, robust findings are limited to rank order interpretations of CBM-R outcomes because the forms are not parallel. That is, the alternate forms are not parallel forms in that they do not yield consistent mean levels of performance. As a result, the forms rank students consistently, but they do not yield consistent raw scores in WRC/min. The raw score is highly dependent on the CBM-R form that is used. As described above, the difference in the raw scores (i.e., absolute interpretation) across two alternate form administrations of grade level materials might approximate 46 WRC/min. Such results establish substantial problems when absolute interpretations of raw scores are used, which is the case for benchmarking (using criterion referents) and progress monitoring.

Results from G-studies indicate that passage selection procedures do influence measurement outcomes. The distribution of variances was optimized by the ED passage selection procedure. The proportion of variance associated with item/passage was minimized in the ED condition (1%) and it was maximized in the random condition (6 to 10%; Table 2). There was a 6 to 10 fold reduction in measurement variance due to passage inconsistencies when ED was used instead of random selection. As compared with the results of random selection, there was a 4 to 5 fold reduction when mean procedures were used; and only a 1 to 2 fold improvement when readability procedures were used.

Dependability studies (D-studies) were used to estimate reliability-like coefficients and SEM for both relative and absolute interpretations (Figs. 3 and 4). Those results provide further evidence to support ED and mean passage set selection procedures. The results presented in the upper and lower left quadrants of Figs. 3 and 4 illustrate that reliability-like coefficients were substantially similar for all of the procedures except the ED passage sets. These results provide support for the conclusion that ED procedures might improve the reliability-like coefficients for rank order interpretations by a magnitude of approximately .01 to .02 above that of other procedures. Moreover, the ED procedure might reduce the magnitude of SEM by approximately 1 WRC/min. It is important to note that these improvements were modest and tended to wane as more measurements were collected. Both reliability-like coefficients and SEM were substantially similar after six or more measurements. However, the results presented in the upper and lower right quadrants illustrate that the differences were substantial in the case of absolute interpretations. In comparison to the random procedure, the ED procedure improved reliability-like coefficients by a magnitude of .05 to .10 and reduced SEMs by magnitudes of 4 to 6 WRC/min. Although the differences were most substantial when only a single measurement was collected, there were consistent differences after as many as 10 measurements were collected. The mean procedure for passage set development also improved the reliability-like coefficients and SEM as compared to both the Spache and random procedures. In summary, the results of this study provide clear and compelling evidence that field-testing and analysis prior to passage selection have tremendous potential to improve the generalizability and dependability of CBM-R.

CBM-R outcomes can be interpreted as both relative and absolute values. The results of this study demonstrate that how a measurement is interpreted is as important as how a measurement is derived. Reliability coefficients, generalizability coefficients, and the corresponding SEMs are most relevant when measurements are used to estimate the rank order and relative standing of students within a group. They are less applicable when
measurements are used to estimate the absolute level of performance in raw score units (WRC/min).

The results of this study and previous studies (e.g., Ardoin & Christ, 2008; Poncy et al., 2005) support the assumption that the reliability of CBM-R is likely to be robust ($r_{xx} > .90$). However, it is technically inaccurate to conclude that CBM-R outcomes are reliable and thereafter ignore error. The reliability of measurement is not a binary concept. Measurements are not either reliable or unreliable. Although some measurements are more reliable than other measurements, all are prone to error and have a restricted level of precision. It is necessary to evaluate and communicate information about the precision of measurement to the consumers of data.

Communicating the precision of measurement usually requires some reference to the actual range of student performances across multiple measurements or some derivation of the probable range of student performances. It has been proposed that one of two methods might be used to estimate the precision of a CBM-R (Christ, Davie, & Berman, 2006; Christ & Silberglitt, 2007). First, a large number of CBM-R passages can be administered and an estimate of central tendency (e.g., mean or median) can be used as an approximation of typical performance. Precision is then estimated with some measure of dispersion, such as the range or standard deviation. Second, a small number of CBM-R passages can be administered and the result from a single administration or the median of three administrations can be used to estimate the typical level of performance. Precision then can be estimated with the SEM, which will be influenced by the universe of generalization, number of items/passages, and mode of interpretation.

There are limitations associated with the current study and therefore results should be interpreted with some caution. Results are limited by a relatively small sample size within each grade as well as the selection of participants from one school. It is possible that exposure to the same curriculum resulted in greater consistency in the relative level of difficulty of each passage for students. It should, however, be noted that the SEM for level reported in the current study is similar to that reported in previous research (Christ & Silberglitt, 2007; Poncy et al., 2005). Another limitation associated with the current study is that examiners were not held constant across all sessions. Although we attempted to keep the number of different examiners to a minimum, most participants read with two different examiners across the five sessions. Attempts were also made to conduct sessions at the same location, but this was not always possible due to space concerns within the school. Unfortunately, we failed to collect data on the administrator and test location and therefore were not able to evaluate examiner and test location as facets within the G-study.

Readers should keep two things in mind when considering the results of this study. First, on average, there was a difference of 46 WRC/min between student performances on the easiest and hardest passages. That is a substantial inconsistency of measurement, which is exclusively due to the characteristics of two alternate forms. Research-based benchmarks for expected performances by the spring of second and third grades approximate 70 and

---

2 *SEM* used in discussion as estimate of precision and likely dispersion. The researchers recognize and note that there are other definitions and proxies to estimate precision.
100 WRC/min respectively (Good, Simmons, & Kame’enui, 2001; Hintze & Silbergliit, 2005). The presumed accuracy and utility of such standards are threatened when student performance fluctuates by 46 WRC/min across alternate forms. The variability in performances across grade level passages that were observed in this study should be no less than a siren call to both researchers and practitioners. Student performance is substantially dependent on the characteristics of individual passages. The consistency or inconsistency of passage characteristics will determine the measured estimate of student performance across administrations.

Second, field-testing did improve the consistency of measurement outcomes. The performance of both random and readability procedures to construct passage sets was inferior to both the ED method and mean method. It seems that those procedures are insufficient. Two alternative methods were presented here, which include field-testing and analysis of student performances across alternate forms. Analysis using both performance means and ED improved measurement outcomes as evidenced by increased reliability-like coefficients and reduced SEM.

Third, the previously summarized observations occurred within the context of robust rank order reliabilities. That is, the range of performance means across alternate passages spanned 46 WRC/min and SEMs were as large as 16 WRC/min. Those outcomes were observed along with alternate form reliabilities that were consistently above .90. Clearly, rank order reliability coefficients provide an incomplete summary of the precision and consistency of measurement. Rank order reliabilities have been reported without reference to parallel test assumptions. It is time to reexamine the evidence-base with regard to the technical and psychometric characteristics of CBM. The simplistic conclusion that CBM is valid and reliable is too broad and should be evaluated with reference to the particular decisions, purpose, and interpretation. The reliability and validity of measurement outcomes are specific to measurement conditions, measurement procedures, instrumentation, and the interpretation of outcomes.

The influence and potential impact of CBM and other similar procedures is substantial. Although CBM was initially developed to guide low-stakes instructional decisions, the scope of potential applications has expanded since the early 1980s. It is very likely that CBM will be used with greater frequency to guide high-stakes diagnostic and eligibility decisions. Ongoing research is necessary to develop and evaluate potential improvements to enhance the quality of CBM-R data. Likely targets for future development and evaluation include both the instrumentation and procedures used for assessment and evaluation. This study examined alternate approaches to establish CBM-R instrumentation. The results support future use of field-testing and systematic evaluation of student performances to enhance the equivalency of performances across alternate forms. Test theory also provides methods to equate performances across less equitable forms (Kolen & Brennan, 2004); and recently published studies establish the feasibility and utility of test equating methods to enhance interpretation procedures for CBM-R outcomes (Betts et al., 2009-this issue; Francis et al., 2008).

It is likely that future developments will substantially improve the quality of both assessment and interpretation to support data-based decision making. Ongoing research, and publications within the professional literature, is necessary to enhance assessment practices as they relate to CBM and other approaches to data collection. The urgency for such work was never greater. The field of school psychology has a dependent variable problem at the
same time that a paradigm shift is underway. Data-based decision making and problem solving are central features to the present zeitgeist in school psychology. However, the field has not yet optimized the methods necessary to realize the full benefits of the ongoing paradigm shift. Research to improve methods of data collection and provide guidance in decision making is critical to helping the field make the most of this unique opportunity.

Appendix A


Appendix B

Instructions administered prior to first three passages on day one:

“For reading this passage well, you can earn 5 cents that you can use at the school store. Beginning reading aloud here and read across and down, if you come to a word that you don’t know I’ll help you. Be sure to do your best reading. Do you have any questions? Begin reading.”

Instructions administered prior to first passage read on days 2–4:

“Here is the first story that I want you to read, beginning reading here, read across and down. If you come to a word you don’t know, I’ll help you. If you read more than ___ words correctly (median score from first day of first three passages) you’ll earn 5 cents. Do you have any questions? Begin reading.”

Instructions read to students prior to all remaining passages:

“Same rules as before, if you read more than ___ words correct you can get 5 cents. Begin reading.”
References


