

# Examining the Reliability of Running Records: Attaining Generalizable Results

PARKER C. FAWSON  
Utah State University

D. RAY REUTZEL  
Utah State University

JOHN A. SMITH  
Utah State University

BRIAN C. LUDLOW  
Alfred University

RICHARD SUDWEEKS  
Brigham Young University

**ABSTRACT** The authors present results of a generalizability study of running record assessment. They conducted 2 decision studies to ascertain the number of raters and passages necessary to obtain a reliable estimate of a student's reading ability on the basis of a running record assessment. Ten teachers completed running record assessments of 10 first-grade students on 2 leveled reading passages. Findings indicate that each student assessed with running records should read a minimum of 3 passages to produce a reliable score. Using a fully crossed design in which all students were rated by each rater on all passages did not provide sizable advantages over a nested design in which students were crossed with passages and nested in raters.

**Key words:** reliability, running record, text leveling

**R**unning records (Clay, 1993) are a widely used reading-assessment process developed originally for children's Reading Recovery programs. However, current data suggest that the use of this diagnostic and progress monitoring process has grown far beyond the boundaries of Reading Recovery. Bean, Cassidy, Grumet, Shelton, and Wallis (2002) noted that 62% of members of the International Reading Association who identified themselves as reading teachers were using running records to assess their students' reading progress. In addition to its original purpose, running records are now broadly used by reading teachers to monitor students' reading progress and to diagnose reading needs in a variety of instructional settings across grade levels (Fountas & Pinnell, 1996).

A *running record* is a test of contextual reading accuracy and student strategy use in which students read leveled connected passages under untimed conditions. The examiner typically makes a record of the types of errors (e.g., deletions, insertions, omissions) that each reader commits during oral reading. Classroom teachers have used the results of running records to establish functional reading levels. Rathvon (2004) identified functional reading levels as independent (beyond 95% accuracy), instructional (between 90% and 95% accuracy), and frustration (below 90% accuracy). Teachers also use running records to identify reading

behaviors of young children in context to guide selection of appropriate instructional interventions. Running records are used as a benchmark or standards-based reading assessment across grade levels in elementary schools.

Running records have been an attractive assessment process for early reading largely because they allow a teacher to capture various reading behaviors that young children exhibit during contextual reading. As such, running records provide teachers with data in which to make informed instructional decisions. Assessment tools and processes that reliably evaluate early reading behavior allow teachers to intervene before a student establishes a pattern of reading failure. Several decades of research have confirmed the importance of early intervention to prevent reading failure in young children (Snow, Burns, & Griffin, 1998). Early intervention is especially critical given the devastating future costs, socially and economically, for children who do not learn to read.

Effective schools research identifies regular progress monitoring and diagnosis as contributors to improved student achievement in reading (Hoffman, 1991; Matsumura, Patthey-Chavez, Valdes, & Garnier, 2002; Ross, 2004; Wharton-McDonald et al., 1997; Wray, Medwell, Fox, & Poulson, 2000). Taylor, Pearson, Clark, and Walpole (2000) and Taylor, Pearson, Peterson, and Rodriguez (2005) found that the most effective schools had a shared system for communicating progress monitoring and diagnostic assessment data within the school. When school professionals used a shared system for communicating student progress data, they worked together systematically to promptly address each student's reading instruction needs.

Pressley and colleagues (2001) found that the most effective primary-grade teachers conducted running record assessments as described in Reading Recovery during student reading instruction (Clay, 1993). Teachers who were

---

Address correspondence to Parker C. Fawson, Utah State University, Department of Elementary Education, Old Main Hill 2805, Logan, UT 84322-2805. (E-mail: [parker.fawson@usu.edu](mailto:parker.fawson@usu.edu))

Copyright © 2006 Heldref Publications



aware of running record information dealing with students' contextual reading were more likely to use these data to match students to appropriate interventions and instruction-level texts. Ross (2004) demonstrated a high correlation between teachers' frequent use of running records and students' reading achievement.

Reading researchers have posited that poor readers rely heavily on context to identify unfamiliar words because they lack the ability to rapidly and accurately decode words (Nation & Snowling, 1998; Share & Stanovich, 1995). Thus, asking struggling readers to read connected text, as is the case in a running record, may provide teachers with a helpful glimpse into how struggling readers are processing written language. Tunmer and Hoover (1992) suggested that using context to identify unknown words in text supports reading acquisition by permitting readers to enhance their basic sound-symbol knowledge with context clues to decode unknown words. Also, measures of oral reading in context, such as running records, provide a more comprehensive glimpse into comprehension processes than do measures using single-word reading (Rathvon, 2004). Running records have traditionally been viewed as producing accurate assessment results because they provide an approximation of authentic school and home reading. However, reliability data have not been conclusive regarding the use of running records (Ross, 2004).

National reading reports, legislative mandates, and competitive government grants focus on improved reading performance among young children. Within that national context of improved reading performance, there is a need to gather accurate assessment data on the most at-risk young readers. One of the most pressing challenges for identifying struggling readers is ensuring that the assessment instruments that teachers use to judge student reading progress produce reliable scores that adequately inform teachers' intervention decisions.

One conspicuously absent finding in the research on running records is the identification of sources of variability within student scores that helps establish the reliability of the running record process. To adequately meet the demands for improved reading-performance assessments, running records should meet the traditionally expected psychometric standards of reliability and validity. Without reliability and validity data on students' running record scores, parents, teachers, administrators, and policy makers cannot confidently use these results to inform teachers' decision making or to determine school-level literacy policy.

Instrument reliability refers to the degree of consistency of the scores obtained from a given measure (Huck, 2004). Traub and Rowley (1991) asserted that running record reliability is related to conditions and factors within the text and also to factors within the student. One factor within the text that affects a running record score is the difficulty level of the passage to be read. Even passages at the same reading level may have differing internal linguistic structures or cognitive concept loads that would cause student

running record scores to vary from one passage to the next. Establishing running record reliability must also take into account the varied experiences of each child that will pose a potential threat to a stable score. That is especially the case with younger children who differ developmentally, experientially, and in general reading ability. In addition to reliability threats present within the text or between students, variability exists across or within the teachers scoring the running records or the raters.

Interscorer reliability refers to the amount of variance in a test related to the variability among the raters. Reynolds (1990) suggested that test reliability requires a report of scoring consistency when raters must make fine-grained distinctions among student responses. When conducting a running record, teachers must make such an analysis to fully and consistently assess each student's reading performance. The raters of running records, typically teachers with varied levels of experience, make decisions on a wide array of reading behaviors that the child demonstrates. Teacher level of sophistication in accurately recording complex reading behaviors creates a potential threat to running record reliability. Variability among raters can result not only from varied interpretations of student responses on running records but also from the accuracy of recording student responses while taking a running record. Past research has shown that rater variance is most likely to occur on tests and in assessment processes that require rapid and accurate scoring of student responses (Reynolds).

Clay (1966) stated that "We have to be concerned with whether our assessments are reliable because we do not want to alter our teaching, or decide on a child's placement, on the basis of a flawed judgment" (p. 8). In an experimental study, Ross (2004) contrasted systematic classroom assessment using running records with a control condition. Teachers in the treatment condition completed six 60-min training sessions in which they learned how to administer and interpret running records to better inform their reading instruction. Student achievement in treatment classrooms outperformed that of the control students. Ross confirmed, however, that there has been little psychometric data reported on running records. Where data have been reported, reliability evidence has been mixed (Chapman, Tunmer, & Prochnow, 2001; Ross). Running record reliability has been especially difficult to establish given the variability in passage difficulty and rater. Chapman and colleagues identified passage difficulty and rater variance as potential threats to the reliability of running record scores. Although those variables likely affected the reliability of running record scores, no researchers have reported their impact on running record reliability. Also, little is known about how the interaction of the variables influences running record reliability.

In summary, given the wide utility and lack of strong, consistent reliability evidence for running records, we explored potential sources of error variance associated with running record administration and scoring caused by two



important sources of error variance: (a) passage difficulty and (b) rater variability. Our purpose in this study was to suggest potential changes needed for scoring running records to make them more reliable.

To investigate the effect on running records scores of the two sources of variability, passage difficulty and rater variability, we used generalizability theory (Cronbach, Gleser, Nanda, & Rajaratnam, 1972). Three research questions guided this study.

1. To what degree do inconsistencies between passages of the same level of difficulty, inconsistencies between raters, and their interactions contribute to discrepancies in scores obtained from a running record assessment (G study to establish reliability)?
2. Can one change the number of passages or the number of raters, or both, to optimize the reliability of the obtained running record scores, and, if so, what changes are needed (D study to affect decisions)?
3. What effect will using a nested design have (students crossed with passages and nested in raters), rather than expecting every rater to rate every passage read by every student, on the error variances and on the overall reliability of the running record ratings (D study to affect decisions)?

## Method

### Participants

*Teachers and raters.* Teacher participants were selected from a pool of all first-grade teachers who taught in a large suburban school district in the western United States. One member of the *research team* (the authors) made contact with the district reading coordinator, who then solicited willing volunteers from a pool of first-grade teachers. From that set, one researcher, in collaboration with the district reading coordinator, identified 10 teachers who represented a wide range of training in reading. The 10 teachers were contacted by a researcher and agreed to participate in this study.

Prior to their selection for participation, all teachers received from 2 to 6 hr of training on running record use to assess student reading performance. The training was provided through their participation in a literacy conference sponsored by a local partner university. Three running record training sessions were presented in this conference. A reading recovery-trained district literacy specialist taught each 2-hr session. Teacher participation in the sessions was voluntary; they could attend as much or as little training as they wanted. As a result of the training, all participating teachers had used running records in their classrooms at varying frequencies and for different purposes.

Three teachers reported using running records in their classrooms daily to assess student progress in reading. Three other teachers applied running records on a weekly basis to appraise student reading. One teacher used running records on a monthly basis as a general check on student reading

progress. The final 3 teachers reported using running records three times per year as directed by district policy. In addition to their level of running record training, the amount of elementary teaching experience also varied from teacher to teacher. Three teachers had more than 15 years of classroom experience, 4 teachers possessed 5 to 15 years of experience, and 3 teachers had less than 5 years in the classroom.

All participating teachers indicated that one of the main purposes for using running records in their classrooms was to monitor student progress toward achieving established reading benchmarks. Each teacher had to use a running record three times per year to assess student reading. However, the three teachers who conducted daily running records also reported that they examined the types of errors that students were making and used this information to adjust their instruction to meet resulting reading gaps.

The teachers indicated that when they used running records to assess benchmarks, they needed to use text identified by their district. Also, teachers who conducted regular running records stated that they selected instruction-level text from their classroom book collection. When selecting the texts, the teachers generally used books with a narrative format.

*Students.* The participants included 12 first-grade students who were selected from a school in a large suburban district. The socioeconomic status of the school was average; the school served a population of 571 kindergarten through fifth-grade students. The ethnic makeup of the school was 6% Asian, 1% African American, 9% Hispanic, 1% American Indian, 2% Pacific Islander, and 81% Caucasian. Students spoke a variety of languages that included Korean, Chinese, Spanish, and Japanese; English was the language of instruction. All students participating in this study spoke fluent English and had recently completed first grade.

An important consideration in this study was the selection of students who represented varied reading-performance levels. We consulted the first-grade teachers to help identify student participants. Classroom teachers assigned each student to one of the three groups on the basis of a set of criteria. The criteria included (a) teacher observation, (b) student performance on classroom assignments, and (c) end of level reading-test scores. Using those criteria, we identified 4 students as above-average readers, 4 students as average, and 4 students as below-benchmark readers. The above-average students were at least one grade above grade level, as measured by the end of level reading-test scores. The average students were within a range of 5 months of grade level. The below-average students were at least one grade below grade level. We required students to read two grade-level narrative passages. We removed 2 students from the below-average group from the sample because they could not read the selected text, leaving a total of 10 student participants—5 girls and 5 boys. (See Table 1 for a summary of student characteristics).



TABLE 1. Participating Student Characteristics

Participant	Gender	Language background	Reading level	Ethnicity
1	F	Spanish	Average	Hispanic
2	F	English	Above average	Caucasian
3	F	Chinese	Above average	Asian
4	F	English	Below average	Caucasian
5	F	English	Average	Caucasian
6	M	English	Average	Caucasian
7	M	English	Below average	Caucasian
8	M	English	Average	Caucasian
9	M	Spanish	Above average	Hispanic
10	M	English	Above average	Caucasian

Note. F = female; M = male.

### Texts

The narrative passages for this study were selected by a veteran Reading Recovery teacher in the participating school district from the district's benchmark running record text collection. The two narrative passages were "George the Porcupine" (1997) and "The Wagon" (1997) and were Reading Recovery Level 14 texts. We used Level 14 texts because they approximate the type of text that an average first-grade student should be able to read toward the end of the year.

"George the Porcupine" is part of the reading recovery text collection and is used widely to assess student reading performance through running records. The passage follows a narrative text structure. A short summary of the story is provided in the reading recovery book collection. "George is a porcupine that lives under Mr. Jay's porch. George has sharp, pointed quills. The quills hurt if you get stuck with them. Mr. Jay petted all kinds of animals, but he didn't pet George".

"The Wagon," used in the Developmental Reading Assessment (DRA; a popular reading test), follows a narrative text structure; it is a Reading Recovery Level 14 text. The DRA summary of the text is provided to each reader. "This story is about what happens to the wagon that Kevin's big brother got to carry his newspapers. Kevin's two brothers and his sister each used the wagon for different things".

Both texts used in this study follow a narrative text structure (Mandler & Johnson, 1977; Thorndyke, 1977). Story structure describes the necessary elements to make a story and the expected sequence for these elements. Researchers generally agree on the following elements and sequence of elements in a story: setting, problem, goal, events, and resolution. The Reading Recovery teacher selected the two texts, in part, because teachers frequently used them to conduct running record assessments. Both texts follow a similar text structure, but the topic addressed in each text is different, potentially requiring varied experience from the reader.

The teacher selected text that was similar in several ways and dissimilar in one. Both texts were similar regarding (a) frequency of use in conducting running records, (b) identified text difficulty level, (c) text type (narrative), and (d) text structure (story). We did not control for the topic addressed in each story. That consideration is important because we wanted to replicate the way that teachers typically conducted running records with existing leveled text and passages that vary by topic.

### Design

We used a fully crossed, three-factor measurement design: 10 students crossed with 10 raters who each rated two passages (Crocker & Algina, 1986; Shavelson & Webb, 1991). We considered the 10 students as the object of measurement (Shavelson & Webb) in the design. The object of measurement and each additional factor, student, rater, and passage represented a potential source of error in the ratings. We treated student, rater, and passage as random factors in the design because we viewed the sample of each as "smaller than the size of the universe" and "exchangeable with any other samples of the same size drawn from the universe" (Shavelson & Webb, p. 11).

Historically, researchers considered that a measurement score is reliable when systematic variance is relatively high and random variance is low (Crocker & Algina, 1986; Shavelson & Webb, 1991). We evaluated two assessment designs in this study to explore optimal generalizability and to reduce random error. Design 1 was a fully crossed, three-factor, Student  $\times$  Passage  $\times$  Rater design. Design 2 was a three-factor nested design with students crossed with passages and nested in raters.

### Procedure

To collect data for analysis, one of the researchers videotaped the 10 participating students as they read both of the narrative, Reading Recovery Level 14 passages. A video



camera closely approximated the angle from which a teacher would typically view a child during the administration of a running record. The teachers used the video recording of the 10 students reading the texts to score the running record for each passage. During the video recording, a table and chairs were set up, and a research team member invited students to sit at the table. The researcher gave the parents the option to sit on a nearby couch, browse through the books in the library, or wait outside the classroom.

Once we collected video data for each participant, a research team member asked the 10 teachers to complete a running record in the standard format (oral reading error rate) on each of the 10 students and to tabulate their data. Each teacher completed a running record assessment on each of the 10 students reading the first narrative, Level 14 passage, "The Wagon." The teachers had time between each students' reading to tabulate their data and calculate the running record score. Next, each teacher completed a running record on each of the students as they read the second narrative, Level 14 passage, "George the Porcupine." All teachers completed their assessments of the students in the same order and within approximately 2 hr and 15 min. They were not allowed to stop the videotape during the time when the student was reading, but they could pause briefly between students.

At the outset of our conducting this study, we faced a significant task. To establish the generalizability of running record results, some means of determining instrument reliability needed to be applied. However, traditional methods of determining the stability of running record results presented several challenges. In classical test theory (CTT), which has been applied traditionally to reliability studies, multiple-reliability coefficients for various sources of possible error can be estimated for each instrument (i.e., interrater or alternate form). However, because the coefficients are rarely equivalent, test users must decide which one is most appropriate for use with their application. An inherent weakness in CTT is its inability to simultaneously analyze multiple sources of error, evaluate the interactions of these errors, or predict the overall impact of measurement error (Eason, 1991). Generalizability (G) theory was developed to help researchers overcome the inherent limitations of CTT (Rathvon, 2004). G theory is used with increasing frequency to establish the dependability of behavioral assessments (Hintze, Owen, Shapiro, & Daly, 2000; Hintze & Petitte, 2001; Wolfersberger, Reutzell, Sudweeks, & Fawson, 2004). When using G theory, multiple sources of error can be considered that may have an impact on student scores, as well as on the interaction effects of various error components or factors (Brennan, 1992; Glissmeyer, 1999).

G theory is a type of conceptual framework that employs statistical methods with analysis of variance (GENOVA) to assess the stability of a measure, instrument, or process. G theory assumes that each student's observed score is com-

prised of a *universe score* (the student's average score over all items of measurement), along with one or more sources of error. Therefore, the power of G theory is that it allows one to evaluate the extent to which generalizations might be made from the observed score of a student to the universe of observations that are confined to the factors or conditions measured in the G study (Rathvon, 2004).

The factors considered in this study include student, rater, and passage because researchers have found that these contribute to the difficulty in establishing instrument reliability (Ross, 2004). Oral reading accuracy scores in a running record may vary tremendously, according to which passage students read or how knowledgeable the rater is. Passage variance is attributable to two sources of variance: (a) linguistic structure and features of a text and (b) each student's variant background in relation to the meaningful content in the text. For example, the leveling procedure used to determine passage difficulty attempts to gauge linguistic factors that may influence a student's reading accuracy and running record score. Also, even though two passages may be labeled as similarly difficult in relation to the number and length of words and sentences, students may exhibit very different running record scores because of the extensiveness of their extant prior knowledge. Furthermore, the knowledge level of the rater may introduce a source of error into running record scoring. For example, a rater who has (a) advanced knowledge of reading development, (b) specific training on using running records, and (c) significant practice taking running records will most likely produce a more accurate score than will a teacher who has limited knowledge, training, or experience with running record scoring.

Teachers must ensure that obtained running record scores are an accurate and reliable estimate of a student's universe score. According to Freund and Wilson (1998), "A general principle in any data-collecting effort is to minimize the error variance, which will, in turn, provide for a higher power for hypothesis tests and narrower confidence intervals" (p. 362). Any time that an assessment is used as a standardized measure for making important instruction or diagnostic decisions, one must determine whether the scores generated by the tool or process are reliable. If teachers are not confident that the score obtained from a running record is a good approximation of the student's true score, then the score limits the usefulness of the measure in tracking reading performance. If the score is unreliable, then one may be able to determine if a source of error could be reduced by changing aspects of the process or instrument under the teacher's control. For example, when a high degree of variance exists between raters of a running record, the amount of error could possibly be reduced by increasing the number of raters and averaging their scores. If a major source of error is the passage being read, then increasing the number of passages read and using an average should produce a score, which is a more reliable estimate of the true score. The degree of



variance across error components is established through a G study.

Making decisions about how error or variance might be reduced is filtered out by conducting a D study. Whereas a G study allows one to pinpoint sources of measurement error, D studies incorporate those data to determine the optimal design. One can use the results of D studies to determine the optimum number of passages and raters that should be used to obtain a score that is the best estimate of a student's universe score. Although it may seem obvious that more passages or raters will produce a more reliable score, D study results indicate whether this is the case. The results of D studies also specify the degree of difference as the condition of each facet is changed. The power of any assessment is critically diminished if teachers cannot make dependable generalizations on the basis of behavioral observations. Knowledge of conditions or circumstances that influence a teacher's ability to make generalizations according to a score from an assessment is important. Oosterhof (1996) reported that a gauge of student performance, in this case a set of scores from a running record, will generalize only when it measures something with consistency.

#### Data Analysis

We used a three-way (Student  $\times$  Passage  $\times$  Rater) random effects ANOVA for the G study to compute estimates of the seven variance components. The components included student, rater, passage, Student  $\times$  Rater, Student  $\times$  Passage, Rater  $\times$  Passage, and residual interactions. We conducted D studies to consider whether we should use alternate designs to determine the optimum number of passages and raters needed to have reliable scores. Resulting data allow one to make absolute decisions based on the level of the student-observed score, without regard to the performance of others (Shavelson & Webb, 1991). We used the computer program GENOVA (Brennan, 1983) to perform the G and D studies. GENOVA is a Fortran 77 program designed for use in conducting generalizability analyses with balanced (equal  $n$ ) designs.

## Results

### G Study

A G study answered the first research question (The degree to which passages and raters contribute to discrepancies in student running record scores) by identifying the interactions present among readers, passages, and raters on running records. The magnitude of all seven variance components is reported in Table 2. Three variance components are large relative to the others. The three variance components are (a) students (53.4%), (b) Student  $\times$  Passage interaction (28.0%), and (c) residual (14.5%). However, we explain the results of all seven variance components.

*Variation among students' reading accuracy.* Students' running record accuracy was the object of measurement; thus, it represented the target population about whom potential users of running record scores intend to make inferences. We calculated the mean rating for each student by averaging that student's ratings across two passages and across 10 raters. The resulting mean rating for each student provided an estimate of that individual's universe or true score. The variance component for students provided an estimate of how much the unknown universe scores varied from one examinee to another examinee in the population of students about whom the test user wanted to make inferences. Ideally, the value of the student mean rating variance component should be large relative to the other sources of variability. Students' mean ratings provide an estimate of their reading ability, as assessed by similar passages and raters that could reasonably be considered to be comparable with the passages and raters that we included in this study.

The fact that the variance component between students reported in Table 2 was so large relative to the other variance components indicates that the 10 first-grade teachers or raters could reliably detect differences in the reading abilities of the individual students, regardless of which passage the student read or which rater rated the passage. We did not consider the variability among students' reading accuracy error variance because it was analogous to the true score variance in classical test theory. The larger the stu-

TABLE 2. Estimated Variance Components and Their Standard Errors

Source of variability	df	Variance component	SE	% (total variability)
Student	9	26.79	14.95	53.4
Rater	8	0.48	0.42	1.0
Passage	1	1.21	2.30	2.4
Student $\times$ Rater	72	0.36	0.89	0.7
Student $\times$ Passage	9	14.03	6.33	28.0
Rater $\times$ Passage	8	0.00	0.25	0.0
Residual	72	7.29	1.20	14.5
Total	179			100.00



dent variance component was in proportion to the total variability in the ratings, the better. If there was no error in the ratings, all the other variance components would have been zero, and the variance component for students would equal the total observed variance in the ratings.

*Student  $\times$  Passage interaction.* Because the variance component for students is considered as true variance, the largest source of error variability in the ratings analyzed in this study was the Student  $\times$  Passage interaction (see Figure 1). The fact that the two variables interacted indicated that the ranking of the students' mean ratings (averaged across all 10 raters) varied from one passage to another. In short, that outcome meant that a student's rated reading ability was affected by differences in passages that he or she was asked to read. Even when passages were identified as being at the same level of difficulty (i.e., Reading Recovery Level 14), subtle variations in the linguistic structure or conceptual content placed differing demands on the reader. The precise meaning of that error variance is difficult for one to determine with absolute certainty. Although both texts were at Reading Recovery Level 14 and were of the same type and structure, clearly this control did not eliminate passage variance, possibly because of topical differences between passages.

If no interaction had occurred between students and passages, then each student would have had the same rank order on both passages. Hence, if there was no interaction between students and passages, the lines connecting the pair of means for each student would be parallel (similar to the lines for Students 9 and 10). The crossing lines illustrate the inconsistencies in the students' relative standing on the two passages.

For example, Student 5 had the highest mean rating on passage 1 but only the third highest mean rating on Passage 2. Similarly, Student 7 had the eighth highest mean rating on passage 1 but the fourth highest mean rating on Passage 2. The number of crossing lines in Figure 1 is indicative of the degree to which a Student  $\times$  Passage interaction was present in the ratings. A larger proportion of crossing lines would have indicated that more of the students were ranked differently from the first passage to the second, and the variance component for the Student  $\times$  Passage interaction would have been larger. Conversely, if a larger proportion of the lines were parallel, more of the students would have had the same relative standing on both passages and the variance component for this interaction would have been smaller.

*Residual variance.* The residual variance component included the three-way Student  $\times$  Passage  $\times$  Rater interaction plus any variation in the ratings resulting from other unidentified sources not included in the two-facet design. In a three-way fully crossed, two-facet design there was no way to remove the confounding in the three-way interaction from the other unidentified sources, so one cannot determine whether the relatively large size of the residual variance was caused by the three-way interaction or other unidentified sources.

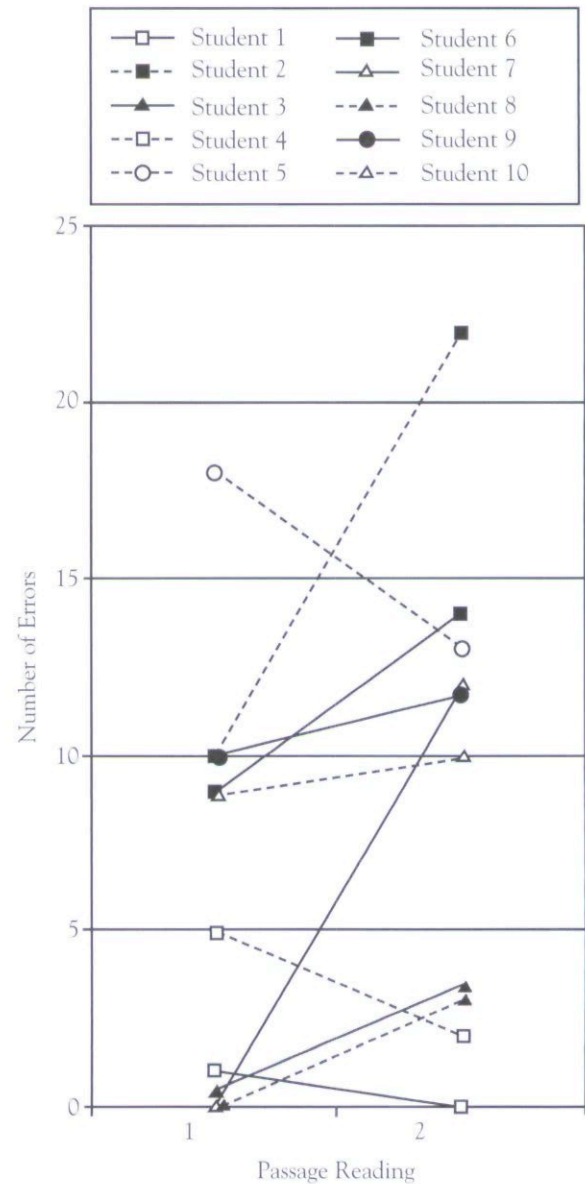


FIGURE 1. Average number of oral-reading errors by student on two passages.

*Variance caused by passages.* As shown in Table 2, the variance component for passages was relatively small (2.4%). That finding indicates that the variability in the overall student means as measured by teacher ratings on the two passages (averaged across all raters) was minimal, providing evidence that the two passages were approximately equally difficult for the group of students as a whole. However, the small main effect for passages obscures the fact that some students found that Passage 1 was more difficult, whereas other students believed that Passage 2 was more difficult, as evidenced by the relatively large Passage  $\times$  Student interaction. Again, the variance may have been caused, in part, by subtle differences in topics addressed in each of the two passages.



*Variance caused by raters.* The variance component for raters provided an estimate for the amount of variability in the mean ratings of the various raters averaged across all students and all passages. Table 2 shows that the estimated variance component for raters accounted for less than 1% of the total variability in the ratings. That finding indicates that the various raters were essentially interchangeable in the sense that a student being rated by only one rater would not make much difference as to which rater did the rating. All of the participating teachers had some experience with running records. In spite of that fact, the degree of teacher experience in using running records may have had a limited impact on the teachers' accuracy when scoring students' running records.

*Student  $\times$  Rater interaction.* The Student  $\times$  Rater interaction also accounted for less than 1% of the variability in the ratings. The negligible size of this interaction effect indicated that the students tended to be ranked in the same order by the various raters. The absence of a large Student  $\times$  Rater interaction is shown in Figure 1 by the relatively small proportion of lines that cross each other.

*Rater  $\times$  Passage interaction.* The smallest variance component reported in Table 2 was the Rater  $\times$  Passage interaction; the standard error for this estimate was smaller than for all other standard errors in this table. The small Rater  $\times$  Passage interaction indicates that raters rated students' readings of each passage consistently.

### D Study

A D study is used for making decisions about the way to reduce or filter out error or variance. We conducted D studies to address the two remaining research questions: (a) What number of passages and raters should one use when administering running records to minimize error variance and optimize the reliability of the resulting ratings? (b) What effect does using a nested design, in which students are crossed with passage and nested in raters, have on the reliability of running record scores?

That analysis was especially important given that some variance between running records was accounted for by passages and raters. D study results helped establish the running record testing conditions that best produced a stable or reliable student reading score. The G study results show clearly that simply providing students with text of the same level, type, and structure in a running record does not sufficiently eliminate error associated with passage. Also, having one teacher conduct and interpret student running records does not eliminate rater variance. Through the D study, we can identify optimal running record assessment conditions for reducing passage and rater variance. Those conditions include using at least two raters and three passages to obtain a student running record score.

We performed two separate D studies to address the issues. We based the first D study on the same fully crossed design that we used in the G study (Design 1). We con-

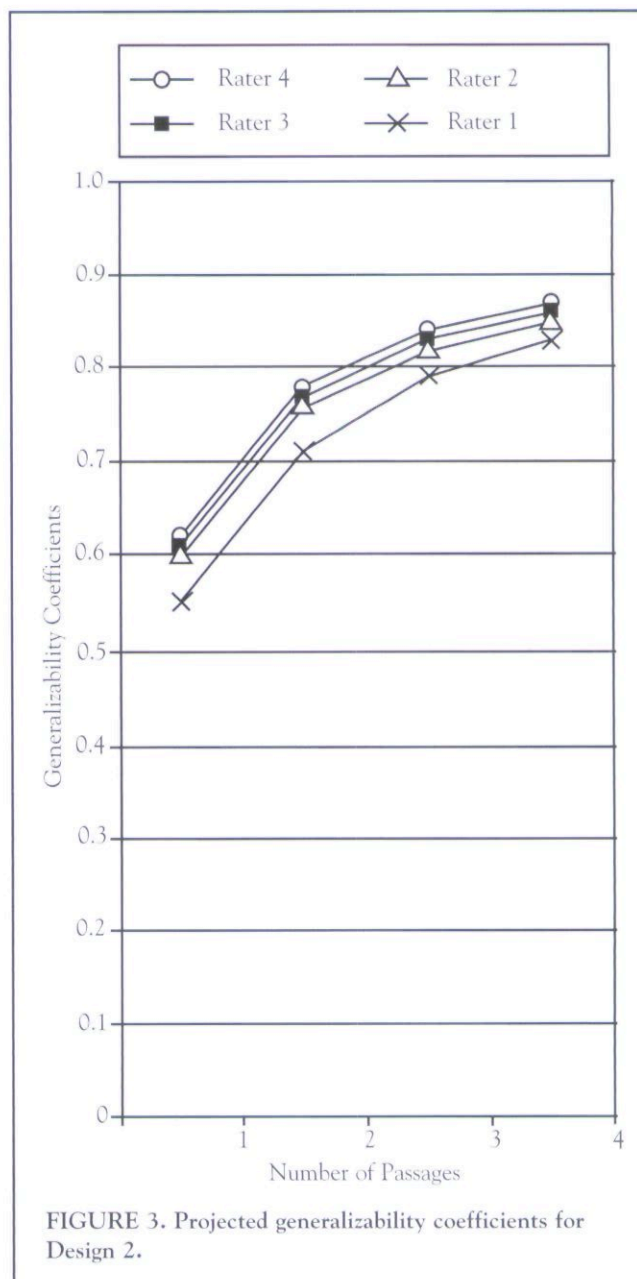
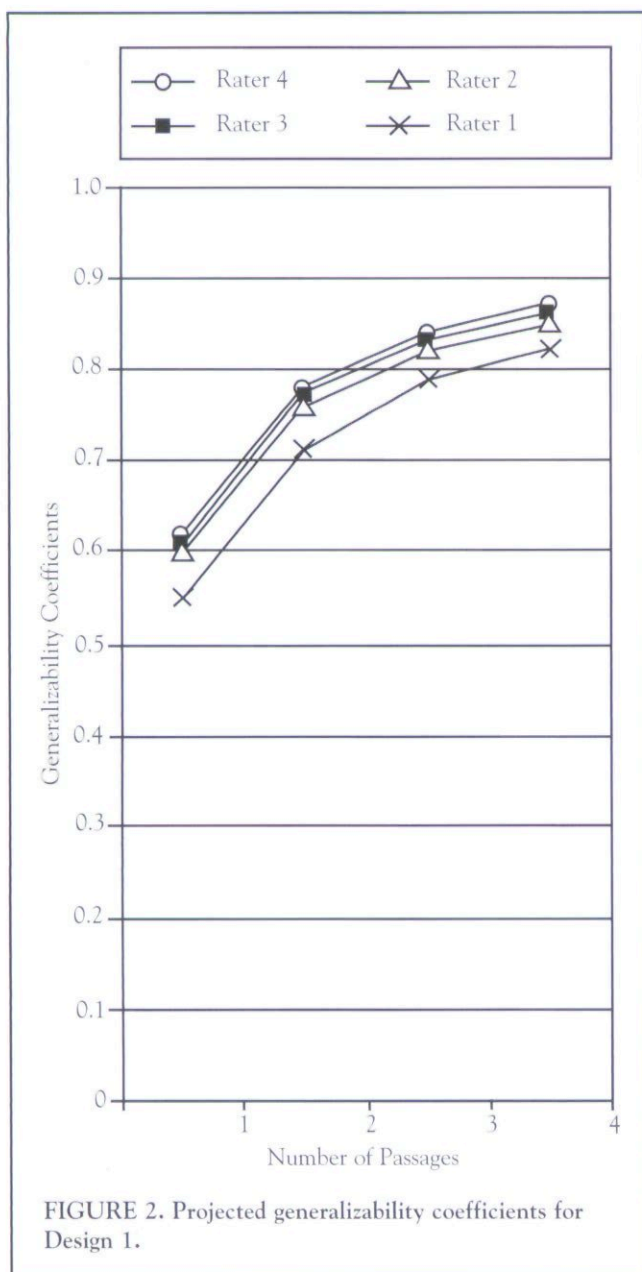
ducted the second D study to examine the consequences of using an alternate nested design (students crossed with passages and nested in raters) for collecting the ratings (Design 2). We used the variance component estimates from the original G study as input values in both D studies. Each D study produced estimated generalizability coefficients and error variances for relative and absolute decisions and for varying numbers of passages and raters.

*Generalizability coefficients.* The generalizability coefficient (G coefficient) provides an estimate of the reliability of using the mean rating obtained by an individual student (averaged across all passages and all raters) as a basis for making decisions about how that student's reading ability compares with the reading ability of other students. The generalizability coefficient describes the reliability of decisions about individual students' relative standing in a group. Therefore, the G coefficient is the reliability estimate that one should use if the obtained ratings are to be employed as a basis for making decisions about which students are better or poorer readers when compared with the mean of a relevant peer group. Decisions of this kind are called *relative decisions* in generalizability theory.

Figure 2 shows the predicted G coefficients that would likely result from varying the number of passages or the number of raters, or both, when one obtains the ratings from the design requiring that each student be rated by every rater on every passage (Design 1, fully crossed). Figure 3 shows the predicted G coefficients that would likely be obtained from using an alternate design in which each rater is expected to rate every student on some passages (Design 2, nested). The positive slope of the lines in Figures 2 and 3 shows the effect of increasing the number of passages that each student is asked to read and the number of raters. Figures 2 and Figure 3 reveal that increasing the number of passages increases the value of the G coefficient more than does increasing the number of raters, as shown in the slopes of each figure. Asking a student to read two passages instead of one passage increases the G coefficient more than does using four raters instead of one rater.

Another important finding is the similarity of the patterns in Figures 2 and 3. The slopes of the four lines and the differences in the elevation of the lines in the two graphs appear to be the same. Inspection of the numerical values of the reliability estimates for corresponding number of raters and passages in the two figures reveal that they are nearly equivalent in every case. When a single passage is used, the values are equivalent to more than five decimal places. However, when more than two passages are used, the predicted values are nearly identical, but they differ slightly in the second or third decimal place. The reliability estimates for Design 2 are typically a few thousandths larger than the corresponding estimates for Design 1 when more than one passage is used, but the differences are typically so inconsequential that one should ignore them for all practical purposes. More important is that Design 2, which is the more practical design, produces G coefficients





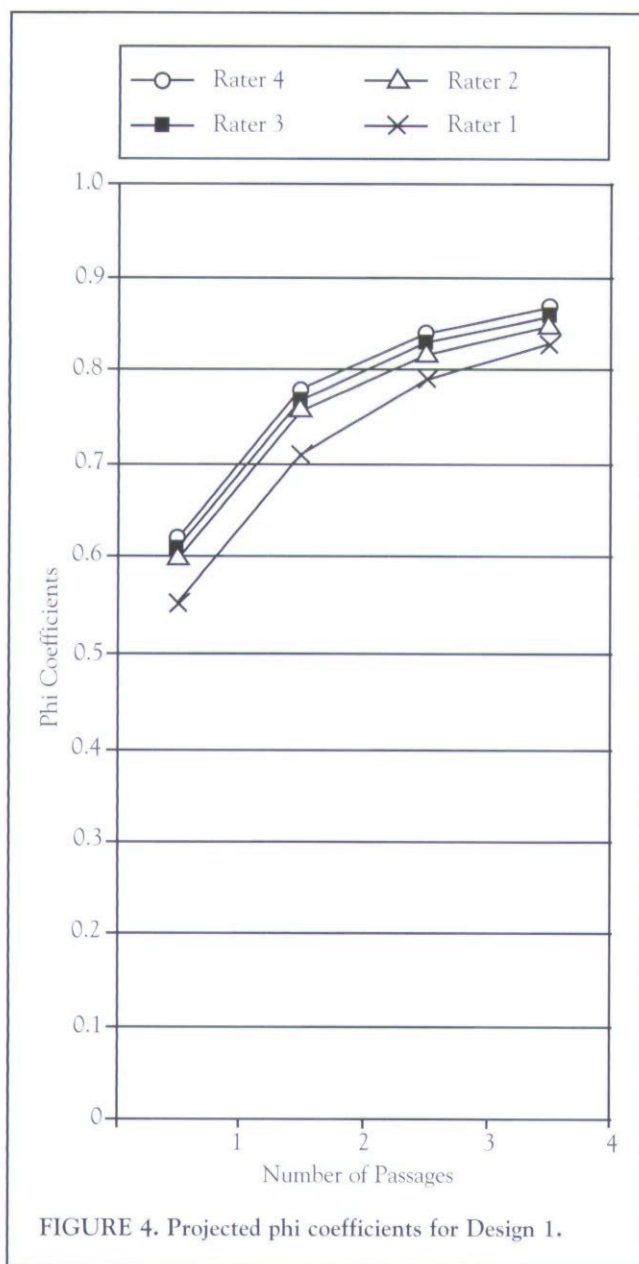
that are equally as good or better than are those obtained from Design 1.

*Phi coefficients.* Instead of assigning meaning to the ratings by comparing students with one another, the ratings obtained from running records can be used to make decisions about an individual student's reading ability compared with some pre-established level or cut score. As Rathvon (2004) suggests, the cut scores for running records are generally separated into three categories: (a) frustration level, below 89% accuracy; (b) instructional level, 90% to 95%; and (c) independent level, above 95%. Such decisions are called *absolute decisions* in generalizability theory because they describe a student's performance in comparison with the cut score, with no consideration for how the performance of an individual compares with the performance of other students. The phi coefficient is a statistic

produced by a D study, which presents the reliability of absolute decisions about individual students.

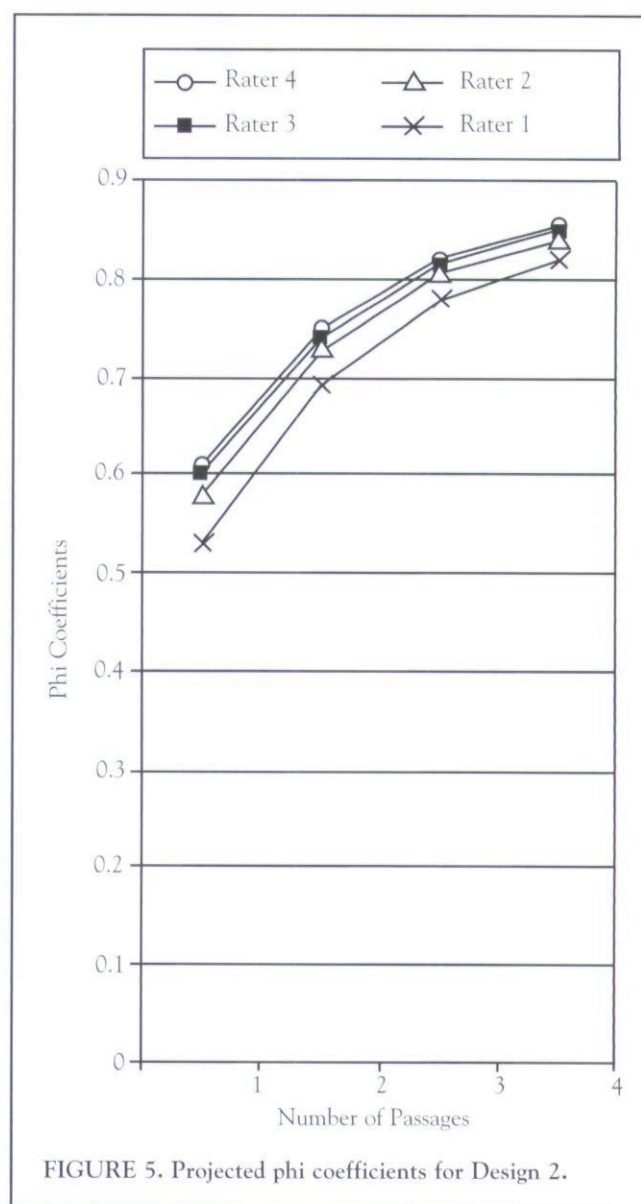
Figure 4 illustrates the predicted phi coefficients for various combinations of passages and raters when the fully crossed design is used to obtain the ratings (Design 1). Figure 5 shows the predicted phi coefficients for various numbers of passages and raters when the partially nested design is used where each rater rates all of the students but only randomly selected passages (Design 2). The phi coefficients in Figures 4 and 5 are smaller in every instance than the corresponding G coefficients in Figures 2 and 3, respectively. One should expect that finding because the phi coefficient accounts for all estimated sources of error in the ratings, whereas the G coefficients account for only the error sources that contribute to differences in students' relative standing within a group.





Similar to the findings with the G coefficients, the phi coefficients obtained from the two designs are essentially identical. The steepness of the slopes of the lines in Figures 4 and 5 indicates that using multiple passages is more efficient than using multiple raters to obtain reliable running record scores on which absolute decisions will be made.

*Standard error of measurement (SEM).* The size of the SEM indicates how much the test score obtained by an individual examinee would likely vary from one testing situation to another if that student were tested repeatedly. The SEM reported here was a derivative of the square root of the relative error variance. That measure of variability is particularly relevant in situations in which ratings are used as a basis for making relative decisions about individual students. The absolute error variance is a descriptive statistic in generalizability theory that one uses to describe the amount

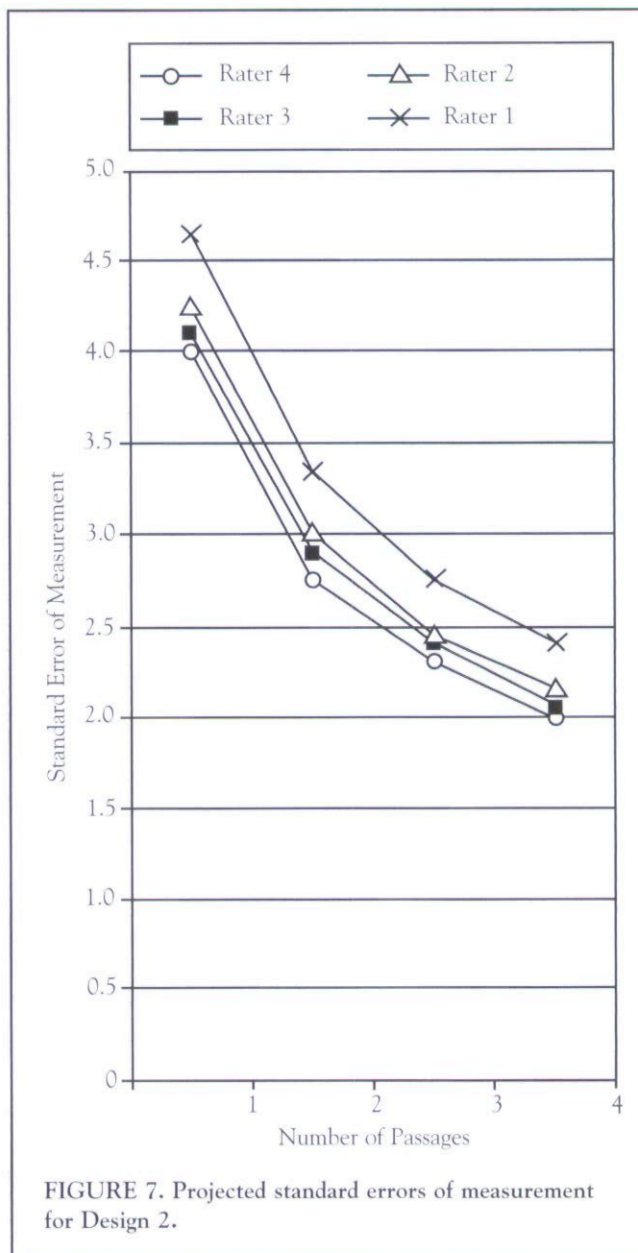
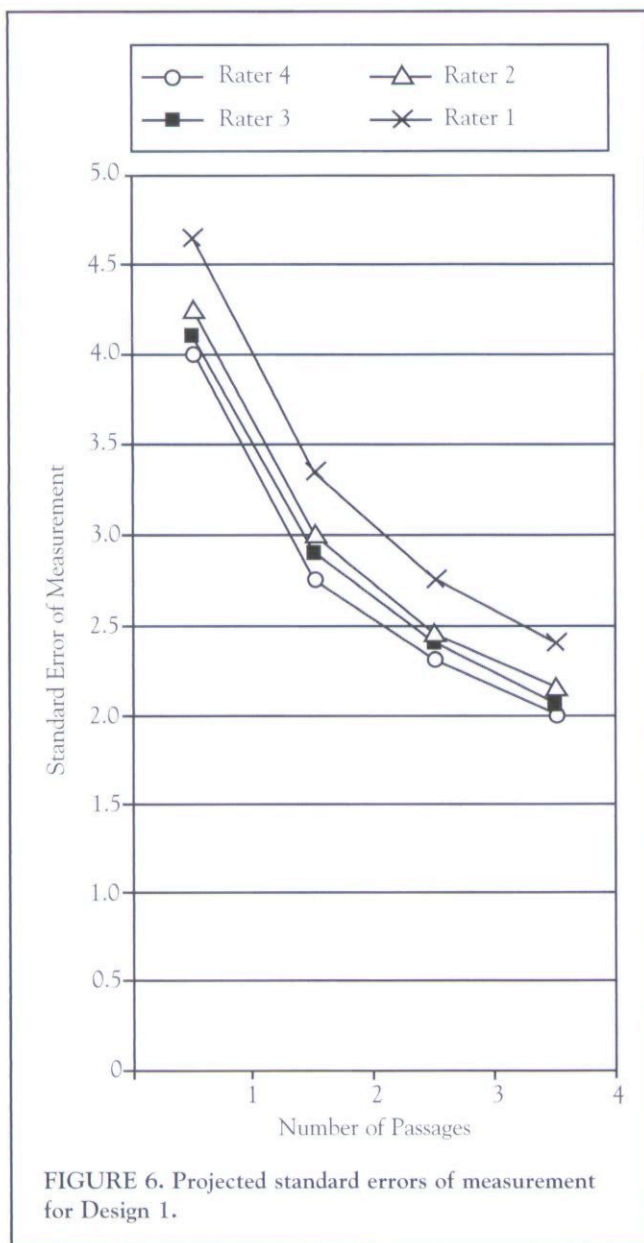


of error variance involved when ratings are used as a basis for making absolute decisions about students' ability.

Figures 6 and 7 show the estimated SEMs for varying numbers of passages and raters for Designs 1 and 2. Both graphs show how the size of the SEM varies as a function of the number of passages and raters. As the numbers of passages and raters increases, the SEM decreases, but increasing the number of passages has a greater effect on decreasing the error than does increasing the number of raters. The graphs for the two designs are essentially identical for the same reasons that the two designs have similar G coefficients.

Perhaps the most important message conveyed by Figures 6 and 7 is the magnitude of the SEM when only one passage is administered to a student and the student's performance on this passage is rated by a single rater. The SEM for this single-passage, single-rater situation is 4.65, as shown in Figures 6 and 7. When interpreting this number, the reader should keep in mind that the SEM is defined in





terms of the same units of measurement as the original scores on which it is based. In that case, the unit of measurement is the number of miscues or errors made by an individual student as he or she reads aloud. D study results indicated that at least three passages rated by two raters were necessary to make an absolute decision.

## Discussion

In education settings, teachers are expected to make instructional decisions regarding their students' abilities, progress, and needs by using reliable reading assessments. Because of federal and state legislation in the last few years, the demand for reliable and valid reading progress assessments has increased. In cases such as with running records, where assessment processes are widely used, it is especially important to establish scoring reliability. Making absolute

decisions with a running record requires the teacher to average student scores on at least three passages with at least two raters.

Our results indicate that the most limiting factor in rendering students' running record scores reliable is the number of passages used. That finding supports the contention of Ross (2004) that passage might exhibit a sizable source of error variance when scoring running records. Using a single score obtained from reading a single passage to portray that student's universe score would be highly questionable. Because the researchers controlled text variation of the passages by selecting texts of the same level, type, and structure, other between-passage differences may have contributed to text difficulty. One likely source of variation among the passages was the story topic or content in which vocabulary served as the proxy representation of topical differences.



Vocabulary difficulty caused by topical variation could be the factor that caused the running record scores to vary when controlling for text level, type, and structure. "Wagon" would be considered a Tier I word, whereas "porcupine" would be a Tier II word (Beck & McKeown, 1985). A first-grade reader might find a word such as porcupine more difficult to read than wagon. A mispronunciation of porcupine could likely translate to additional reading errors throughout the text. Also, the reader would have to know something about the concepts of a porcupine or a wagon to accurately understand the text. The central finding of this study is that despite topical variation between passages, the use of at least three passages from the same level, type, and structure can nevertheless produce a stable running record score.

A careful analysis of the data indicates that student scores varied considerably from passage to passage, directly influencing their relative standing or rank order. Student performance on each passage was not consistent, even though they read two passages identified at the same level of difficulty. That finding seems to call into question the popular Reading Recovery or A-Z leveling of text (Fountas & Pinnell, 1996). Hoffman, Roser, and Salas (2001) found that teachers using the Fountas and Pinnell leveling structure can reliably level text. However, when texts leveled in that manner assess student reading performance, they produce highly unreliable results. Running record scores that are acquired from a single-leveled text reading would not necessarily represent a student's true reading level.

One of the robust findings of this study is that given the sizable weakness of leveled passages in predicting student reading performance on a running record, reliable results can occur. That effect would require that the student read at least three similarly leveled passages. Students' error scores would then be averaged, producing a reasonably accurate representation of the true score.

Whereas the number of raters did not appear to account for much variation in the G study, the D study indicated that more reliable absolute decisions can be made by having at least two raters review each running record session. Oosterhof (1996) indicated that "Because the usefulness of assessments are significantly reduced if our observations fail to generalize beyond what we observe, it is important to be aware of the conditions that reduce generalizability" (p. 45).

Ross (2004) suggested that the rater could potentially be an additional source of sizable error in running records. That is an important finding because the raters possessed varying levels of experience with reading and running records. However, all of the raters had received some training in taking running records. If we had included raters with little training or specialized knowledge of taking running records, then a different result might have occurred. Although the variance components for raters and the interaction effect of raters crossed with students were relatively small, a difference existed between using one, two, or three raters in both designs. Improved coefficients resulted from including additional raters in the model. Given the practi-

cal constraints placed on teachers who may use running records, we suggest that they use at least three passages in running record assessment to obtain reliable results. That seems appropriate because most of the variance is accounted for by the passage. Given the small amount of variance accounted for by raters, teachers may choose not to add additional raters. In a practical context, it would seem to be easier for a teacher to use three passages without having to add raters as well.

### Limitations

There is undoubtedly some difference between performing a running record on a student sitting next to you and on one on television. For example, a teacher sitting next to a student during the assessment may inadvertently give verbal or nonverbal cues, which could either help or hinder a student's performance on an assessment.

We investigated student running record scores by using narrative text with students who were progressing normally in reading. We did not assess students with other text types or assess running record scores for students who were struggling readers. In addition, the topic of the text may have introduced a source of variation in students' performance on running record assessments. Thus, text type, structure, topic, and student background may influence the reliability of a running record score.

Another limitation is that students may perform better or worse, depending on their relationship with the individual(s) giving the test or on the way that they respond to the presence of a camera. In a typical classroom setting in the region in which we conducted the study, classroom teachers, rather than individuals who the students had never met, performed the running records.

One major limitation in this study was that only 10 students and 10 raters participated. Using a small sample size cannot be generalized to all texts, students, and teachers. Additional replication of this research with a larger number of randomly selected participants and texts would expand the generalizability of these findings.

### Instructional Implications

Given the popularity of running records, we tentatively made several important instructional implications from this study. Teachers should be very cautious when they determine instructional placement of students into reading groups with a *single* running record score. Although accurate scoring of running records can identify some sources of between- and within-student differences, single scores are not as sensitive to between-text variations, such as level, structure, type, and topic. All of those text characteristics can contribute to running record score variability. Between-text topical differences may also contribute to inaccurate instructional group assignment for students. For example, a student may have a good grasp of basic reading



skills, but when confronted with a text that uses a difficult or unfamiliar vocabulary word, a false score would result. The teacher may be interpreting that the child struggles with word recognition, when the reader is in fact requiring vocabulary support.

Teachers should recognize that traditional text-leveling procedures do not fully account for all factors that affect the difficulty of a text. Even when controlling for text level, type, and structure, there are still naturally occurring topical variations between texts that will render one more difficult than another. However, a teacher can accommodate topical differences by taking running records with at least three passages and averaging the three scores. A key finding is that using the average score from the three running records can produce a reliable student score. Also, students should be carefully matched to text. Teachers should account for the topical variation between texts and ensure that students are scaffolded into the topic of texts. If limited topical knowledge is present, the teacher should help the child understand the topic of the text. When children are reading, their success will be strongly affected by the material that they are reading. Teachers must be highly involved in helping students understand what they will read to increase the likelihood that the child will be a confident reader.

## Conclusions

Our results indicate that under certain conditions, running records can produce scores that closely approximate a student's universe or true score. The scores generated through the use of running records by teachers with varying degrees of experience and expertise are reliable when at least three passages are administered and scores are averaged. Although there is a large difference between using one and two passages versus two and three passages, the difference between using three and four passages is less substantial.

The finding that students' scores may vary considerably depending on which passage they read does not bode well for the use of Reading Recovery or A-Z leveling structures because text levels may be confounded by failure to consider topical differences that contribute to text difficulty. Those techniques do not appear to produce reliable text levels, which makes it difficult for one to predict student reading performance.

Given the wide use of running record assessments and the national focus on use of reliable assessment tools, our data provide some utility. We suggest that teachers consider these data when they administer running records to increase the reliability of this high-utility assessment tool.

## REFERENCES

- Bean, R. M., Cassidy, J., Grunert, J. E., Shelton, D. S., & Wallis, S. R. (2002). What do reading specialists do? Results from a national survey. *The Reading Teacher*, 55, 736-744.
- Beck, I. L., & McKeown, M. G. (1985). Teaching vocabulary: Making the instruction fit the goal. *Educational Perspectives*, 23, 11-15.
- Brennan, R. L. (1983). *Elements of generalizability theory*. Iowa City, IA: American College Testing Program.
- Brennan, R. L. (1992). The context of context effects. *Applied Measurement in Education*, 5, 225-264.
- Chapman, J. W., Tunmer, W. E., & Prochnow, J. E. (2001). Does success in the Reading Recovery program depend on developing proficiency in phonological-processing skills? A longitudinal study in a whole language instructional context. *Scientific Studies of Reading*, 5, 141-176.
- Clay, M. M. (1966). *Emergent reading behaviour*. Unpublished doctoral dissertation, University of Auckland Library, New Zealand.
- Clay, M. M. (1993). *An observation survey of early literacy achievement*. Portsmouth, NH: Heinemann.
- Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Orlando, FL: Holt, Rinehart and Winston.
- Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. New York: Wiley.
- Eason, S. (1991). Why generalizability theory yields better results than classical test theory: A primer with concrete examples. In B. Thompson (Ed.), *Advances in educational research: Substantive findings, methodological developments* (pp. 83-98). Greenwich, CT: JAI.
- Fountas, I. C., & Pinnell, G. S. (1996). *Guided reading: Good first teaching for all children*. Portsmouth, NH: Heinemann.
- Freund, R. J., & Wilson, W. J. (1998). *Regression analysis: Statistical modeling of a response variable*. San Diego, CA: Academic Press.
- George the Porcupine. (1997). Glenview, IL: Scott Foresman.
- Glissmeyer, C. B. (1999). *Oral retelling as a measure of reading comprehension: The generalizability of ratings of college-aged second language learners reading expository text*. Unpublished doctoral dissertation, Brigham Young University, Provo, UT.
- Hintze, J. M., Owen, S. V., Shapiro, E. S., & Daly, E. J. (2000). Generalizability of oral reading fluency measures: Application of G theory to curriculum-based measurement. *School Psychology Quarterly*, 15, 52-68.
- Hintze, J. M., & Peritte, H. A. P. (2001). The generalizability of CBM oral reading fluency measures across general and special education. *Journal of Psychoeducational Assessment*, 19, 158-170.
- Hoffman, J. V. (1991). Teacher and school effects in learning to read. In R. Barr, M. L. Kamil, P. B. Mosenthal, & P. D. Pearson (Eds.), *Handbook of reading research* (Vol. 2, pp. 911-950). New York: Longman.
- Hoffman, J. V., Roser, N., & Salas, R. (2001). Text leveling and "little books" in first-grade reading. *Journal of Literacy Research*, 33, 507-528.
- Huck, S. W. (2004). *Reading statistics and research*. Boston: Allyn & Bacon.
- Mandler, J. M., & Johnson, N. S. (1977). Remembrance of things parsed: Story structure and recall. *Cognitive Psychology*, 9, 111-151.
- Matsumura, L. C., Patthey-Chavez, G. G., Valdes, R., & Garnier, H. (2002). Teacher feedback, writing assignment quality, and third-grade students' revision in lower- and higher-achieving urban schools. *Elementary School Journal*, 103, 3-26.
- Nation, K., & Snowling, M. (1998). Individual differences in contextual facilitation: Evidence from dyslexia and poor reading comprehension. *Child Development*, 69, 996-1011.
- Oosterhof, A. (1996). *Developing and using classroom assessments*. Englewood Cliffs, NJ: Merrill.
- Pressley, M., Wharton-McDonald, R., Allington, R., Block, C. C., Morrow, L., Tracey, D., et al. (2001). A study of effective first-grade literacy instruction. *Scientific Studies of Reading*, 5(1), 35-58.
- Rathvon, N. R. (2004). *Early reading assessment: A practitioner's handbook*. New York: Guilford.
- Reynolds, C. R. (1990). Conceptual and technical problems in learning disability diagnosis. In C. R. Reynolds & R. W. Kamphaus (Eds.), *Handbook of psychological and educational assessment of children: Intelligence and achievement* (pp. 571-592). New York: Guilford Press.
- Ross, J. A. (2004). Effects of running records assessment on early literacy achievement. *The Journal of Educational Research*, 97, 186-194.
- Share, D. L., & Stanovich, K. E. (1995). Cognitive processes in early reading development: Accommodating individual differences into a model of acquisition. *Issues in Education*, 1, 1-57.
- Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer*. Newbury Park, CA: Sage.
- Snow, C. E., Burns, M. S., & Griffin, P. (1998). *Preventing reading difficulties in young children*. Washington, DC: National Academy Press.
- Taylor, B. M., Pearson, P. D., Clark, K., & Walpole, S. (2000). Effective school and accomplished teachers: Lessons about primary-grade reading

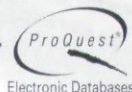


- instruction in low-income schools. *Elementary School Journal*, 101, 121-166.
- Taylor, B. M., Pearson, P. D., Peterson, D. S., & Rodriguez, M. C. (2005). The CIERA school change framework: An evidence-based approach to professional development and school reading improvement. *Reading Research Quarterly*, 40(1), 40-69.
- Thorndyke, P. W. (1977). Cognitive structures in comprehension and memory of narrative discourse. *Cognitive Psychology*, 9, 77-110.
- Traub, R. E., & Rowley, G. L. (1991). Understanding reliability. *Educational Measurement: Issues and Practice*, 10, 37-45.
- Turner, W. E., & Hoover, W. A. (1992). Cognitive and linguistic factors in learning to read. In P. B. Gough, L. C. Ehri, & R. Treiman (Eds.), *Reading acquisition* (pp. 175-214). Hillsdale, NJ: Erlbaum.
- The Wagon*. (1997). Glenview, IL: Scott Foresman.
- Wharton-McDonald, R., Pressley, M., Rankin, J., Mistretta, J., Yokoi, L., & Ettenberger, S. (1997). Effective primary-grades literacy instruction = balanced literacy instruction. *The Reading Teacher*, 50, 518-521.
- Wolfersberger, M. E., Reutzel, D. R., Sudweeks, R., & Fawson, P. C. (2004). Developing and validating the classroom literacy environmental profile (CLEP): A tool for examining the "print richness" of early childhood and elementary classrooms. *Journal of Literacy Research*, 36(1) 83-144.
- Wray, D., Medwell, J., Fox, R., & Poulson, L. (2000). The teaching practices of effective teachers of literacy. *Educational Review*, 52(1), 75-84.

## How is this publication thinking about the future?

**By becoming part of the past.** This publication is available from ProQuest Information and Learning in one or more of the following ways:

- **Online, via the ProQuest® information service**
- **Microform**
- **CD-ROM**
- **Via database licensing**



For more information, call  
**1-800-521-0600, ext. 2888 (US) or 01-734-761-4700 (International)**  
[www.il.proquest.com](http://www.il.proquest.com)

From: ProQuest  
COMPANY



Copyright of Journal of Educational Research is the property of Heldref Publications and its content may not be copied or emailed to multiple sites or posted to a listserv without the copyright holder's express written permission. However, users may print, download, or email articles for individual use.