Credibly assessing reading and writing abilities for both elementary student and program assessment

John Burgin, Gail D. Hughes *

Department of Educational Leadership, DKSN 419, 2801 S. University Avenue University of Arkansas - Little Rock, Little Rock, AR 72204-1099, USA
Available online 6 February 2009

Abstract

The authors explored the credibility of using informal reading inventories and writing samples for 138 students (K–4) to evaluate the effectiveness of a summer literacy program. Running Records (a measure of a child’s reading level) and teacher experience during daily reading instruction were used to estimate the reliability of the more formal Developmental Reading Assessment scores. Training of scorers was used to increase the reliability of writing scores; a second scoring was used to estimate the reliability of the scores. The results suggested that with minimal modifications to administration and scoring procedures, scores from both reading inventories and writing samples can be a dependable source of data for teachers, administrators, and policy makers. This result is significant because it suggests that formative literacy assessments can be reliably used instead of standardized multiple-choice tests to make more credible summative decisions without taking time away from instruction, and can truly match curriculum, instruction, and assessment. © 2009 Published by Elsevier Ltd.

Keywords: Assessing reading and writing; Running Records; Multiple-choice tests; No Child Left Behind legislation

America’s No Child Left Behind (NCLB) legislation requires each state to create a plan to ensure that all children are reading on “grade level” by the year 2014. The federal government must approve each state’s goals, or benchmarks, and the means they use to assess progress toward the benchmarks by fourth-grade. The federal government’s approval requires that states use standardized multiple-choice tests. Schools that do not make “Adequate Yearly Progress” on these tests for two years must offer “vouchers” to attend another school (if one is available). Schools that do not make AYP for five consecutive years face closure, being taken over by the state, and/or the reassignment of teachers and administrators (see Sanders (2008) for legal commentary). Many educators believe that the high-stakes testing requirements of NCLB represent a simplistic solution

* Corresponding author. Tel.: +1 501 569 8944; fax: +1 501 569 3547.
E-mail address: gdhughes@ualr.edu (G.D. Hughes).

1075-2935/$ – see front matter © 2009 Published by Elsevier Ltd.
to a complex problem: improving literacy instruction and student achievement. Yet high-stakes testing is not unique to the U.S.; usage and consequences of such tests are broader areas of concern for the entire educational community.

Educators often refer to the social sciences as the “soft sciences,” however, both researchers and practitioners can attest to the fact that the social sciences are often the “hardest sciences” of all since the object of measurement is often an abstract construct rather than something that we can easily measure with a ruler or scale. Nevertheless, data driven decisions require measurement of these constructs and educators must continue to refine our measurement abilities. Unfortunately, the pressures of “high-stakes” testing have caused local school districts to adopt a “kitchen sink” approach to assessment that takes precious time away from instruction and does not generate results that help teachers plan instruction. Therefore, it is important for stakeholders to periodically re-examine the number and quality of the assessments in use.

The purpose of this study was to generate assessment results that could be used to evaluate a kindergarten through fourth-grade summer literacy program; plan instruction; and explore the possibility that informal reading and writing assessment results alone could be used to make more credible summative decisions in a timely manner. The purpose was not to examine the validity of decisions made from test scores in the strict quantitative sense; rather to explore the credibility of decisions made from scores in the more qualitative sense of believability, plausibility, and authenticity. The purpose of the summer literacy program was to support the participating school’s efforts to make AYP toward the goals required by the federal government’s NCLB legislation.

1. The problem

In the United States, most schools are not year-round and 100% of the 100 largest school districts in the United States have some type of summer school program to fill the void (Borman, 2001). Unfortunately, measuring success is difficult and comparing the relative effectiveness of different programs is even more complex. If the definition of success is narrowed to student gains in reading and writing, what is the best way to measure those abilities?

The results of multiple-choice tests cannot provide the depth of feedback that teachers need to improve instruction, nor that administrators need to evaluate program effectiveness because a score on a sub-test does not directly describe nor adequately measure literacy; it merely denotes test-taking behaviors that are correlates of literacy (Farr & Carey, 1986). Furthermore, the validity of the conclusions drawn from such data is confounded by the significant impact of low SES on standardized multiple-choice test results (Goodwin & Goodwin, 1997; Patton, 1990).

Unlike the results of multiple-choice tests, the performance data generated by informal reading and writing inventories can be used to design individualized instruction that is based on students’ actual literacy behaviors, and it provides more credible results (Farr & Carey, 1986; Patton, 1990; Wortham, 2005). Historically, the problem with using informal reading and writing inventories is the inability of scorers to consistently agree on scores for the same sample – reliability (Huot & Neal, 2006).

1.1. The complexity of measuring reading and writing

In the United States, achievement testing is a growing industry that has emerged as a “cornerstone of educational practice and policy making” (Paris & Hoffman, 2004, p. 199). Achievement testing is dominated by the belief that knowledge can be reduced and measured. Multiple-choice tests gained popularity because they were less costly and less time consuming than more direct methods like scoring writing samples (Linn & Gronlund, 2000), and are “believed to be objective
and fair because they were seen as independent of the bias of the human decisions of individual teachers” (Williamson, 1994, p. 151). However, should efficiency be the ultimate criterion? The fact that a measure is free from bias does not ensure that decisions made from that measure have a high degree of validity (Huot & Neal, 2006).

The problem with validity is further compounded by the format of most objective tests. “Before age 8, standardized achievement measures are not sufficiently accurate to be used for high-stakes decisions. Therefore, high-stakes assessments intended for accountability purposes should be delayed until the end of third grade or preferably fourth grade.” Shepard, Kagan, and Wurtz, 1998, pp. 20–21).

Ideally, literacy assessments should look like good instruction (Farr & Tone, 1994). Rather than striving for efficiency, we should strive to enhance performance and encourage learning with assessments. Not only do students learn from the process of test preparation, they learn the larger lesson of what is important by what they are asked to do on the test. If testing is what one does apart from teaching, then validity and educative power will always be sacrificed to reliability (Wiggins, 1994).

Kohn (2000) distilled the issue “we measure most what matters least.” While standardized tests have some utility for administrators and policy makers in informing them about how their students compare with other students, scores from norm-referenced tests have little validity in informing a teacher about what their students know and can do. Instead, measures such as informal reading inventories (IRIs) and writing samples should be considered. The larger question is whether the numerical results from informal literacy assessments can be used to credibly aggregate data across students, classrooms, and schools to provide administrators and policy makers with the data they need for decision-making.

1.2. Reliability of informal literacy assessments

If the ultimate purpose of the assessment is to determine how well a student can read or write, whether the concern is for a single student or to aggregated across classrooms and schools, the most valid assessment would be one that actually measures how well a child can actually read and write. Nonetheless, reliability is a necessary condition for validity. Studies exploring the reliability of scores from informal literacy inventories like the DRA and scored writing samples have mixed results.

By the 1990s, the paradigm had shifted away from multiple-choice tests to measure writing ability. Direct writing assessments were believed to have “demonstrated the kinds of technical capabilities needed for large-scale assessment” (Gearhart, Herman, Novak, & Wolf, 1995, p. 207), and direct writing assessments were added to state’s benchmark exams (Farr, 1990), the Standard Achievement Test (SAT), and the Graduate Record Exam. Critics countered that the more student writing assessments are controlled for testing purposes, the less they resemble the natural writing process (Moss, 1994; Wiggins, 1994). At the classroom level, Mann (1988) found inter-rater reliability to be so low that he advised avoiding analytic scoring unless adequate time was taken for extensive training on the categories of the rubric.

The reviews of IRIs are also mixed. In the 1990s, a variety of new reading assessments were developed for both formative and summative functions. A majority of these IRIs depend upon the administrator of the test to create a “Running Record” (Clay, 1993). As a child reads a portion of a text, the teacher marks word substitutions, self-corrections, repetitions, omissions, and words “told” to the student. An accuracy rate is then computed to guide future book selection, and an analysis of the pattern of errors or “miscues” is completed to plan instruction.
Following reading, the child is asked to retell the story. During more formal IRIs, like the DRA (Beaver, 1997), a retelling score is generated during a formal interview. The child’s accuracy rate, retelling score, and a description of the patterns of miscues are then compared to the reading rubric to identify the child’s “instructional reading level” on a scale of 1–44.

Paris and Hoffman (2004) believe that not only are IRIs effective, but they could also “provide multiple indicators of children’s oral reading, including rate, accuracy, prosody, retelling, and comprehension” (p. 207). In addition to providing multiple indicators of performance, IRIs can be utilized by multiple audiences. IRIs are “appropriate tools for classroom teachers,” and for administrators and policy makers to determine the effectiveness of instructional programs in producing growth over time compared with other strategies (Davidson & Myhre, 2000, p. 26).

Critics of IRIs like the DRA focus on their vulnerability to inter-rater reliability issues (Rathvon, 2004). These issues arise principally because administration of the test requires: (a) expertise in choosing the appropriate text to begin testing, (b) familiarity with the standardized reading passage’s content to score a child’s retelling, (c) the time and effort to master the miscue scoring system, and (d) teachers to appropriately prompt students while they read and during the retelling interview – an ability that varies considerably.

1.3. Conclusion

Measuring such abstract criteria as how well a student can read and write remains a challenge. However, the literature reveals a growing support for direct measures of reading and writing abilities for assessment (Davidson & Myhre, 2000; Gearhart et al., 1995; Paris & Hoffman, 2004). While reliability is not the ultimate criterion, overcoming the problems in this area would enhance validity. The extension of construct validity to encompass further “relevance and utility, value implications, and social consequences of the assessment” (Gearhart et al., 1995, p. 236) encourages researchers to consider also what assessments can best serve students, teachers, administrators, and policy makers not only in their decision making, but also in a way that does not disrupt learning. Therefore, studies are needed to address these critical issues of reliability and validity of data from direct assessment.

With a specific summer program in mind, the questions that this study sought to answer were: (a) How can reading and writing abilities be measured so that the data provide high enough inter-rater reliability and validity to be used for program evaluation, and are rich enough so that teachers can use the results to inform instruction? and (b) How might we administer assessments so that the normal educational process is undisturbed?

2. Method

2.1. Pilot study

In addition to the norm-referenced multiple-choice tests and the state’s benchmark tests NCLB required, the state’s plan called for classroom teachers to administer a battery of informal assessments in the fall and spring. During the first year of the summer program, the results of the classroom teacher’s spring administration of the Developmental Reading Inventory (DRA; Beaver, 1997) were used to make summer reading groups. In comparison to the informal reading inventories (IRI) summer instructors collected, the school’s DRA scores were much higher. The summer teachers believed that classroom teacher’s subconscious biases (prior knowledge, high-stakes pressures) inflated the DRA scores.
Conversely, the writing samples the schools provided were a more credible source of information – more credible in that they are authentic and more concrete in that writing samples can be independently evaluated at a later date. Following a two-month pilot summer session, teachers believed that students’ reading ability increased, but not dramatically. Interestingly, instructors thought that students’ writing improvement was more significant than their reading gains. Without reliable data, a control group, or fall scores, the program could not be credibly evaluated. In order to justify the expense of the program, better data were needed. This research project began the following summer. The methods used to improve the reliability of the DRA and writing scores form the basis of this article.

2.2. Participants

This convenience sample consisted of student scores from a summer literacy program that serves two large elementary schools annually. The schools were comparable schools in the same high poverty neighborhood with 92% of students receiving free or reduced lunch, and both schools having been awarded “Reading First” grants from the Arkansas Department of Education using NCLB funds. Consequently, both schools had a well-trained literacy coach, an excellent set of books for each classroom, and an uninterrupted literacy block each morning. The student body at each school was approximately 78% African-American, 20% Hispanic, and 2% White/Other.

All students in the participating schools were invited to attend the summer program. Seventy students who attended summer school 70% of the time were matched to 70 students who did not attend but returned to the same school in the fall.

In an effort to control for variables other than summer instruction, control group students were matched by: home school, grade level, DRA scores, writing scores, primary language, and gender. The final sample size was 138 students with 70 in the “treatment group,” and 68 in the “control group” (clerical errors resulting in missing data excluded two control group members). Answers to the research questions addressed in this paper do not require the distinction between the “treatment” and “control” groups; however, the authors used these terms to convey the original nature of the data.

2.3. Instrumentation

The district used the Developmental Reading Assessment (Beaver, 1997) to make summative decisions at the end of each year, and Running Records (Clay, 1993) as a day-to-day formative assessment. Both assessments are based on the same practice of marking errors while students read, then figuring an accuracy rate, and analyzing the types and qualities of students’ errors or “miscues.” The DRA differs from Running Records in that administration includes a formal interview about the content of the text read, and a scoring rubric that uses both the descriptions of students’ miscues and the results of the interview to generate a “reading level.” Although, teachers taking Running Records briefly ask students about what happened in the story, the DRA interview is much more rigorous, and scoring is much more time consuming.

Following the district’s protocol, the writing samples collected were based on one piece of writing; written in one sitting; in response to a grade-level prompt. Rather than disrupt the daily routine like the administration of a standardized norm-reference test, the authors had teachers administer the test during the time in the daily schedule that children normally had “writer’s workshop.” Kindergarten through second-grade students had 30 min; second- and third-graders had 45 min; however, these limits were unnecessary as students finished writing before the time
limit was reached. To ensure anonymity during scoring, students wrote their names on the back of their examination paper rather than on the front.

The writing prompt for each grade-level and the directions were similar in format and structure. Students were allowed to have a copy of the prompt on their desk and it served as a dictionary of sorts. For example, in first-grade the prompt provided to students was in the following format:

Think about your favorite animal. It may be a pet, an animal at the zoo, or an animal on a farm. Before you write, think about what your favorite animal looks like. Why is it your favorite animal? Write about your favorite animal. Be sure to describe it and give details that explain why this animal is your favorite one.

The school district had a different writing rubric for each grade and provided anchor papers to illustrate the criteria within each rubric. The kindergarten rubric was on a 5-point scale (Heenman, 1985), “1” = Scribble Stage, “2” = Isolated Letter, “3” = Transitional, “4” = Stylized Writing, and “5” = Writing. First- and second-grades’ rubric mirrored the state’s benchmark score categories of “1” = Below Basic, “2” = Basic, “3” = Proficient, and “4” = Advanced. The third category “proficient” had a checklist of several writing behaviors in each of five categories: Content, Style, Organization, Language/Word Choice, and Sentence Structure. For example, on the first-grade rubric, under the descriptor “Organization,” raters had two boxes that could be checked: “Student has a clear beginning, middle, and end,” and “The events are in a logical order.” Raters then added the number of checked boxes to determine the overall score.

Scores for the third and fourth-grade students were based on a rubric that was similar to the one used by the state department of education to score the fourth-grade benchmark exam. Samples were scored across six domains: Content, Style, Sentence Formation, Usage, Mechanics, and Spelling. Scores ranged from “1” = Little or No Control, “2” = Inconsistent Control, “3” = Reasonable Control, and “4” = Consistent Control. Each domain had detailed quality descriptors to guide the scoring process across the four quality point values. For example, under the domain “Content” four areas were listed: Central idea, Elaboration, Unity, and Organization. Attributes of each area were described across the rubric in terms of the amount of control “1, 2, 3, or 4.”

2.4. Procedures

In the summer literacy camp, kindergarten through fourth-grade students attended school from 8:30 am until 11:30 am Monday through Friday for one month (20 contact days). Instruction focused exclusively on literacy using a curriculum based on Linda Dorn’s conception of a “comprehensive literacy model” (Dorn & Soffos, 2001) that forms the basis of the Reading First initiative in Arkansas. In general, classrooms were organized by grade level, but students were also sorted by reading level to balance the number of guided reading groups in each classroom. Graduate students taking a reading practicum course and classroom teachers from the two host schools staffed summer classrooms. Each classroom had two teachers and an average of 10 kindergarten through fourth-grade students.

Teachers in the host schools had been collecting writing samples and DRA scores for several years. Rather than supplant the school’s good faith efforts to collect high quality data, the authors took a less obtrusive approach, and looked for ways to facilitate the school’s efforts and assist the teachers with the task. Had the authors adopted a best practices approach to rubric development and scoring (Moskal & Leydens, 2000), the school’s routine would have been disrupted, and the data would not have represented students’ daily behavior. Instead, this study only used data that the schools were already collecting, and the rubrics and protocols the local district provided.
During the pilot summer session, instructors were concerned that the DRAs were not administered properly and questioned the credibility of the DRA scores provided by the schools. As a result of the pilot study, two changes were made to the schools’ assessment practices in scoring the DRAs. First, to avoid personal biases that can occur when assessing one’s own classroom, classroom teachers did not administer the DRA to their own students. Instead, teachers traded classrooms and administered the DRA to students from other classrooms. Second, the DRA scores were spot-checked by Reading Recovery teachers using one or two children from each classroom.

At the beginning of the summer program, graduate students collected Running Records (Clay, 1993) from all the children attending summer school. Running Records (RRs) were not collected for students in the control group. Although RRs are a much more informal method of assessment, these scores were used to estimate the credibility of the spring DRA scores. By using the two sets of reading scores to calculate reliability, the authors were able to estimate the credibility of DRA scores; generate data that reflected the quality of matches between summer participants and the control group; and address the issue of using DRA scores as a summative assessment.

In an effort to maximize the credibility of writing scores, the authors provided the teachers with in-service training. Sessions were held prior to administering the writing assessment to assure that the school district’s testing protocols were followed, and again at the beginning of scoring to calibrate the use of the district’s writing rubric. The administration of the writing prompt and collection of writing samples were monitored by the authors and Title I reading teachers.

To increase both the reliability and validity of the writing scores, all writing samples were scored by grade-level teams who discussed differences in writing samples and reached a consensus when questions arose rather than receiving a single score from a person working alone. The first scoring of each writing sample was conducted within each of the host schools. Next, the two participating schools traded writing samples and scored each sample a second time. The scores for each grading of a writing sample were averaged to provide a measure of students’ writing abilities at the beginning of the summer program.

3. Results

3.1. Reliability

RRs rather than a formal DRA test were used by summer school teachers to estimate the reliability of DRA scores. The administration of the DRA would have taken an exorbitant amount of time during the first week of summer school, time that would have significantly taken away from instructional contact hours. An experienced teacher can take a Running Record in less than five minutes. The DRA often takes 30 min or more to administer and score the results. Both instruments are based on the same process, but the DRA is much more formal. The summer school teachers were very experienced with both instruments; therefore, it was decided that time was the more important factor in the data collection protocol. Baseline RR scores used for comparison were not scores given for a single reading; rather the scores were the product of several days of instruction utilizing more than one RR from a variety of texts.

Spearman’s rho was used to calculate correlations between the spring DRA scores host schools provided and the Running Records summer teachers collected. The correlations can be interpreted like classical reliability coefficients that measure equivalence, stability, and inter-rater reliability because different forms of IRIIs were administered, with time elapsed between administrations and two sets of raters rated each sample (Linn & Gronlund, 2000). While there is no absolute
standard for desirable values, many researchers strive for values of at least .70 and others above .75. Using either criterion, the squared correlations for the data were all high for first- through fourth-grades, thereby indicating a strong level of equivalent forms reliability for the scores used to match the treatment and control groups, and a high level of consistency by the 35+ classroom teachers who administered and scored the spring DRAs (K, \( \rho^2 = .61 \); 1st, \( \rho^2 = .92 \); 2nd, \( \rho^2 = .88 \); 3rd, \( \rho^2 = .92 \); 4th, \( \rho^2 = .81 \)).

The inter-rater reliabilities of the writing samples scores were not as high as the DRA scores. Even with the best rubrics and training, evaluation of student writing maintains at least some level of subjectivity; therefore, each student’s writing sample was scored twice by rater pairs to reduce error variance. Correlations for the spring scorings of the kindergarten, first- and second-grade, and third- and fourth-grade writing samples for inter-rater pairs within grades were .76, .08, and .32 respectively and for the fall correlations were .58, .62, and .57. Unfortunately, only one of the scorings exceeded .75. It is likely that the restricted range of scores awarded to the samples hindered the correlations; scores awarded to the kindergarten writing samples ranged from two to five, scores for First- and second-grade only from one to four, and scores for third- and fourth-grade from one to four with decimal values awarded on third- and fourth-grade rubrics.

In light of these findings, and the restriction of range issue, the data were re-examined. Therefore, in what percent of cases were the ratings exact matches and in what percent of cases were they within one point? Table 1 presents the percent of time that the scores from the two ratings were an exact match and the percent of time that the scores were within one point of each other. Of the six cases considered, raters were within one point on their ratings at least 86% of the time. However, the question of whether ratings within one point of each other are accurate enough is a matter of validity.

Reliability of the writing scores was further examined using generalizability theory. Generalizability theory is a powerful methodology for addressing issues of rater agreement (Allen & Yen, 1979; Crocker & Algina, 1986) and superior to simply calculating classical reliability coefficients because it offers a prescription for improving the situation. In generalizability theory the systematic sources of variance in measurement are considered, and ways of estimating the variance attributed to these sources are provided. Generalizability coefficients are variance ratios comprised of the variance due to the objects of measurement (in this study, the scores of the writing samples), the total variance due to both the objects and conditions of measurement (in this study, the raters), and the variance from the interaction. The researchers conducted a two factor ANOVA using examinees and raters as factors and the scores on the writing sample as the dependent variable to obtain the scores necessary for computing the generalizability coefficients. Table 2 presents the
Table 2
Mean squares and generalizability coefficients.

<table>
<thead>
<tr>
<th>Grade/ Term</th>
<th>MS examinee</th>
<th>MS rater</th>
<th>MS residual</th>
<th>1 Rater</th>
<th>2 Raters</th>
<th>3 Raters</th>
<th>4 Raters</th>
<th>5 Raters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kindergarten Spring</td>
<td>0.57</td>
<td>0.74</td>
<td>0.30</td>
<td>0.76</td>
<td>0.86</td>
<td>0.91</td>
<td>0.93</td>
<td>0.94</td>
</tr>
<tr>
<td>Fall</td>
<td>0.63</td>
<td>0.73</td>
<td>0.17</td>
<td>0.58</td>
<td>0.73</td>
<td>0.80</td>
<td>0.85</td>
<td>0.87</td>
</tr>
<tr>
<td>First/second Spring</td>
<td>0.40</td>
<td>0.09</td>
<td>0.34</td>
<td>0.08</td>
<td>0.15</td>
<td>0.21</td>
<td>0.26</td>
<td>0.31</td>
</tr>
<tr>
<td>Fall</td>
<td>0.78</td>
<td>1.63</td>
<td>0.19</td>
<td>0.60</td>
<td>0.75</td>
<td>0.82</td>
<td>0.89</td>
<td>0.88</td>
</tr>
<tr>
<td>Third/fourth Spring</td>
<td>0.57</td>
<td>0.30</td>
<td>0.08</td>
<td>0.31</td>
<td>0.48</td>
<td>0.58</td>
<td>0.65</td>
<td>0.70</td>
</tr>
<tr>
<td>Fall</td>
<td>1.13</td>
<td>0.12</td>
<td>0.31</td>
<td>0.57</td>
<td>0.72</td>
<td>0.80</td>
<td>0.84</td>
<td>0.87</td>
</tr>
</tbody>
</table>

mean squares for examinees, raters, and residuals along with the generalizability coefficients for using from one to five raters.

In this case, the G-study coefficients appear under the heading “2 Raters” as the study used two pairs of raters to score each writing sample. The next step was to use the results from the G-study to conduct a D-study (i.e., decision study) to ascertain the number of raters needed to obtain acceptable reliability levels. Again there are no established rules for what constitutes acceptable values; therefore, the commonly used .75, or what Crocker and Algina (1986) denote as a “substantial generalizability coefficient” (p. 167), was used as the criterion. For the spring scorings, only the scores from the kindergarten writing samples reached this level and, according to the coefficients from the D-study, could do so with only one rating. For the fall scorings, all three groups were close to .75 and could exceed that criterion if three ratings were obtained for each writing sample.

4. Conclusions

4.1. Validity

The content-related validity of both the DRA and RR scores for the purpose of assessing reading ability are believed to be quite high. The scoring procedures are designed for the exact purpose of determining how well a child can read a given passage. A key distinction between the two measures is in scoring reading comprehension. The DRA rubric includes the score from a formal reading comprehension interview and, thus, takes much longer to administer than a RR. Reading comprehension is measured in RR with a short retelling of the story by the child and is a factor in the reading material selected by the teacher.

Therefore, if the purpose of the assessment is to measure the mechanics of reading then scores from both measures appear to have a high degree of content-related validity. However, if the goal of the assessment is to measure a child’s reading comprehension and make summative decisions, the DRA has a more rigorous approach. The fact that the DRA and RR are essentially two different tests with similar procedures and the same objective provides a means for exploring criterion related validity. The high reliabilities between scores on the two measures provides strong evidence for high criterion-related validity of the shorter RR assessments.

The low inter-rater reliability estimates for scores of the writing samples in the spring hinders estimates of validity, as reliability is a necessary yet not sufficient criterion for validity. However, as was mentioned previously there were possible restriction of range issues on the scores for the writing samples. If measuring students writing abilities within one point on the given rubrics is
sufficient enough to allow teachers and others to make decisions, then the scores are reasonably reliable. However, persons utilizing the scores must be aware of this limitation in their interpretation. The values from the D-study for the fall scorings indicate that the scores for the writing samples should have reasonable reliabilities if two ratings are used.

With respect to content and criterion-related validity, this approach to measuring writing appears to be sound; though not without flaws. Because a sample of student writing is the object under evaluation, then it would seem to have a strong degree of content-related validity for how well a student can write. A potential concern here is a mismatch between the assigned topic and the interests and experiences of the child.

The adaptation of collecting student writing samples during the students’ normal writing workshop time strengthens the validity of the conclusions drawn from the writing samples. Collecting the samples during this time should reduce such threats as the Hawthorne effect and novelty effects, and reduce test anxiety; thereby, resulting in more valid scores. The fact that teachers did not see a need to take a second writing sample, implying that the sample was indeed representative of the students’ abilities, provided evidence of the samples’ validity. Nevertheless, the writing samples still represent only what a child was able to write on a given topic, within a specified time frame, with no opportunity for feedback or revision. Ultimately, the researchers believe the content-related validity of such scores is so high, due to the ultimate match with the criterion, that no better measure could be found for which to examine criterion-related validity.

5. Discussion

The reliability estimates of the scores from the IRIs were encouraging with correlations ranging from .78 to .96. This indicates equivalence between scores for RR and DRAs that can have a significant impact on assessment due to the shorter evaluation time for the RR. Summer teachers’ perceptions that the DRA scores provided by the schools were much more accurate than the pilot year, are also noteworthy. Having teachers switch classrooms to administer the DRA had two effects: (a) DRA records were more complete and the protocol was followed more closely. In effect, colleagues served as monitors for each other because of the social pressure of having to submit the scores to someone else who cared a great deal about the results; and (b) it reduced the effects of bias caused by prior knowledge of the students’ level.

From the standpoint of validity, the conclusion by Paris and Hoffman (2004) that IRIs provide indications of multiple dimensions of reading abilities contributes significantly to the validity equation. Moreover, IRIs conducted by teachers provide scores with a high degree of reliability and validity without disrupting normal classroom procedures and, therefore, should be considered as one of many strategies for both formal program evaluation and more informal classroom assessment.

Measures of inter-rater reliability for scoring writing samples were not as high as the reading scores. The authors facilitated the scoring of writing samples, and were satisfied that the scoring of writing samples was as consistent as possible within the requirements of the local school district. Unlike the district’s protocol, each writing sample was scored twice by rating teams using the same grade-level rubric. It was thought that the team approach to scoring would increase both the reliability and validity of the scores as any issues or uncertainties could be resolved between the two raters before awarding a score to the writing sample.

Two scorings were used to reduce the error variance. However, the only inter-rater reliability to exceed .75 was the spring scoring of the writing samples for the kindergarten group. If, however, a score is usable if it is within 1 point on the rating scale; then 86% or more of all scorings met these
criteria. The decision about whether a score within one point of a child’s true score is accurate enough for decision-making would depend on the specific purpose of the evaluation. Contrary to the findings from the classical score reliability model, results from the generalizability study were more promising. Specifically, the D-study indicated that two ratings per sample should provide adequate reliability.

Of further interest is the pattern of inter-rater reliability estimates across the analyses that revealed more reliable scores in the fall than in the spring. There are two potential causes for this effect. First, the teachers could have become more proficient at scoring the writing samples in the fall than they were in the spring. However, the procedures were the normal procedures followed in the district with only the additional training that was given each semester.

The most likely reason for the more reliable scores was less teacher fatigue. In the spring, teachers scored writing samples for all 850 students twice. The scoring sessions involved 35 teachers and lasted as long as two hours. This was an enormous grading burden added to the teachers’ already busy schedules. In the fall, the teachers only scored the writing samples for the students who attended the summer program and the students in the control group; or 140 writing samples scored twice. The fall scoring only required 10 teachers and took about an hour.

In light of the differing reliability results, we believe that the move to writing samples over objective tests cannot be made without allowances for the time and effort required for effective scoring. Nonetheless, if teachers were relieved of the burden of administering the entire battery of tests that the district required, and did not have to alter dramatically their schedules to prepare for and administer standardized multiple-choice tests then fatigue, stress, and anxiety would be significantly reduced.

From a validity perspective, there are several suggestions for improvement. The first is in the area of content-related validity. Offering only one writing prompt allows a greater possibility for a mismatch between the topic and background of the student. Therefore, we believe that the content-related validity could be improved by offering two prompts from which the students may choose to write. To dispel the criticism that writing samples merely measure writing at one point in time, teachers could collect two writing samples from each child and score the better of the two samples to provide a more credible measure of the child’s abilities.

While this recommendation came out of the study, rather than being applied during the study, the concern of poor student performance on a given day was addressed by instructing teachers to have students submit a replacement writing sample if the teacher knew of students who “had issues” the day the sample was taken. Only one teacher felt compelled to do so for one child. It is believed that collecting data during the time in the school day when students normally write alleviated this concern.

The varying quality of different rubrics the district used to score the writing samples for kindergarten, first- and second-grade, and third- and fourth-grade was also a factor that affected the reliability of the scores. Inter-rater reliability estimates were much higher for the holistic rubric with anchors for kindergarten, and for the fully developed analytic rubric used for third- and fourth-grade. The poorest reliability estimates were generated by the holistic checklist (without anchor papers) used with samples in first- and second-grade.

6. Summary and recommendations

The data indicated that reading and writing abilities can be measured with the reliability and validity of the data high enough for program evaluation and rich enough to assist teachers with instructional planning. This study provides additional evidence that IRIs can be reliably
administered, and the literature clearly supports the authors’ beliefs that IRIs are a far more credible direct measure of reading ability than standardized multiple-choice tests.

The results of this study suggested that with a few modifications to the administration and scoring procedures, writing samples have the potential to be a dependable source of data for teachers, administrators, and policy makers. To increase the inter-rater reliability of scores from writing samples, teachers should work in rater pairs; receive training on the rubric; each writing sample should be scored twice; and teacher workload must be monitored throughout the grading process. The authors’ experiences also suggested that use of a common scoring rubric for kindergarten through fourth-grade would further increase reliability. Ideally such a rubric would utilize holistic categories (content, style, etc.), descriptions for each quality rating one through four, and anchor papers.

Suggestions for improving the content-related and criterion-related validity of the writing samples include providing students with two prompts and allowing them to write on the prompt with which they more readily identify; collecting two writing samples from each child and scoring the better sample; and collecting data during the time that the children normally participate in those activities. The processes utilized in this study quickly generated results that could be used to make placement decisions, plan instruction, and evaluate programs. Rather than using a battery of tests, schools could focus on collecting reading and writing scores that are credible, and generate results that could be used for both formative and summative decisions. The streamlined evaluation would dramatically decrease the instructional time lost during test preparation and administration and provide teachers with the information they need for immediate decision-making. The teacher training required to obtain reliable results would also create more qualified teachers, and hopefully more assessment driven instruction.

References


**John Burgin** is an assistant professor of Early Childhood Education at the University of Arkansas at Little Rock. Previous to his tenure at UALR, he was a reading specialist, a Kindergarten teacher, and a reading recovery teacher in the public schools for 20 years.

**Gail Hughes** is an associate professor of Educational Leadership at the University of Arkansas – Little Rock teaching graduate courses in Statistics and Research. She serves on the editorial board for the Journal of Educational Research and for Educational Technology, Research, and Development; and contributes as copy editor for Research in the Schools.