DOCUMENT RESUME

ED 352 371 TM 019 254

AUTHOR Barrett, Thomas J.
TITLE Implementation of an Integrated Language Arts Performance Assessment in a Large Urban School District: Technical Issues in Aggregating and Reporting Results.
PUB DATE Apr 92
NOTE 37p.; Paper presented at the Annual Meeting of the American Educational Research Association (San Francisco, CA, April 20-24, 1992).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Educational Assessment; Elementary Education; Elementary School Students; Equated Scores; *Integrated Activities; *Language Arts; Reading Achievement; Reading Tests; Scaling; School Districts; School Surveys; Scoring; *Student Evaluation; Test Reliability; *Test Validity; Urban Schools; Weighted Scores; *Writing Evaluation
IDENTIFIERS Aggregation (Data); Alternative Assessment; *Performance Based Evaluation; *Riverside Unified School District CA

ABSTRACT

This study assessed how some key measurement issues were considered in the context of a specific integrated reading and writing assessments conducted at each of seven grade levels in the Riverside (California) Unified School District. Data are reported for the following parameters: (1) test reliability; (2) test validity; (3) scoring; (4) scaling; (5) weighting; and (6) equating. Issues are considered in light of validity criteria proposed by R. L. Linn (1991). In general, reliability and validity reported for the assessment are encouraging, although test-retest reliability estimates and generalizability coefficients would add to evidence for the generalizability of the assessment. Evidence is also presented from teacher surveys that supports the meaningfulness of the assessment and positive impacts on instructional practices. Efficiency and cost effectiveness are also considered acceptable. Recommendations are given for alternative assessment methods. It is concluded that instruction might be best served by allowing schools to use performance assessment results in a more informal manner, with more subjective judgments about the process. Twelve tables present study data. (SLD)

IMPLEMENTATION OF AN INTEGRATED LANGUAGE ARTS PERFORMANCE ASSESSMENT IN A LARGE URBAN SCHOOL DISTRICT:

TECHNICAL ISSUES IN AGGREGATING AND REPORTING RESULTS

Thomas J. Barrett, Ph.D.

Department of Research and Evaluation
Riverside Unified School District
3380 14th St.
P.O. Box 2800
Riverside, CA 92516

Paper Presented at the Annual Meeting of the American Educational Research Association, San Francisco, April, 1992.

Implementation of an Integrated Language Arts Performance Assessment in a Large Urban School District: Technical Issues in Aggregating and Reporting Results

Thomas J. Barrett, Ph.D.

Department of Research and Evaluation
Riverside Unified School District

Introduction

Much of the discussion of technical measurement issues pertaining to authentic (or performance) assessment has been general in nature. Lorrie Shepard (1991) in an interview focused on alternative assessments acknowledges the success that direct assessments such as the College Board's Advanced Placement tests have had in establishing reasonable reliability and validity. James Popham (1991) has addressed some of the potential threats to the validity of authentic assessments and cites the example of Tennessee's teacher career-ladder portfolio program (demands by teachers to have the assessment criteria publicized diminished the validity of the assessment).

Others, like Mehrens (1991) and Williams, Phillips, and Yen (1991) discuss issues in reliability, validity, scoring, scaling, and comparability that must be addressed in order for performance assessments to meet schools' many evaluative needs. Linn, Baker, and Dunbar (1991) cogently set down several categories of criteria against which performance assessments should be measured in order to certify their validity. Such validity concerns as (1) consequences of the assessment, (2) fairness, (3) generalizability of outcomes, (4) adequacy of cognitive complexity of tasks, (5) content quality, (6) content coverage, (7) meaningfulness of the results, and (8) efficiency and cost effectiveness, are intelligently discussed. Yet in all of these references, few details are provided to give the practitioner in the field a really concrete appreciation of how these issues might relate to a real-world implementation of a performance assessment.

This paper is an attempt to illustrate some of the key measurement issues by describing how they were considered in the context of a specific integrated reading and writing assessment. The assessment was conducted at each of seven grade levels in the Riverside Unified School District in Riverside, California. Approximately 2200 students were assessed at each grade level in the spring of 1991. After reporting data on reliability, validity, scoring, scaling, weighting schemes and equating, the author attempts to summarize some of these considerations in light of the eight criteria delineated by Linn et. al. (1991) and draw some conclusions about recommended directions for school districts to follow.

Description of Assessment

The Riverside Unified School District has adopted a commercially developed integrated reading and writing performance assessment--the Language Arts Performance Assessment (LAPA)--for use at grades 1 through 7 (The Psychological Corporation, 1990). Those wanting a more complete description of the assessment procedures used in our district are referred to an article by Ferrett (1991). During the assessment, students read for the purpose of writing. The reading materials include many types of passages--fiction, poetry, plays, expository selections and many more (see Table 1). Many of the passages contain excerpts from the works of well-known children's authors.

Each reading/writing prompt is designed to elicit higher-order thinking. Some prompts, for example, require students to compare and contrast, while others require students to synthesize information from multiple sources. Still others engage students in analyzing the actions and motives of characters and predicting what will happen next. The prompts themselves elicit different types of writing; informational reports, story ending, writing directions, persuasive essay, or character analysis.

Unlike many writing tasks that call for students to produce a final draft in a brief time period, the reading/writing prompts take students through all the steps of the writing process: planning, editing, revising, and production of a final paper. Collaboration and peer editing are encouraged while the teacher may serve as a coach to motivate, prod and encourage students to produce their best work without doing it for them. The assessment is not a timed task, but students are not to take more than four class periods to complete the assessment.

For each grade level, three reading/writing tasks are available. Each prompt elicits a different type of reading, thinking, and writing. Early in the year, the district's Research and Evaluation Department randomly assigned each teacher one of the three prompts to be used as the required language arts performance assessment to be administered in April. Because all of our schools have at least three classes per grade, each prompt was assigned to at least one classroom in a school. In schools with more than three classrooms, the number of students taking each of the prompts differed depending on the number of classrooms taking a given prompt.

Teachers were trained for the administration of the assessment through use of a "trainer-of-trainers" model. Each elementary school sent a primary and upper grade representative to a two hour district conducted inservice in November. Middle schools sent English and math department representatives. These representatives then went back to their schools and were responsible for training the school's staff.

Following the assessment, from seven to ten teachers per prompt were given four hours of training to evaluate the papers. Scoring was held on two Saturdays during which all of the student papers were read once during a four to

six hour session. Teacher packets of papers were opened one at a time and distributed around the table to the readers so that all readers were involved in the scoring of a given classroom set of papers. A sample of about thirty papers were selected for each prompt and sent to the test publisher's scoring service for a second reading in order to estimate inter-rater reliability.

Description of Scoring

The scoring is done on a four-point scale along three separate dimensions. The dimensions are (1) Response to Reading (RR) (2) Management of Content (MC), and (3) Command of Language (CL). Response to Reading encompasses the amount of information included from the passages, the accuracy of that information and the relevancy of the information to the task. Management of Content includes the level of organization and focus, development, and the degree to which the task was accomplished. Command of Language assesses sentence structure, word choices, and grammer/usage/mechanics.

Although the detailed scoring rubric differed somewhat from prompt to prompt depending on the specific task to be accomplished, the description of the three scoring dimensions was invariant across the twenty-one prompts used. Because of the emphasis placed on a similar dimension in the California state assessment program (CAP), we decided to weight Management of Content double that of the other two dimensions in arriving at a weighted average total score for each student.

Analysis of Results

Reliability--

One of the advantages of using a commercially developed assessment is that substantial technical information is available through the publisher's pilot and field test process. However, it also behooves districts adopting commercial materials to study how locally implemented variations of the assessment (eg. local training and scoring, local differences in student performance, etc.) affects the technical characteristics of the measure.

The publisher reports inter-rater reliability as measured by correlations between raters to be in the .80s and .90s (Table 2). Exact agreement between scorers was over 75 percent on nearly every dimension for each prompt. As can be seen in Table 3, exact agreement was in the 80 to 90 percent range in many cases and differences greater than one score point (requiring resolution) occurred only 11 times in grades one through seven.

In order to estimate the inter-rater reliability of the local scoring in our district, we sent a sample of about thirty papers for each prompt to the test publisher's national scoring center. Correlations between our raters and those of the scoring service are given in Table 4. It can be seen that although the correlations ranged from a low of .48 (prompt 702) to a high of .95 (prompt 301), the

average correlation of .88 indicates strong agreement between our ratings and those of the scoring service.

Comparison of the percentage of agreement between our readers and the publisher's readers (Table 5) shows that while the percentage of exact agreement tended to be lower in our sample than that reported in the publisher's field test, the number of scores differing by more than one point was very similar (11 per publisher vs. 12 per RUSD). In 15 out of 63 categories reported (prompt by dimension combinations), the percentage of exact agreement found between our local ratings and the scoring service was higher than that found between the two readers in the publisher's field test.

We also ran a contingency table analysis (crosstabulation) between dimension scores and readers for each prompt. Since papers can be considered to be approximately randomly distributed among the readers, the expectation is that the proportion of papers scored 1,2,3 and 4 would be similar if scorers are rating papers uniformly. Table 6 shows that for the twenty-one prompts scored on three dimensions, Cramer's V ranged from .12 to .54. Cramer's V ranges from 0 to 1 with 0 indicating no relationship between reader and the scores assigned. Thus, the lower the value of this statistic the more we can be assured that scores given on a particular dimension are independent of the reader.

Although there were some instances in which a given reader departed from the scoring pattern produced by the group as a whole, this analysis shows that our readers were reasonably consistent in their scoring at grades 1 through 6. At grade seven, similar score by reader independence was found for prompt 702 while Cramer's V for prompts 701 and 703 was in the .50's. At this point, we cannot explain this anomoly. There were no indications from the teacher adminstration training or the scorer training to suggest a greater inconsistency of ratings for these two seventh grade prompts. On the contrary, correlations between RUSD scoring and the scoring service was actually quite high for prompts 701 and 703.

Generalizability theory is another, more sophisticated, approach to examining the reliability of performance measures and the appropriateness of making broader inferences about achievement from a limited sample of behavior. By designing and carrying out a well conceived study, it is possible to partition the sources of variation in performance assessment scores to arrive at a generalizability coefficient representing the degree to which one is justified in generalizing assessment results. With this method the investigator can isolate the variance in scores attributable to such things as differences in tasks, readers, testing occasions (test-retest), and even levels of teacher preparation in administering the assessment. These are all components of variance that warrant caution in interpreting performance assessment results but that may be addressed rigorously with generalizability theory.

To date no generalizability studies have been conducted in our district. However, others have reported findings that indicate a substantial amount of variability attributable to the specific task performed (Shavelson, Mayberry, & Li &

Webb, 1990; Swanson, Norcini, & Grosso, 1987; Shavelson, Baxter, and Pine, 1990). Those who would like a relatively readable introduction to generalizability theory are referred to the paper by Webb, Rowley, and Shavelson (1988).

Validity--

The publisher also reports data on content and construct validity in its technical manual. Evidence for content validity includes teacher feedback regarding the appropriateness of the materials used during the field test. The authors report that approximately 75 percent of the teachers thought the LAPA prompts were reasonable assessment activities for their instructional programs. Identical surveys used in our district showed similar results.

Perhaps more importantly, the publisher surveyed leading language arts instructional programs prior to development of the materials as well as various state and school district curriculum guides. Based on this input and the professional status of the LAPA authors, it is reasonable to conclude that a significant level of content validity inheres.

Although no correlations were reported between the LAPA assessment and multiple choice measures of writing, it was assumed that correlations would be similar to those found in other studies of direct writing. These correlations typically range between .40 and .60. Although the author was not able to correlate the LAPA results with a multiple-choice writing test, we were able to correlate each dimension of the LAPA with the reading comprehension subtest of the Stanford Achievement Test, 8th Edition, Abbreviated (The Psychological Corporation, 1989). As shown in Table 7, correlations between Response to Reading and the Stanford ranged from .25 (prompt 301) to .47 (prompt 202). It was also found that reading comprehension on the Stanford often correlated less with the response to reading dimension than it did with the other two dimensions. Since response to reading is measuring what a student brings to his or her writing from the reading selections, it should be considered to be measuring a different facit of reading comprehension than does a standardized multiple choice reading test. In light of this, school districts might want to seriously consider continued use of a multiple choice reading test to compliment LAPA response to reading scores. While the LAPA's approach to integrating reading and writing is to be commended, it probably ought not stand alone as the sole measure of reading.

Although the above data are relevant, the issue of content validity encompasses more than just whether the materials used are appropriate measures. A larger question is how well these tasks sample from the universe of possible writing tasks? The California CAP includes as many as eight genres or modes of writing in its assessment and writing research has shown that at least four factors should be judged to fully assess writing: content, organization, sentence structure, and mechanics (Freedman, 1979). To have adequate content validity, an authentic assessment program in writing should certainly address the adequacy of the sampling of types of writing as well as the appropriateness of the scoring rubrics. Although the LAPA prompts include many types of writing

when all grade levels are considered, it is unclear whether the three tasks available for each grade adequately sample from the universe of possible writing modes.

The construct validity of the measure was assessed by the publisher through use of Campbell & Fiske's (1959) "multitrait-multimethod" procedure to show that the test not only correlates highly with other variables with which it should correlate (convergent validity), but also that it does not correlate with variables with which it should differ (discriminant validity). The three dimensions-Response to Reading, Management of Content, and Command of Language-were considered to be the "traits" assessed and scorers were considered the "methods" of assessing those traits. Correlations between different scorers on the same trait ranged from .51 to .96 although most ranged between the mid-.70's to high .80's and .90's providing strong support for convergent validity (Farr, R. and Farr, B., 1991).

Likewise, correlations between readers on different traits and correlations between different traits for the same readers was lower than the correlations between different readers on the same trait. Thus, there was also substantial evidence for discriminant validity.

Inter-dimensional correlations reported by the publisher showed that Response to Reading was generally more highly related to Management of Content than with Command of Language (Table 8). This is presented as evidence of validity in that one would expect this pattern to hold given the anticipated stronger relationship between the reading dimension and the organization dimension than between reading and writing conventions. Also, Command of Language shows a stronger correlation with Management of Content than with Response to Reading--another expected pattern.

Another important point to note is that over 70 percent of the first graders assessed in our district achieved a weighted average score of 3.75 or 4.00. This finding is consistent with high average scores reported by the publisher at grade one and indicates minimal discrimination (see Table 9). It should be noted, however, that the grade one rubric is different than that used at the other grades with management of content reflecting the ability of the student to produce written print rather than the ability to organize and develop a writing sample. Students are assigned a score of 4 if nearly all of the responses show an attempt to communicate with print. This is clearly an easy criterion which most students can meet if they are reasonably attentive to the requirements of the task. It would probably be more appropriate to use the grade one prompts at the end of kindergarten or early in the first grade experience as a measure of writing readiness. It should also be pointed out that our weighting management of content double that of the other two dimensions results in an even greater skewing of the scores.

Comparability-Aggregation-Reporting

Because we wanted to have enough time to thoroughly analyze the assessment data prior to reporting results, we chose not to publicly release results of this first year of the LAPA assessment. Ideally, we wanted to be able to use the results for a variety of purposes including: (1) evaluation of the effectiveness of our language arts writing curriculum at the classroom, school, and district level (2) assessment of individual student competency in writing at the elementary level, and (3) re-designation of Limited-English Proficient (LEP) students to Fluent-English Proficient (FEP). At the same time, we realized that using the assessment as we did (ie. one prompt per class) could potentially lead to some sticky issues regarding establishment of score comparability as well as aggregation and reporting of results.

In our district, we first considered developing a standard score scale with a mean of 250 and standard deviation of 50 to mimic the scale developed by the California State Department for its direct writing assessment (CAP). We abandoned this idea when it became clear that such a standard score conversion would adjust only for differences in means and variances between prompts but would not adjust for differences in the shapes of the distributions. Table 9 shows the distributional characteristics of the scores obtained in our district and demonstrates that the distributions tend to differ in shape as well as means and variances. We also wanted to be able to report scores on the original unequated 1 to 4 scale in order to provide information regarding the district's relative strength and weakness on the different types of writing assessed. By setting means and variances to 250 and 50 respectively, differences in district performance on the prompts would be masked.

It is also important for the test results to make sense to teachers and parents. Clearly standard scores are more difficult to understand than the 1 to 4 scale utilized in the scoring rubric. Furthermore, our scale would really not be equivalent to that of CAP anyway. The California State Department had thousands of writing samples available and went beyond conventional linear equating by using a sophisticated IRT calibration technique for Likert type scales (Muraki, 1990) -- a methodology which is beyond the resources of most school districts.

Still, in order to give ourselves some assurance that decisions regarding level of writing competency at the individual student level was not a function of the writing prompt that the students took, we established prompt specific weighted score cutoffs for LEP/FEP re-designation at grades 1 through 7 and for elementary competency assessment at grades four and five. This represents an equipercentile (curvilinear) equating at the cutpoints and is appropriate when the shapes of the score distributions for the measures differ. Such an equating is meaningful, of course, only if the groups taking the different prompts district-wide are of equal ability.

Although we attempted to randomly assign prompts to classrooms within a school, there were some instances in which schools made their own assignments in order to accommodate various logistical problems. In order to be assured that the groups taking the three prompts at each grade level district-wide were essentially equivalent, we looked at their performance on the Reading Comprehension subtest of the Stanford.

With the exception of prompt 303 at grade three, the range of the average NCE's for the grade level groups was only two points. This gave us reasonable assurance that the groups taking the different prompts at each grade were substantially equivalent. Prompt 303 had an average NCE nearly seven points higher than 301 which implies that the students taking this prompt had a higher level of ability than the other third grade groups district-wide. Consequently, equating results for this prompt should be regarded cautiously.

Historically, we have used passing scores of about the 23rd percentile for elementary competency and the 36th percentile for redesignation from limited-English proficient (LEP) to fluent-English proficient(FEP). As can be seen in Table 10, we were generally able to establish competency cut points that passed similar numbers of students for grades four and five. For LEP/FEP re-designation purposes, however, the percentage of students passing a given prompt ranged from 60 percent to 72 percent at grade seven and 62 percent to 72 percent at grade six. These two grades presented the biggest problem in establishing comparable cutoffs.

One of the difficulties encountered in establishing comparable scores was in part a result of the selection of weighting factors in computing the weighted average writing score. By weighting Response to Reading and Command of Language dimensions by twenty-five percent each and Management of Content by fifty percent, we had only thirteen possible scores on which to base an equipercentile equating. According to Anghoff (1971), at least thirty score points are generally recommended to accomplish a solid equipercentile equating throughout the scale. Had we weighted all three dimensions differently, we might have generated a broader distribution of scores and possibly improved the equating.

Table 11 shows how the cutoff scores would look for LEP/FEP re-designation and elementary competency if we weighted the three score dimensions as follows: RR(.20) MC(.50) CL(.30). (Incidentally, in the California Assessment Program, writing is scored on two six point dimensions- (a) rhetorical effectiveness, and (b) conventions, with the dimensions weighted 85 percent and 15 percent respectively. The resulting thirty-six possible scale points coupled with far more papers available for the study, allows for greater accuracy of equating than does our weighted average scale.)

Notice that for grades six and seven a cutoff can now be established which results in a more uniform number of students passing. The range of passing percentages is now only four points instead of twelve at grade seven and the range was reduced from ten to six points at grade six. It was observed, however, that

even though we increased the number of possible score points to 31 with the .20, .50, .30 weighting scheme, we still tended to have very few cases at many of the score points and none at others. Any attempt to accomplish a horizontal equating of prompts throughout the scale at a given grade level would be limited as a result. Therefore, the decision was made not to attempt an equipercentile equating of the entire weighted average scale in our district.

Similar difficulties arise when attempts are made to conduct vertical equating studies to construct a longitudinal scale for facilitating gain score analysis from year-to-year. With three prompts per grade, the number of cases needed to equate from all three prompts at one grade to any of the three prompts at the next lower grade implies an elaborate design. In order to adjust for differences in distributions between all combinations of prompts that students might take from year-to-year would require that a design be used like the one illustrated in the top section of Chart 1. For each grade, we would need to have students take one of nine combinations of prompts. This would clearly be cumbersome to administer given the amount of teacher attention required for each task. More importantly, since only one-ninth of our students would get a given two-prompt combination, we would face an even greater problem with inadequate score density at the points in the scale than we had in our horizontal equating attempts.

A preferred design is given in the bottom section of Chart 1. Here, each classroom would need to administer only two prompts-the one assigned for the class's current grade level and one of three from the next lower grade. The first step would involve equating prompts 102 and 103 to 101. Next, prompt 201 could then be equated vertically to 101. Then prompts 202 and 203 would be equated to 201. This same procedure would then be followed for the other grades. This "linked" equating is possible but it compounds the equating error by requiring the "equating of equated scales". Although this is commonly done in conventional observed score equating with multiple choice tests, given the restricted scale of the LAPA, it would appear that any attempt to conduct such a vertical equating in our district could lead to unacceptable levels of error.

An additional consideration in establishing comparability of scores from year-to-year involves potential drift in grading standards. That is, the pool of scorers may change how they judge identical papers from one year to the next. As Linn et. al. (1991) points out, it is desirable to rotate a sample of papers in and out of the scoring pool each year to adjust for this drift (the CAP writing program includes such a step in its scaling procedures.)

Aggregating and Reporting-

If conventional observed score equating procedures seem limited for use with RUSD's LAPA program, how then might scores be aggregated for school-to-school comparisons? One idea that we explored in Riverside was to arrive at an overall score for a grade level at a school by computing the percentage of students getting weighted average scores above a cutoff for each prompt and then averaging these percentages. This would allow for more fair comparisons

between schools since the percentage scoring above the cutoff would be independent of the proportion of students taking a given prompt at a particular school.

This can also be problematic, however, unless the classrooms taking the three prompts can each be assumed to include students selected randomly from all students at that grade and school. If classes taking the prompts differ in ability level, the above procedure would be appropriate only if there is no substantial Ability X Prompt interaction effect. That is, we would not want a school's overall performance to depend on which prompt was assigned to which classroom. To test for this possibility, we separated students at each grade level into four ability groups based on percentile ranks on the Stanford Reading Comprehension test. The groups were defined by percentile ranges as follows: Level 1=1-10; Level 2=11-50; Level 3=51-89; Level 4=90-99. We chose to define the highest and lowest groups to be extreme performers since it would be most typical for extreme classrooms in our district to be comprised of either a significant number of special education, limited-English proficient, or gifted and talented students.

As is evident in Table 12, the Prompt X Ability interaction effect was statistically significant at grades five and six only. Inspection of the cell means revealed that the interaction effect is most evident for the highest ability group. This group found prompt 503 to be the most difficult although it was the easiest prompt at fifth grade overall. At grade six, prompt 601 was easiest overall yet the highest ability group scored best on prompt 603. Still, it does not appear that the differences in the scoring pattern is dramatic for the various ability groups. Therefore, based on this analysis, we believe that averaging percents scoring above a given criterion on each prompt is a legitimate way to report school level performance.

While an ability by prompt interaction effect is without question a potential threat to comparability, a related but probably more crucial threat involves the interaction of classroom instruction with the prompts. To the extent that a given teacher has focused on a particular genre for his/her classroom's writing instruction, the extent to which the assigned prompt matches this preparation is a key consideration. In our district, we assigned prompts to classrooms in November and informed the schools so that teachers would be able to integrate the requirements of the assessment with their lesson plans throughout the year. Although this may serve to dampen the effect of any instructional mismatch and strengthen the "fairness" element of validity, this could also lead to an undesirable narrowing of the instructional focus; a situation which would tend to decrease validity related to educational "consequences" (Linn,et.al.,1991).

Another potential threat to the comparability of classroom level and school level scores involves the procedures used for distributing papers to the trained readers. In our district we chose to have papers for a given class distributed to all of the readers for that particular prompt. By doing this, we are in effect controlling for any "reader" effect on classroom and school level aggregation. We can be assured in comparing a classroom average to the district-wide average for

the same prompt that both averages reflect the ratings of all readers. While some readers will be somewhat lenient and others will be somewhat stringent, on average, these slight biases will tend to cancel each other out.

Although some teachers have argued that a perceived inconsistency in reader ratings makes it difficult at times to understand why apparently similar papers in their classes received different scores, it is not recommended to have a single individual reading the papers for an entire classroom. We know from the contingency table analysis reported earlier (Table 6) that some readers do in fact depart from the district average scores. To have a single reader responsible for the reading of an entire classroom's papers introduces an occasional reader bias that can detract from the ability to compare results from classroom-to-classroom and even from school-to-school.

Due to the limitations inherent in using one prompt per classroom, we chose to provide reports that allow teachers to compare the performance of their classrooms with other school-wide results and with district-wide averages for the same prompt . Direct comparisons between groups taking different prompts are, lacking adequate equating procedures, considered to be inappropriate.

Discussion

The following discussion represents an attempt to pull our findings together and to discuss them in the context of the eight validity criteria set forth in the Linn et. al. (1991) article. In addition, a number of recommendations are made for how districts can improve on the way performance assessment was carried out in RUSD.

In general, we feel confident that the aspects of reliability and validity reported for the LAPA assessment are encouraging. However, test-retest reliability estimates and generalizability coefficients would certainly add to the evidence supporting the "generalizability" of results, as would criterion validation with other direct writing assessments. With students taking only one prompt per grade, it is likely that a well designed generalizability study would call into question how well the assessment results reflect individual student writing achievement more broadly defined. In fact, other generalizability studies have found that limiting examinees to one performance task substantially detracts from generalizability. While there are threats to the criterion of "fairness" as referenced earlier, we believe that our assessment program leads to relatively fair comparisons at least at the school level. The "cognitive complexity" and "content quality" of the tasks appear to us to be quite good although there is some question about how well the three prompts per grade lead to adequate "content coverage".

There is also evidence from teacher surveys, both from the publisher's field test and our district, in support of the "meaningfulness" of the assessment and the positive impact on instructional practices ("consequences"). Some of the typical answers given to the question posed in our district survey "How do you think teachers can best prepare students for this kind of assessment?" included recommendations for having students write on a daily basis using varied styles

(modes) of writing, establish standards for each grade level, use all steps of the writing process, weave writing into all subject matter, and using a variety of literature in the classroom. In effect, teachers recommended that they do all of the things widely recognized to represent exemplary writing instruction!

The final element of validity delineated by Linn et. al.(1991) refers to the "efficiency and cost effectiveness" of the assessment. The trainer-of-trainer model used to deliver inservice to the teachers for test administration, we believe to be an efficient way to train. However, it should be noted that in point of fact, some schools conducted a more thorough and timely training program than others. With regard to scorer training, the five hours per teacher for seven to ten teachers per prompt we believe was a reasonable expenditure. And while the lion's share of expense went toward the actual scoring of the papers, implementation of score moderation techniques for adjusting teacher ratings based on results from a sample of papers has great potential in controlling overall scoring costs (this technique is explained later in this section).

Recommendations

There is an alternative assessment design that would have resulted in students continuing to take only one prompt, but that would have allowed us to make better comparisons at the classroom level. In this design, each classroom would be divided randomly by the teacher into thirds with each group writing to one of the three prompts at that grade level. While this could get overly complicated, especially with combination classrooms, flexible grouping patterns might be used at a school that would allow for students to be assessed by teachers other than their regular instructor.

When attempting to assess gains in writing over time (and to make comparisons between classrooms at a particular grade level), a potentially useful approach involves the establishment of local percentiles. Here, a normative score could be developed based on the local cumulative frequency distributions for each prompt during the baseline administration of the assessment. Then, gains could be calculated relative to the baseline regardless of which prompt was taken the next year. This is really nothing more than an equipercentile approach for group data. (It would still be difficult to assess individual student growth without adequate vertical and horizontal equating procedures for the entire scale.) Although one could convert local percentiles to NCE's and measure group progress in terms of NCE gains, an alternative that is probably more meaningful to non-technically oriented users involves the definition of one or more "performance criteria" based on the local percentiles. In California, plans are underway to report assessment results in terms of percentages of students doing "commendable" work (above the 75th percentile) and percentages doing "adequate" work (above the 35th percentile).

To illustrate, assume that 10 percent of the scores for a group of second grade Chapter 1 students is above the 35th local percentile in the base year. This percentage can be computed across all prompts taken since the local percentile is computed independently by prompt. Gains from year 1 to year 2 could be

measured in terms of the percentage above the 35th local percentile for second graders in the base year (once again with the base year local percentile computed on a prompt-by-prompt basis). Not only does this approach allow for reasonably fair comparisons , it also reports results in a way that may be more meaningful to users--especially the public. The percentage above a certain established criterion is easier for parents (and often teachers) to understand than a derived scaled score or perhaps even an average score on the scale defined by the scoring rubric.

In trying to establish score comparability for students and groups taking different prompts, the realization that we are probably measuring somewhat different abilities with the various tasks suggests that for group reporting, it may be inappropriate to attempt an equating of the prompts in the first place. Rather, it may be more reasonable to report scores separately by prompt and leave the scores on the original scale defined by the scoring rubrics.

For all levels of aggregation and especially for evaluating individual student performance, the preferred alternative to equating would be to administer all three prompts at a grade level to every student. Given the constraint of being limited to only one sample of writing per student, we felt that we were forced to establish equipercentile equating at the cutpoints for purposes of elementary competency certification and re-designation from LEP to FEP because we could not tolerate pass/fail decisions that are a function of the specific prompt taken. However, using multiple writing samples for each student would not only preclude the need for equating, but would also enhance the reliability and content validity of the assessment by having more "items" in the measure. Although at first glance this appears costly, there are ways to accomplish it without increasing the cost of scoring in the district beyond current levels.

To do this would require that teachers do their own scoring and that the district obtain a sample of papers from each class to be scored centrally by expert raters. Wilson (1991) describes a method for adjusting (or "moderating") teacher ratings by including information from a sample of papers. "Suppose that, for each student i in class c, we have two measures: X_{ci} which is the teacher's rating of student i in class c, and Y_{ci} which is the external rating of the same student. Then produce a moderated rating Z_{ci}, according to the following formula:

$$Z_{ci} = \bar{Y}_c + \frac{(X_{ci} - \bar{X}_c) S_{Yc}}{S_{Xc}}$$, where

$\bar{X}_c$ and $\bar{Y}_c$ are the means of X and Y within class c, and S_{Xc} and S_{Yc} are the standard deviations of X and Y in class c. . . . ."

In effect, a moderated score on each prompt could be derived for each student by sampling as many as a third of the students in each teacher's classroom for scoring by the experts. This should be more than adequate for making the adjustments in teacher scores without increasing the total number of papers centrally scored beyond current levels (scoring three papers for one third of the students is equivalent to scoring one paper for all students).

Conclusions

Given our experience in using an integrated reading and writing measure to serve multiple purposes in both focusing instruction and providing several levels of accountability reporting, it seems clear that the preferred assessment design requires multiple samples of behavior from each student. This would not only increase the reliability and content validity of the assessment but would also greatly simplify the ways in which we can appropriately aggregate and report results.

In the absence of such an option (ie. where only one sample is available per student), use of local norms established for each task administered is a viable alternative. Finally, a simple, yet reasonable way to report grade level results for schools is to average percentages above a non-equated cutpoint for the three prompts.

In closing, it is suggested that, in the broader picture, instruction might be better served by allowing schools who know their program best to use performance assessment results in a more informal manner and make more subjective judgements about progress rather than by forcing instructionally oriented assessments into the mold of psychometric technologies that, depending on the specific context of the assessment, are often limited in their usefulness.

When assessment can be designed to be as unobtrusive as possible and to look as much like good instruction as possible, there should be little in the way of road blocks to keep us from gathering the results of multiple samples of authentic performance for students. And when teachers can be trained to reliably assess their own students' work (perhaps with the aide of the above mentioned moderating techniques), then the final logistical impediment can be overcome. Performance assessment might then become a truly integrated assessment of instructional effectiveness and realize its full potential.

Table 1
LAPA Prompts
(FROM PUBLISHER'S MANUAL)

Code	Grade	Title	Type of Reading	Type of Thinking	Type of Writing
101	1	Bill's Hat	Picture Story	Predicting/Inferring	Sentences
102	1	Silly Animal Stories	Predictable Picture	Predicting/Inferring	Sentences
103	1	Robert the Robot	Poetry	Analyzing/Extending	Sentences
201	2	Nancy	Fiction	Extending/Predicting	Story Ending
202	2	The White Cat	Fiction	Extending/Predicting	Story Ending
203	2	Turtle's House	Fiction	Analyzing/Evaluating	Friendly Letter
301	3	At Mary Bloom's	Fiction	Modeling/Applying	Pattern Story
302	3	Baking Day	Fiction	Analyzing/Sequencing	Directions
303	3	Dinosaurs	Nonfiction	Comparing/Contrasting	Report
401	4	A Desert Garden	Nonfiction	Analyzing/Sequencing	Directions
402	4	What's Up Amelia?	Poetry	Analyzing/Evaluating	Letter
403	4	Heavy Steps	Fiction	Predicting/Extending	Story Ending
501	5	Learning About Mammals	Nonfiction	Analyzing/Summarizing	Report
502	5	The Wizard of Oz	Fiction	Analyzing/Evaluating	Persuasive Paper
503	5	Lost in a Storm	Fiction	Analyzing/Synthesizing	Newspaper Story
601	6	Needing a Friend	Fiction	Analyzing/Evaluating	Letter
602	6	Famous Flights	Nonfiction	Applying/Synthesizing	Feature Article
603	6	The Last of the Light	Fiction	Extending/Predicting	Story Ending
701	7	The Sinking of the Titanic	Nonfiction	Extending/Applying	Story
702	7	TV Viewing	Nonfiction	Analyzing/Evaluating	Persuasive Essay
703	7	Best Friends	Fiction	Analyzing/Critiquing	Report

Table 2

Interrater Reliability Estimates by Dimension and Prompt for LAPA
(PUBLISHER'S FIELD TEST)

		(Pearson)			(Spearman)		
Prompt	N	RR	MC	CL	RR	MC	CL
101	200	.89	.98	.89	.85	.95	.89
102	211	.89	.95	.84	.88	.97	.78
103	292	.94	.93	.87	.94	.95	.85
201	321	.85	.84	.89	.82	.82	.88
202	179	.94	.92	.94	.93	.90	.94
203	173	.67	.72	.77	.63	.71	.77
301	200	.93	.84	.79	.92	.84	.78
302	211	.96	.96	.94	.96	.95	.93
303	428	.97	.96	.97	.97	.95	.97
401	254	.96	.96	.95	.96	.96	.95
402	218	.84	.85	.81	.84	.83	.81
403	320	.86	.85	.84	.85	.84	.84
501	202	.78	.77	.70	.77	.76	.71
502	298	.90	.93	.93	.88	.93	.93
503	320	.90	.85	.85	.88	.85	.85
601	268	.86	.86	.82	.85	.84	.80
602	179	.79	.75	.51	.78	.74	.52
603	145	.94	.89	.91	.95	.90	.91
701	164	.89	.86	.91	.89	.86	.90
702	150	.84	.86	.84	.83	.83	.82
703	155	.95	.93	.91	.95	.93	.92

Table 3

Percentage of Agreement Between Readers
for Field Test Scoring of LAPA
(PUBLISHER'S FIELD TEST)

		Response to Reading			Management of Content			Command of Language		
Prompt	N	Exact Agreement	Adjacent Agreement	Differences Requiring Resolution	Exact Agreement	Adjacent Agreement	Differences Requiring Resolution	Exact Agreement	Adjacent Agreement	Differences Requiring Resolution
101	200	84	16	0	100	0	0	79	20	1
102	210	81	19	0	97	3	0	80	20	0
103	292	94	6	0	98	2	0	83	17	0
201	322	82	17	1	75	25	0	83	16	0
202	179	94	6	0	90	10	0	92	8	0
203	173	69	30	0	71	28	1	77	20	3
301	200	82	18	0	73	27	0	74	25	1
302	211	92	8	0	92	8	0	90	10	0
303	427	96	4	0	95	5	0	97	3	0
401	254	94	6	0	95	5	0	92	8	0
402	260	76	24	0	78	22	0	73	27	0
403	330	77	23	0	78	22	0	69	31	0
501	202	68	31	0	69	30	0	54	42	4
502	298	85	15	0	89	11	0	90	10	0
503	320	87	13	0	83	16	1	80	19	1
601	268	81	19	0	81	19	0	77	23	0
602	179	76	24	0	74	25	1	55	37	8
603	146	95	5	0	85	15	0	89	11	0
701	161	80	20	0	75	25	0	84	16	0
702	150	75	24	1	83	17	0	77	23	0
703	151	94	6	0	91	9	0	87	13	0

Table 4

Inter-rater Correlations-Weighted Scores
RUSD and Scoring Service

Prompt Number	Prompt Title	Number Of Students In Study	RUSD Average Weighted Score	RUSD Standard Deviation	PSY. CORP. Average Weighted Score	PSY. CORP. Standard Deviation	Correlation Between RUSD & PSY. CORP.	Significance Of Correlation
101	Bill's Hat	33	3.8030	0.3046	3.8864	0.2081	0.8373	< .01
102	Silly Animal Stories	33	3.6818	0.6473	3.7424	0.5745	0.9334	< .01
103	Robert The Robot	31	3.8790	0.2651	3.7903	0.2150	0.7827	< .01
201	Nancy	29	2.6207	0.7399	2.3707	0.6499	0.6974	< .01
202	The White Cat	29	2.4828	1.2025	2.3793	1.1252	0.9387	< .01
203	Turtle's House	33	2.1515	0.9415	2.2803	0.7441	0.8414	< .01
301	At Mary Bloom's	33	3.0759	0.7868	3.1288	0.8056	0.9548	< .01
302	Baking Day	33	3.1061	0.5591	3.0303	0.4134	0.7968	< .01
303	Dinosaurs	33	3.0227	0.6039	2.9545	0.5845	0.9492	< .01
401	A Desert Garden	23	3.4674	0.4479	3.4674	0.6275	0.5318	< .01
402	What's Up Amelia?	33	3.6439	0.3646	3.4394	0.3752	0.8083	< .01
403	Heavy Steps	33	2.8485	0.5960	2.9091	0.4993	0.8317	< .01
501	Learning About Mammels	32	2.4219	0.6612	2.6406	0.6348	0.7571	< .01
502	The Wizard Of Oz	32	2.6172	0.7540	2.6172	0.6779	0.8400	< .01
503	Lost In A Storm	17	3.3824	0.7348	2.9412	0.6094	0.5419	< .01
601	Needing A Friend	30	3.1833	0.6226	3.2000	0.5887	0.8373	< .01
602	Famous Flights	30	2.2583	0.6314	2.5250	0.5348	0.8802	< .01
603	The Last Of The Light	33	2.6439	0.7655	2.7045	0.5641	0.8524	< .01
701	The Sinking Of The Titanic	33	2.9091	0.6898	2.9318	0.5938	0.9237	< .01
702	TV Viewing	33	2.6288	0.6647	2.9167	0.5291	0.4757	< .01
703	Best Friends	30	3.3083	0.6148	3.4333	0.5331	0.8606	< .01
	All Prompts	646	2.9957	0.8347	3.0147	0.7548	0.8761	< .01

(produced by Gary Badarak-Research and Evaluation)

Table 5

Percentage of Agreement Between Readers at RUSD on LAPA
and the Publisher's Scoring Service

		Response to Reading			Management of Content			Command of Language		
Prompt	N	Exact Agreement	Adjacent Agreement	Differences Requiring Resolution	Exact Agreement	Adjacent Agreement	Differences Requiring Resolution	Exact Agreement	Adjacent Agreement	Differences Requiring Resolution
101	33	85	15	0	100	0	0	58	42	0
102	33	82	18	0	94	6	0	88	12	0
103	31	35	65	0	97	3	0	81	19	0
201	29	52	45	3	55	42	3	69	28	3
202	29	69	31	0	62	35	3	45	48	7
203	33	61	39	0	73	27	0	70	30	0
301	33	94	6	0	82	18	0	85	15	0
302	33	73	27	0	76	24	0	79	21	0
303	33	79	21	0	91	9	0	79	18	3
401	23	78	22	0	52	48	0	57	39	4
402	33	91	9	0	76	24	0	70	30	0
403	33	73	27	0	79	21	0	79	21	0
501	32	63	37	0	63	37	0	69	25	6
502	32	84	16	0	72	25	3	78	22	0
503	17	47	47	6	47	41	12	59	35	6
601	30	90	10	0	77	23	0	90	10	0
602	30	70	30	0	70	30	0	70	30	0
603	33	70	27	3	70	30	0	61	39	0
701	33	76	24	0	91	9	0	73	27	0
702	33	45	39	15	48	46	6	70	27	3
703	30	80	20	0	90	10	0	77	23	0

Table 6

Cramer's V for Reader by Score
Contingency Tables (LAPA)
(RUSD)

Prompt	Response to Reading	Management of Content	Command of Language
101	.20	.15	.18
102	.14	.12	.17
103	.13	.13	.15
201	.15	.18	.20
202	.24	.15	.17
203	.13	.20	.18
301	.17	.13	.19
302	.13	.16	.15
303	.18	.17	.16
401	.19	.16	.17
402	.16	.15	.13
403	.16	.18	.18
501	.19	.15	.20
502	.18	.15	.16
503	.22	.21	.27
601	.22	.18	.19
602	.17	.19	.16
603	.18	.19	.20
701	.54	.53	.53
702	.19	.25	.28
703	.53	.53	.53

Table 7

Correlations of Stanford Reading Comprehension with LAPA
RUSD

Prompt	N	Response to Reading	Management of Content	Command of Language	Weighted Average
101	724	.34	.20	.40	.35
102	709	.39	.22	.41	.38
103	723	.28	.06	.34	.24
201	830	.35	.42	.50	.47
202	660	.47	.51	.53	.55
203	777	.42	.45	.47	.49
301	761	.25	.30	.36	.32
302	675	.27	.38	.41	.40
303	797	.28	.27	.41	.36
401	741	35	.40	.49	.46
402	843	.35	.44	.49	.48
403	796	.26	.26	.37	.32
501	823	.37	.40	.47	.45
502	740	.30	.40	.47	.44
503	693	.37	.45	.48	.48
601	749	.27	.32	.34	.34
602	732	.36	.36	.44	.44
603	677	.43	.50	.54	.56
701	710	.27	.43	.48	.47
702	443	.31	.44	.55	.53
703	507	.41	.47	.49	.51

Table 8

Inter-dimensional Correlations by Prompt for LAPA
PUBLISHER'S FIELD TEST

101	RR	MC	CL
RR	1.00	.38	.78
MC		1.00	.26
CL			1.00

102	RR	MC	CL
RR	1.00	.43	.67
MC		1.00	.38
CL			1.00

103	RR	MC	CL
RR	1.00	.26	.40
MC		1.00	.39
CL			1.00

201	RR	MC	CL
RR	1.00	.77	.62
MC		1.00	.71
CL			1.00

202	RR	MC	CL
RR	1.00	.81	.72
MC		1.00	.69
CL			1.00

203	RR	MC	CL
RR	1.00	.77	.61
MC		1.00	.63
CL			1.00

301	RR	MC	CL
RR	1.00	.66	.63
MC		1.00	.68
CL			1.00

302	RR	MC	CL
RR	1.00	.88	.74
MC		1.00	.78
CL			1.00

303	RR	MC	CL
RR	1.00	.72	.65
MC		1.00	.74
CL			1.00

401	RR	MC	CL
RR	1.00	.84	.67
MC		1.00	.69
CL			1.00

402	RR	MC	CL
RR	1.00	.76	.43
MC		1.00	.56
CL			1.00

403	RR	MC	CL
RR	1.00	.82	.60
MC		1.00	.65
CL			1.00

501	RR	MC	CL
RR	1.00	.68	.52
MC		1.00	.55
CL			1.00

502	RR	MC	CL
RR	1.00	.67	.54
MC		1.00	.57
CL			1.00

503	RR	MC	CL
RR	1.00	.73	.60
MC		1.00	.66
CL			1.00

601	RR	MC	CL
RR	1.00	.70	.47
MC		1.00	.68
CL			1.00

602	RR	MC	CL
RR	1.00	.63	.56
MC		1.00	.58
CL			1.00

603	RR	MC	CL
RR	1.00	.84	.58
MC		1.00	.62
CL			1.00

701	RR	MC	CL
RR	1.00	.68	.59
MC		1.00	.74
CL			1.00

702	RR	MC	CL
RR	1.00	.81	.65
MC		1.00	.76
CL			1.00

703	RR	MC	CL
RR	1.00	.82	.58
MC		1.00	.64
CL			1.00

Table 9

Characteristics of Prompt Distributions
Weighted Average Scores on LAPA
RUSD

Prompt	Mean	Median	S.D.	Kurtosis	Skewness
101	3.67	3.75	.61	15.39	-3.54
102	3.60	4.00	.73	7.56	-2.58
103	3.62	4.00	.72	12.41	-3.29
201	2.55	2.50	.82	-.59	.01
202	2.43	2.25	.86	-.53	.01
203	2.48	2.50	.90	-.69	-.03
301	3.06	3.25	.90	2.19	-1.37
302	3.01	3.25	.71	2.19	-1.10
303	2.51	2.50	.75	.10	- .19
401	2.89	3.00	.73	-.01	-.48
402	2.93	3.00	.75	.12	-.49
403	2.67	2.75	.74	-.13	-.28
501	2.40	2.50	.83	.97	-.58
502	2.53	2.50	.85	-.69	.01
503	2.60	2.50	.81	-.24	-.25
601	2.68	2.75	.82	-.175	-.31
602	2.42	2.50	.77	.10	-.32
603	2.54	2.50	.75	.02	-.09
701	2.61	2.50	.81	.59	-.44
702	2.52	2.50	.72	-.48	.26
703	2.67	2.75	.81	.03	-.29

Table 10

PERCENT OF STUDENTS PASSING* ON
LAPA (WEIGHTED SCORE)

Lep/Fep Re-Designation(Weighted Raw Score Units)

GRADE	PROMPT			PERCENT PASSING (BY PROMPT)		
	01	02	03	01	02	03
1	3.75	3.75	3.75	75	71	72
2	2.25	2.25	2.25	66	60	62
3	3.00	3.00	2.25	69	68	69
4	2.75	2.75	2.50	69	68	63
5	2.25	2.25	2.25	65	64	71
6	2.50	2.25	2.25	62	65	72
7	2.50	2 25	2.25	60	69	72

Elementary Competency (Weighted Raw Score Units)

GRADE	PROMPT			PERCENT PASSING (BY PROMPT)		
	01	02	03	01	02	03
4	2.50	2.50	2.25	75	74	76
5	2.00	2.00	2.00	82	77	84

* Cutoffs for LEP/FEP Redesignation were selected to approximate a local percentile of 36. For Elementary Competency, cutoffs were selected to approximate a local percentile of 23.

Table 11

PERCENT OF STUDENTS PASSING* ON
LAPA (WEIGHTED SCORE)
(Using .20; .50; .30 Weighting Scheme)

Lep/Fep Re-Designation(Weighted Raw Score Units)

	PROMPT			PERCENT PASSING (BY PROMPT)		
GRADE	01	02	03	01	02	03
1	3.70	3.60	3.70	75	71	72
2	2.20	2.00	2.10	66	60	63
3	3.00	3.00	2.30	67	63	63
4	2.80	2.70	2.50	60	68	63
5	2.20	2.20	2.30	65	64	62
6	2.30	2.20	2.30	62	65	59
7	2.40	2.30	2.40	60	64	63

Elementary Competency (Weighted Raw Score Units)

	PROMPT			PERCENT PASSING (BY PROMPT)		
GRADE	01	02	03	01	02	03
4	2.30	2.30	2.10	77	76	76
5	2.00	2.00	2.00	82	76	81

* Cutoffs for LEP/FEP Redesignation were selected to approximate a local percentile of 36. For Elementary Competency, cutoffs were selected to approximate a local percentile of 23.

Chart 1

Equating Design for Vertical Equating of LAPA
(adjustments made for all grade-to-grade combinations)

Grade	Prompt Combinations
2	101,201; 101,202; 101,203 102,201; 102,202; 102,203 103,201; 103,202; 103,203
.	.
.	.
.	.
.	.
7	601,701; 601,702; 601,703 602,701; 602,702; 602,703 603,701; 603,702; 603,703

Equating Design for Vertical Equating of LAPA
(linked equating)

Grade	Prompt Combinations
2	101,201; 102,202; 103,203
.	.
.	.
.	.
.	.
7	601,701; 602,702; 603,703

Table 12

Prompt By Ability (Reading Comprehension)Interaction Effects
(LAPA and Stanford 8th Edition, Abbreviated)
RUSD

Grade	Significance of Interaction (Prompt X Ability)
1	.16
2	.21
3	.78
4	.24
5	.02*
6	.002**
7	.98

* $p < .05$

** $P < .01$

Table 12 (cont.)

Prompt By Ability (Reading Comprehension)Interaction Effects
(LAPA and Stanford 8th Edition, Abbreviated)
RUSD

CELL MEANS FOR SIGNIFICANT INTERACTIONS

Grade	Level	Prompt 501 X	501 N	502 X	502 N	503 X	503 N
5	1	1.96	(100)	1.94	(88)	1.88	(67)
	2	2.33	(377)	2.41	(364)	2.43	(314)
	3	2.70	(241)	2.82	(226)	2.94	(249)
	4	3.33	(77)	3.38	(58)	3.14	(59)
	Total	2.49	(795)	2.55	(736)	2.62	(689)
		601 X	601 N	602 X	602 N	603 X	603 N
6	1	2.20	(55)	1.81	(55)	1.92	(76)
	2	2.51	(353)	2.26	(328)	2.38	(310)
	3	2.96	(250)	2.70	(236)	2.76	(199)
	4	3.01	(87)	2.85	(106)	3.29	(91)
	Total	2.70	(745)	2.45	(725)	2.56	(676)

REFERENCES

Angoff, W. H. Scales, Norms, and Equivalent Scores. (1971) In R. Thorndike (Ed.), Educational Measurement, Second Edition. Washington: American Council on Education.

Baxter, G.P., Shavelson, R.J., Goldman, S.R., Pine, J. (1990) Evaluation of Procedure-Based Scoring for Hands-On Science Assessment. White Paper. (Obtained from Shavelson, U.C. Santa Barbara).

Beck, M.D. Authentic Assessment for Large-Scale Accountability Purposes: Balancing the Rhetoric. Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, 1991.

Bennett, R.E., Rock, D.A., and Wang, M. (1991) Equivalence of Free-Response and Multiple-Choice Items. Journal of Educational Measurement, V28, 1, 77-92.

Campbell, D.T., and Fiske, D.W. (1959) Convergent and Discriminant Validation by the Multitrait-multimethod Matrix. Psychological Bulletin, V56, 81-105.

Charney, Davida. (1984) The Validity of Using Holistic Scoring to Evaluate Writing: A Critical Overview. Research in the Teaching of English. V18, 65-81.

Dodd, B.G., Koch, W.R., De Ayala, R.J. Validity of Combined Essay and Multiple Choice Achievement Scores: Classical Versus IRT Approaches. Paper Presented at the Annual Meeting of the American Educational Research Association, San Francisco, March, 1989.

Farr, Roger and Farr, Beverly. (1991) Technical Report, Integrated Assessment System. The Psychological Corporation. San Antonio, Texas.

Ferrett, R. T. (1991) Integrating Reading and Writing. Thrust for Educational Leadership, Oct. 1991, 38-41.

Freedman, S.W. (1979) How Characteristics of Students' Essays Influence Teachers' Evaluation. Journal of Educational Psychology. V71, 328-38.

Harris, D.J. (1991) A Comparison of Angoff's Design I and Design II for Vertical Equating Using Traditional and IRT Methodology. V28, 3, 221-235.

Huot, Brian. (1990) Reliability, Validity, and Holistic Scoring: What We Know and What We Need to Know. College Composition and Communication. V41, 2, 201-258.

Huot, Brian. (1990) The Literature of Direct Writing Assessment: Major Concerns and Prevailing Trends. Review of Educational Research. V60, 2, 237-263.

Kolen, M.J. (1991) Smoothing Methods for Estimating Test Score Distributions. Journal of Educational Measurement. V28, 3, 257-282.

Linn, R.L., Baker, E.L., and Dunbar, S.B. (1991) Complex, Performance-Based Assessment: Expectations and Validation Criteria. Educational Researcher, November 1991, 15-21.

Lord, F.M. (1982) The Standard Error of Equipercentile Equating. Journal of Educational Statistics, V7, 3, 165-174.

Mehrens, W.A. (1991) Using Performance Assessment for Accountability Purposes: Some Problems. Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, 1991.

Meredith, V.H. (1984) Issues in Direct Writing Assessment: Problem Identification and Control. Educational Measurement: Issues and Practice. Spring, 1984, 11-15.

Mullis, I.V. (1984) Scoring Direct Writing Assessments: What Are the Alternatives? Educational Measurement: Issues and Practice. Spring, 1984, 16-18.

Muraki, Eiji. (1990) Fitting a Polytomous Item Response Model to Likert-Type Data. Applied Psychological Measurement. V14, 1, 59-71.

Popham, James. (1991) Interview on Assessment Issues with James Popham. Educational Researcher, March 1991, 24-27.

Quellmalz, E.S. (1984) Toward Successful Large-Scale Writing Assessment: Where are we now? Where do we go from here? Educational Measurement: Issues and Practice. Spring, 1984, 29-32.

Shavelson, R.J., Mayberry, P., & Li, W., & Webb, N.M. (1990) Generalizability of military performance measurements: Marine corps infantryman. Military Psychology.

Shavelson, R.J., Carey, N.B., Webb, N.M. (1990) Indicators of Science Achievement: Options for a Powerful Policy Instrument. Phi Delta Kappan. May, 1990, 692-697.

Shavelson, R.J., Baxter, G.P., and Pine (1990). What alternative assessments look like in science. Paper presented at Office of Educational Research and Improvement Conference, The Promise and Peril of Alternative Assessment, Washington, D.C.

Shepard, Lorrie. (1991) Interview on Assessment Issues with Lorrie Shepard. Educational Researcher, March 1991, 21-27.

Swanson, D., Norcini, J. & Grosso, L. (1987). Assessment of clinical competence: Written and computer-based simulations. Assessment and Evaluation in Higher Education, 12, 220-246.

The Psychological Corporation (1989). Stanford Achievement Test, 8th Edition, Abbreviated.

The Psychological Corporation (1990). Integrated Assessment System: Language Arts Performance Assessment.

Webb, N.M., Rowley, G.L., and Shavelson, R.J. (1988) Using Generalizability Theory in Counseling and Development. Measurement and Evaluation in Counseling and Development, 21, 81-90.

Williams, P.L., Phillips, G.W., and Yen, W.M. (1991) Measurement Issues in High Stakes Performance Assessment. Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, 1991.

Wilson, Mark. (1991) Implications of New Perspectives on Student Assessment for Chapter I and its Evaluation: Educational leverage from a Political Necessity. Paper presented at the annual meeting of the California Educational Research Association, San Diego, November, 1991.