Resource Consultant Training Program
Research Report No. 10

RCTP

Measurement Precision for Screening-Eligibility Decisions: An Application of Writing CBM

Richard Parker
Gerald Tindal
Jan Hasbrouck

University of Oregon, Division of Teacher Education, Special Education Area,
Eugene, Oregon, 97403-1215

Published by
Resource Consultant Training Program
Division of Teacher Education
College of Education
University of Oregon

Parker, Richard, Tindal, Gerald, & Hasbrouck, Jan
Measurement Precision for Screening-Eligibility Decisions:
An Application of Writing CBM
Research Report No. 10

Staff

Gerald Tindal, Program Director
Jerry Marr, Editor
Denise Styer
Donna Jost
Clarice Skeen
Mike Rebar

Acknowledgments

Preparation of this document was supported in part by the U.S. Department of Education, grant numbers G008715106-89 and G008715710-89. Opinions expressed herein do not necessarily reflect the position or policy of the U.S. Department of Education, and no official endorsement by the Department should be inferred.

Cover Design: George Beltran

Measurement Precision for Screening-Eligibility Decisions: An Application of Writing CBM

Richard Parker
Gerald Tindal
Jan Hasbrouck
University of Oregon

Abstract

Five countable indices of writing quality were examined for suitability in making special education screening-eligibility decisions. Writing samples for 2,160 students in Grades 2 through 11 from two school districts were collected and four analyses were performed. Histograms and percentile line graphs with standard error bands were used to examine the sensitivity of the writing indices around potential screening cut-off points. In addition, criterion validity within each grade level was assessed by correlating the writing indices with teachers' holistic judgments of writing quality. Of the five indices, the percent of words spelled correctly in a three-minute writing sample (%CSWd) showed greatest measurement sensitivity for screening applications. If Grade 2 were excluded, the percent of correct word sequences (%CWSeq) could be recommended as an alternative scoring method. Standard errors of measurement were large; scores 30 to 40 percentile points apart could not be reliably differentiated. Use of the writing indices for screening-eligibility decisions must acknowledge this large amount of measurement error.

The expressive writing of students with mild learning handicaps may display several kinds of deficiencies. Problems include illegibility (Tindal & Parker, 1989), mechanical errors (Thomas, Englert & Gregg, 1987), inability to conform to a topic (Englert & Thomas, 1987), inability to produce a cohesive story (Barenbaum, Newcomer, & Nodine, 1987), inability to use organizing strategies (Englert, Raphael, Fear, & Anderson, 1988), and low productivity (Nodine, Barenbaum, & Newcomer, 1985). Writing has not been an instructional priority for these students (Leinhardt, Zigmond, & Cooley, 1980), and also has been neglected in their IEPs (Schenck, 1981).

A logical reason for this lack of instructional emphasis is the paucity of efficient, classroom-based assessment tools (Isaacson, 1985; Phelps-Gunn & Phelps-Terasaki, 1982). Whereas informal reading inventories are widely accepted as classroom assessment tools for reading, no comparable procedures exist for writing assessment. Writing assessment tools are needed for three purposes in special education: (a) screening-eligibility, (b) diagnosis for individualized program planning, and (c) progress monitoring and evaluation (Moran, 1987). Screening-eligibility decisions require less measurement sensitivity than do progress monitoring decisions. Similarly, screening-eligibility decisions do not demand that a test be diagnostically or instructionally useful. For these reasons, screening-eligibility is probably the easiest application to justify for a new assessment procedure.

In response to the need for acceptable, efficient classroom-based writing assessment tools a program of research was conducted at the University of Minnesota's Institute for Research in Learning Disabilities (IRLD) in the early 1980s. The IRLD research centered on the validity, reliability, and efficiency of three countable features of students' writing samples. Nine studies (see Marston, 1989) were conducted on "total words written" (TotWd), "correctly spelled words" (CSWd), and "correct word sequences" (CWSeq), all of which were produced in creative writing samples from a "story starter" collected within a 3-minute period. Two later studies at the University of Oregon reported on these three indices and two percent-based variations: "percent of correctly spelled words" (%CSWd) and "percent of correct word sequences" (%CWSeq), also based on a 3-minute timed sample. The suitability of these five indices for special education screening-eligibility decisions is the focus of this report.

To use countable writing indices for special or remedial education screening-eligibility decisions requires (a) comparison standards or norms based on students in the regular program, and (b) a cut-off score for identifying those students who warrant more detailed individual assessment (Elliot & Bretzing, 1980; Kamphaus & Lozano, 1984). For stable district norms, large student samples (at least 100 per grade level) are required across several grades (Salvia & Ysseldyke, 1988). The present study includes samples of over 400 students at each of Grades 2-5, and nearly 100 students each in Grades 6 and 8.

In addition to stability, normative distributions must show sensitivity, i.e. suitable dispersion (rather than clustering) of students. Usually, a normal bell-shaped curve is desired (Tindal, 1989). More specifically, the instrument should show sensitivity (i.e., scores should be dispersed) at the segment of the score scale where decisions must be made (Salvia & Ysseldyke, 1988). For screening-eligibility use, the critical part of the scale lies near the cut-off used to identify low-performing students requiring further individual assessment. Typically, this cut-off score is around the 35th percentile for remedial programs, and lower for programs serving students with severe learning disabilities. A grade distribution which is positively skewed (a large cluster of lower scores) will be relatively insensitive to screening out high-risk students. Visual displays help us examine the clustering or dispersion of scores in the cut-off area. A histogram with a superimposed normal curve (derived from the distribution means and standard deviations) clearly depicts deviations from normality and score dispersion at various parts of the score scale (Freedman, Pisani, & Purves, 1980).

Test sensitivity around the cut-off score can be described more precisely in terms of how well two low scores (e.g., the 20th and 30th percentiles) can be differentiated within a grade level. Especially for younger students, within-grade level screening appears to be common practice. Therefore, within- rather than across-grade score differentiation should be emphasized (Brown, 1983). Cross-grade sensitivity, reliability, and validity estimates are generally inflated over within-grade estimates because cross-grade scores are more widely dispersed on a broader scale (Brown, 1983; Cronbach, 1984). Therefore, indices of test sensitivity or reliability calculated across grades are not directly applicable to screening-eligibility decisions made within a grade level (Sabers, Feldt, & Reschly, 1988). For depicting score differentiation within a grade (and across grades), the percentile line graph is unequalled (Cleveland, 1985). If score differentiation across grade levels is desired, multiple grades can be plotted together on the same graph.

Although the overlap and proximity of neighboring percentile scores is informative, more precision in judging test sensitivity is gained with the standard error of measurement (SEm), which provides a band of confidence around individual scores (Lord, 1984). The SEm formula includes a test reliability coefficient (e.g. inter-scorer, internal consistency, test-retest, parallel form, stability) which is related to the test's intended use (Brown, 1983). Screening-eligibility decisions assume either parallel form or test-retest reliability. Of

these two, test-retest reliability is the more defensible because parallel form reliability is difficult to apply to creative writing samples. If a cut-off score at the 30th percentile has a wide SEm of 20 percentile points, we can be reasonably certain (with 68% confidence) that students at the cut-off score have true scores between the 10th and 50th percentiles. This degree of uncertainty reflects an inefficient screening device, as we would be likely to "miss" too many low scoring students and include in our screening net too many high scoring students. SEm bands can be applied directly to percentile line graphs.

Although not the primary focus of this paper, any writing screening test must also be valid for a particular purpose or decision. Several validation criteria are available, including other formal and informal writing measures, prediction of concurrent or future special program placement, and teachers' holistic judgments of student writing quality. Of these validation criteria, teachers' holistic ratings are pre-eminent when classroom acceptability of new assessment procedures is a major concern. Among the reasons for teacher rejection of standardized tests for decision making is the lack of agreement between test results and teachers' holistic judgments (Burry, Catterall, Choppin, & Dorr-Bremme, 1982; Salmon-Cox, 1981; Sproull & Zubrow, 1981).

The countable writing indices developed at the IRLD have been recommended for screening-eligibility purposes in a number of publications (Shinn, 1988; Shinn, 1989; Shinn, Ysseldyke, Deno & Tindal, 1986; Shinn & Marston, 1985). These and other articles related to CBM for special/remedial education identification were reviewed by Shinn, Tindal, & Stein (1988). Only one IRLD study (Tindal, Marston, & Deno, 1983) was based on samples sufficiently large to produce stable distribution shapes. That study included approximately 95 students at each of Grades 1-6, but score distributions were not analyzed. None of the reviewed articles directly addressed the problem of score dispersion around a screening cut-off. Nor did any of the reviewed articles examine the sensitivity of these indices for differentiating among low scores within a grade level.

The most recent procedural article on norming with CBM for screening-eligibility (Shinn, 1989) emphasized score comparisons across CBM indices, and across grades for a single index. Mean scores were found to increase from one grade to the next. Within-grade score comparisons were limited to the 25th, 50th, and 75th percentiles. The SEm error band of these scores was not considered. In a second article on the use of decision-making from CBM district norms, Tindal (1989) noted that the particular metric used for a curriculum based measure has a great impact on the measure's sensitivity, and that bell-shaped distributions without ceiling effects were commonly noted with rate-based indices beyond Grade 1. Although he warned that "scoring systems must be analyzed by their effect on the distribution of scores" (p. 211), no guidelines for analysis were offered.

Studies on the reliability of countable writing indices have some relevance to the problem of screening-eligibility. Only two studies (Shinn, 1981; Tindal, Germann, & Deno, 1983) calculated reliabilities *within* rather than across grade levels. These studies yielded retest coefficients of .51 to .71 for TotWds, .52 to .74 for CSWd, and .55 to .73 for CWSeq. These data are valuable in calculating SEm bands around obtained scores.

Among the writing validation studies summarized by Marston (1989), one is directly relevant to the present investigation, as teachers' holistic judgments were used as a concurrent criterion measure (Videen, Deno, & Marston, 1982). Correlations of .85 and .84 with holistic ratings were found for TotWd and CSWd, respectively, but those results were based on a *cross-grade* (Grades 3-6) sample of 50 students, and are not directly applicable to decision making *within* a grade level (Sabers, Feldt, & Reschly, 1988). Schools need within-grade validation evidence based on teachers' judgments of writing quality.

In summary, the countable writing indices developed and studied at the University of Minnesota IRLD are often recommended for screening-eligibility decisions (Shinn, 1989; Shinn, Tindal, & Stein, 1988). However, the research in support of this application is limited in several important respects. First, distribution shapes using sufficiently large, stable student samples have not been studied. Second, the sensitivity of these indices to score differences in the critical region around the cut-off score has not been investigated. Third, reliability and validity indices typically have been calculated across rather than within grade levels. Finally, teachers' holistic judgments rarely have been used to help validate the countable writing indices to ensure their classroom acceptability.

This paper directly addresses these four deficiencies. Results are presented for five countable indices of writing quality from the Universities of Minnesota (TotWd, CSWd, CWSeq) and Oregon (%CSWd, %CWSeq). Scores from two studies totaling 2,160 elementary and secondary writing samples were analyzed to help ascertain the utility of the five indices for making special or remedial education screening-eligibility decisions.

First, mean score comparisons provided gross measures of score increase over a single year and from one year to the next. For a more fine-tuned analysis, histograms displayed score dispersion in the lower segment of the distribution. A more detailed analysis of the differentiation of individual percentile scores was provided through percentile line graphs. The greatest precision was obtained when SEm bands were applied to the percentile graphs. Finally, teacher holis-

tic judgments on the quality of each writing sample were summarized by grade level and correlated with each countable writing index.

METHOD

The Studies

Study #1 (Grades 2-5)

The largest study was conducted in 20 elementary schools within two west coast school districts, one rural and one suburban, both located in lower-middle SES communities. During October and May writing samples were collected within both districts from 79 randomly selected Grade 2 (n = 449), Grade 3 (n = 575), Grade 4 (n = 447), and Grade 5 (n = 446) classrooms—a total of 3,834 writing samples from 1,917 students. All participating students were in attendance in regular Chapter 1 compensatory and special education programs during the day of the assessment.

Study #2 (Grades 6, 8, 11)

The second set of writing samples was collected in the spring of the following year from middle and high schools only within the rural school district. From two middle schools and one high school, 12 classrooms were sampled at Grade 6 (n = 91), Grade 8 (n = 89), and Grade 11 (n = 63) levels—243 students in all. Students in Chapter 1 compensatory programs were included in the sample, but not those in special education. Writing samples were collected in the spring only.

Procedures

All students completed timed, 6-minute creative writing samples from a story-starter, following a modification of procedures outlined by Videen, Deno and Marston (1982). Representative story starters included Grade 2: "Mr. Brown opened the front door very carefully and ..."; Grade 3: "One day our teacher was sick; we had another teacher and ...", Grade 4: "One day my mom surprised me by bringing home ..."; Grade 5: "Walking slowly downstairs, Greg felt the hairs on the back of his neck stand up ...", and Grade 6; "It was the night before Halloween, and all the students...". At the end of three minutes, students were asked to quickly draw a star on their paper, then continue writing for the remaining 3 minutes. The star allowed the writing samples to be scored in two 3-minute sections, and a form of split-half reliability to be calculated.

Objective Scoring

The writing samples were analyzed both subjectively, using teachers' holistic judgments, and objectively, using five different countable indices, defined as follows:

1. *Total Words written (TotWd).* The sum of all word-like units containing letters physically grouped together; correct spelling, usage, and syntax were disregarded. Symbols and numbers were not counted as words.

2. *Correctly Spelled Words (CSWd).* The sum of all words spelled correctly; homonyms had to be spelled according to the usage in the sentence.

3. *Correct Word Sequences (CWSeq)* (Videen, Deno, & Marston, 1982). The sum of all immediately adjacent, correctly spelled word pairs that are syntactically correct together, given the context of the sentence. At the start and end of sentences, correct beginning and ending punctuation replaced correctly spelled words for scoring purposes.

4. *Percentage of Correctly Spelled Words (%CSWd).* The ratio of the number of words spelled correctly (CSWd) to the total number of words written in the composition (TotWd).

5. *Percentage of Correct Word Sequences (%CWSeq).* The number of correct word sequences (CWSeq) divided by the total number of possible word sequences.

Research conducted at the University of Minnesota IRLD with small student samples supports the reliability and validity of three of the five indices: CWSeq, TotWd, and CSWd (Deno, Marston, & Mirkin, 1982; Marston & Deno, 1981; Marston, Lowry, Deno, & Mirkin, 1981; Videen, Deno, & Marston, 1982). The other two indices, %CSWd and %CWSeq, are counterparts of CSWd and CWSeq which are not influenced by the length of the writing sample (Tindal & Parker, 1989; Parker, Tindal, & Hasbrouck, in press).

Scoring Procedures

Objective scoring was completed by four graduate students in education after a 2-hour training session. Scorers were blind to student names, schools, grade levels, and program placement. Reasonable interrater agreement (Pearson r) was reached on representative groups of 30 papers selected within grade levels: Tot.Wd: .99; CSWd: .98; CWSeq: .87; %CSWd: .98; %CWSeq: .87. After practicing on several papers, scorers required about 6 minutes per paper to compute the five indices.

Holistic Judgments

Writing samples were also holistically rated according to their communicative effectiveness, scaled 1 (very poor) to 7 (very effective), with no intermediate descriptors. Range finders for each point on the scale were chosen separately for each grade on the basis of identical ratings on sample papers by two practicing teachers and two of the authors.

In addition to the range finders, raters were assisted by the following definition of good writing, produced by consensus among a team of four practicing teachers: "Good writing clearly communicates to the reader the ideas/story of the writer. Good writing requires legible handwriting or printing, as well as distinguishable words, phrases, and sentences. Coherent linking of ideas from one sentence to the next also contributes to good writing." (Hasbrouck, 1987, p. 2).

Holistic rating was completed by the same four graduate students after a 1-hour training session which included discussing the definition and range-finders, and obtaining interrater reliability on two sample sets of papers. Interrater agreement on a representative cross-grade set of 30 papers was r = .81, and r = .86 for

30 papers selected within a single grade. After initial practice, holistic rating required less than 1 minute per paper.

Analyses

Five descriptive analyses were completed: (a) Mean scores were compared across grades and from fall to spring within a year; (b) histograms with normal curves superimposed were produced for Grades 2-5 to describe score distributions; (c) percentile ranks for CWSeq and %CWSeq were compared across Grades 2-11; (d) for Grade 5, SEm bands were placed on the percentile graphs; and (e) at each grade level, the five countable indices were correlated with teachers' holistic ratings of writing quality.

Results

Mean Score Differences

We expected basic writing skills to increase over the grades due to both learning and maturation. However, to the extent that students challenge themselves to write with more complex words and sentences, skill improvement may not be reflected in score increases across the grades. Summary statistics were tabulated for Grades 2-5 fall and spring assessments, and the Grades 6, 8, and 11 Spring assessments (see Table 1).

Between fall and spring of any given school year, mean scores increased for each countable index at every grade level. Scores also generally increased from one fall to the next and from one spring to the next. The lone exception is the *decrease* in Grade 4 to Grade 5 spring scores for TotWd, CSWd, and CWSeq; students temporarily regressed before progressing further.

Distribution Shape

A sensitive screening tool shows good score dispersion in the bottom 30-40 percent of the distribution. For Grades 2-5, fall scores for the five indices were plotted as histograms. Score frequencies (vertical axis) were standardized to permit comparisons across indices and years. In addition, a normal curve derived from the distribution mean and standard deviation was superimposed on each histogram, using normal density smoothing (Freedman, Pisani, & Purves, 1980) (see Figure 1).

In the absence of artificially imposed scoring limits, a large sample should produce a normal, bell-shaped distribution. A non-normal, skewed distribution reflects a test that lacks precision in measuring either high or low scores. An undesirable clustering of scores at the lower end of the scale indicates that the test cannot distinguish well among students with neighboring

Table 1. Summary Statistics for Four Countalbe Indices of Writing at Eight Grade Levels

	TotWd		CSWd		CWSeq		%CSWd		%CWSeq	
	M	SD	M	SD	M	SD	M	SD	M	SD
Gr.2—fall	12.4	6.6	8.2	5.7	5.0	5.0	.62	.22	.38	.28
Gr.2—spring	30.9	11.4	27.3	11.6	22.7	12.0	.87	.11	.71	.19
Gr.3—fall	23.7	10.0	19.4	9.8	14.6	9.5	.80	.15	.64	.23
Gr.3—spring	35.5	13.3	32.4	13.3	28.7	13.2	.90	.10	.80	.18
Gr.4—fall	27.6	11.6	24.2	11.1	19.9	10.3	.87	.10	.76	.17
Gr.4—spring	43.6	12.6	40.2	12.6	37.7	13.1	.92	.08	.86	.13
Gr.5—fall	36.1	11.8	32.4	12.1	28.1	12.5	.89	.11	.79	.19
Gr.5—spring	41.5	12.4	38.8	12.5	35.7	12.1	.93	.07	.87	.11
Gr.6—spring	46.2	13.4	42.5	12.7	34.2	11.5	.92	.05	.84	.10
Gr.8—spring	52.2	15.2	48.9	14.3	42.3	13.4	.94	.06	.88	.10
Gr.11—spring	58.5	17.2	55.7	16.9	50.9	16.3	.95	.04	.90	.08

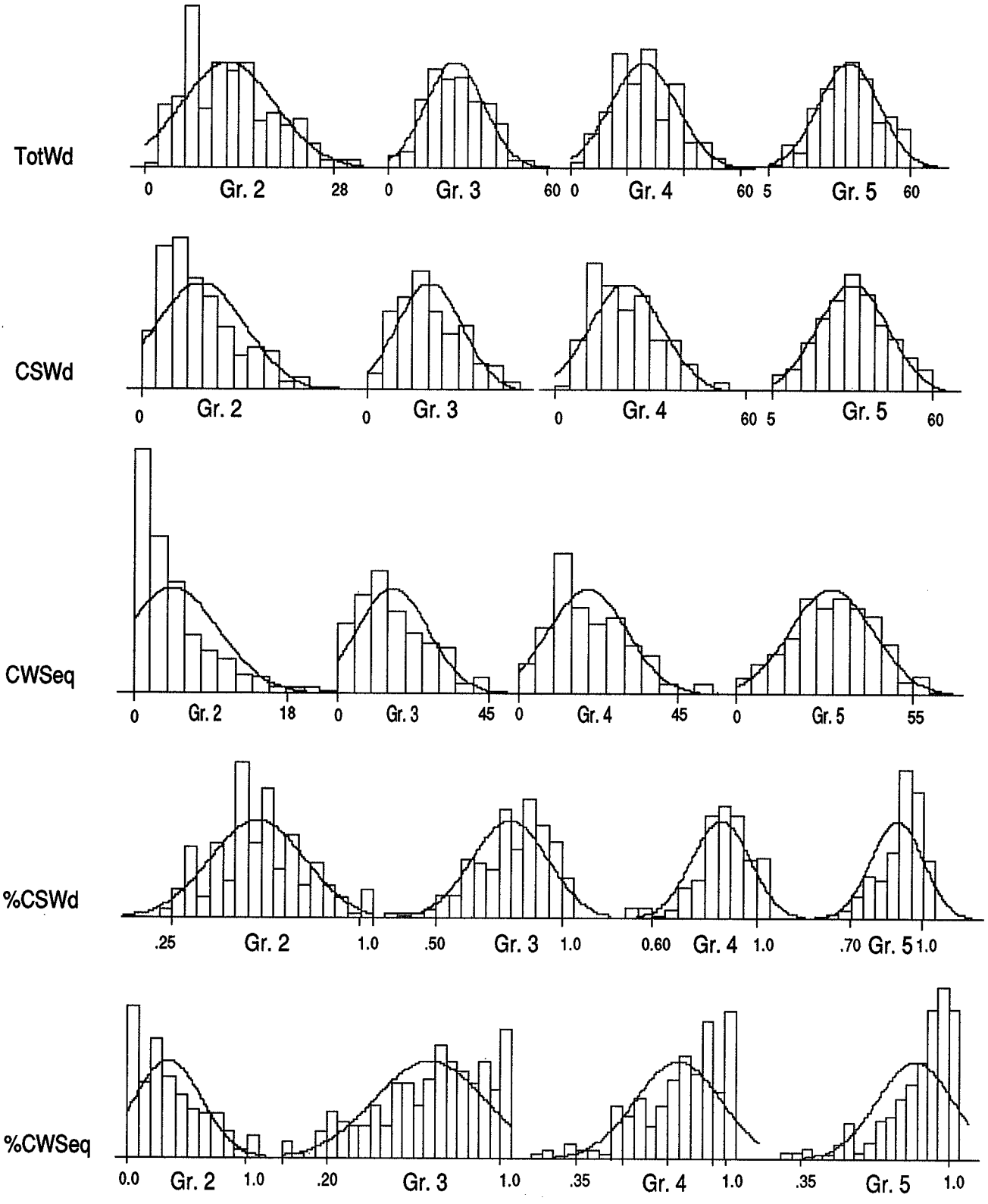

Figure 1. Large-sample Histograms for Five Countable Indices of Writing Quality at Four Grade Levels: Grade 2 (n = 449), Grade 3 (*n* = 575), Grade 4 (*n* = 447), and Grade 5 (*n* = 446).

scores in that region. When students are clustered at the lower end of the score scale (a positively skewed distribution) a small change in the designated cut-off score may reclassify a large number of students.

In Figure 1, the three rate-based indices (TotWd, CSWd, CWSeq) show similar distributions, as do the two percent-based indices (%CSWd and %CWSeq), across the four grades. The rate-based indices differ only in Grades 2 and 3. TotWd shows a small amount of score clustering at the lower scale (0-6 words) in Grade 2, while CWSeq shows extreme skewness at that grade, and CSWd is moderately skewed. CWSeq also shows some undesirable skewness in Grade 3.

Of the two percent-based indices, %CSWd shows desirable distribution of low-range scores at all grade levels, while %CWSeq is suitable for Grades 3-5 only. Both %CSWd and %CWSeq are negatively skewed (score clustering toward the top of the score scale), but that characteristic should not detract from the use of the indices for screening.

Percentile Graphs

A more detailed view of test sensitivity was obtained through percentile line graphs (Cleveland, 1985) (see Figure 2). Percentile scores across seven grade levels were plotted for CWSeq and %CWSeq on two graphs. Percentiles were plotted at 10-point intervals, and intermediate values interpolated. A line slope close to horizontal reflects lack of test sensitivity in that raw score region, while a steep slope reflects high sensitivity. The distance between adjacent lines indicates how well the test distinguishes between students at different grade levels.

For %CWSeq , Grade 2-5 lines are generally steep from the 1st to 90th percentiles, showing good discrimination among low and middle scoring students. The notable exception is the bottom range (1-10 percentile) of Grade 2. %CWSeq shows generally poor discrimination (less slope) at the Grade 6 level and above. %CWSeq also is a poor discriminator above the 90th percentile for Grade 4 and above.

The CWSeq graph shows generally poorer discrimination (less slope) than %CWSeq at the lower grade levels (Grades 2-5) and for low score ranges at all grade levels. CWSeq shows better discrimination in the upper score ranges at all grade levels, and is a stronger measure in Grades 8 and 11 at middle and high score ranges.

In differentiating identical percentile scores between grades, the two indices also perform differently. For CWSeq there is virtually no score overlap from the 10th to 99th percentiles. For CWSeq there is considerable overlap among Grades 4-11. CWSeq clearly outperforms its percentile-based counterpart in grade level discrimination. As was mentioned earlier, however, between-grade score discrimination does not address the needs of within-grade screening.

SEm Bands

Overlap of neighboring percentile scores is a more precise gauge of test sensitivity than skewness of a histogram. Even more precision is gained by applying standard error of measurement (SEm) bands (Lord, 1984) to percentile line graphs (see Figure 2). For the sake of clarity, only the Grade 5 line was bracketed by a shaded ±1 SEm band of confidence. Although neighboring Grade 4 and 6 scores were not shaded, one can imagine a similar shaded band around those lines. We can be reasonably confident (68% certainty) that a student's true score falls within the 1 SEm confidence band. For example, a Grade 5 obtained CWSeq score of 28 (50th percentile) is probably between 21 and 34. We can also say with 68% certainty that two scores are different if their confidence bands do not overlap (Salvia & Ysseldyke, 1988). For example, confidence bands for Grade 5 CWSeq 20th and 50th percentile scores do overlap slightly, so these two percentile scores cannot be distinguished with reasonable confidence. In calculating confidence bands, the rather optimistic reliability estimate of .75 was used. As was noted earlier, within-grade test-retest reliabilities were generally in the .50 to .75 range.

Score discrimination for CWSeq at Grade 5 is poor in the lower percentiles. Tenth and 40th percentile SEm bands overlap, indicating an inability to distinguish between these two scores with reasonable confidence. For Grades 2-4, CWSeq shows even less sensitivity, while for the upper grades (Grade 6, 8, & 11) the index differentiates scores somewhat better.

%CWSeq shows greater sensitivity in discriminating among scores in the lower percentiles; the 10th and 40th percentiles can be differentiated with reasonable certainty. However, the 10th and 30th percentiles cannot be reliably differentiated. The sensitivity of %CWSeq is greatest at the low extreme of the scale; the 1st and 5th percentiles can be differentiated, as can the 5th and 20th percentiles. %CWSeq shows greatest sensitivity at the lower half of the scale, in Grades 2-5. Sensitivity is lost in Grades 6, 8, and 11, although at these upper grades %CWSeq is still more sensitive than CWSeq for low scores. Percentile line graphs were also produced for CSWd and %CSWd, with results paralleling those for CWSeq and %CWSeq in Figure 2 (see Figure 3). TotWd percentile graphs were also produced, yielding slopes very similar to CSWd (also in Figure 3).

Although between-grade score differentiation is not a focus of this paper, that information is available from the percentile graphs with SEm bands. CWSeq is clearly the better index for discriminating among grade levels, although there is no reliable difference between neighboring grade levels (e.g., Grades 5 and 6 or Grades 5 and 4). Grade 5 and 8 CWSeq scores appear to be differentiated only marginally, although an SEm band should be computed for Grade 8 scores to be certain.

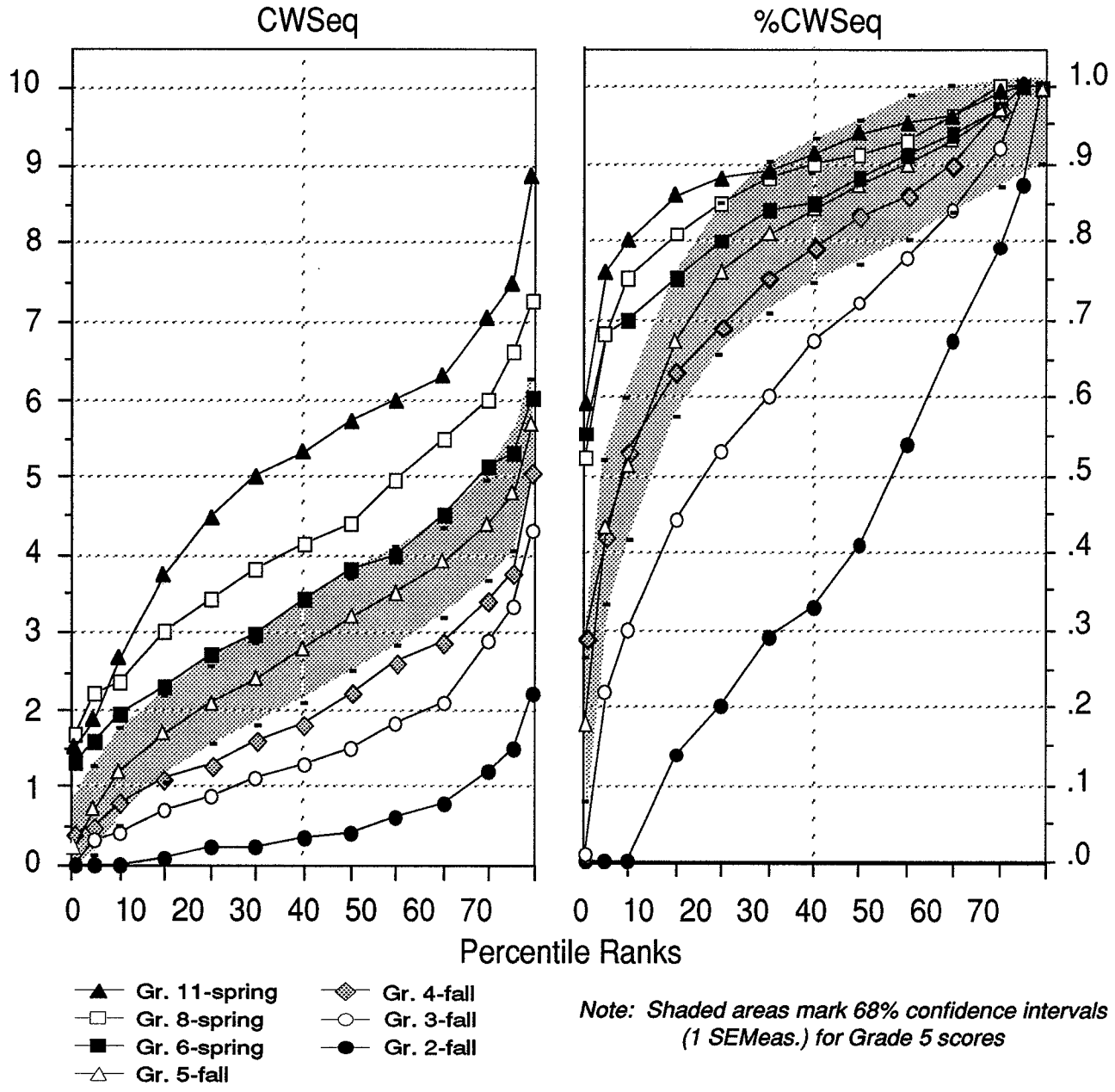

Figure 2. Percentile Line Graphs for "Number of Correct Word Sequences Written" (CWSeq) and "Percent of Correct Word Sequences" (%CWSeq) Across Seven Grades.

For %CWSeq there is no practical differentiation of grades above grade 4 and above the 35th percentile.

Validation by Teachers' Holistic Judgments

Because of the importance of classroom acceptability of any curriculum-based screening tool, teachers' holistic judgments of the quality of each writing sample were correlated with the five countable indices at each of seven grade levels (see Table 2).

Mainly low-moderate size correlations were obtained between the countable indices and teachers' holistic judgments of the same writing samples. This finding was expected, as weak reliability places an artificial ceiling on validity coefficients, and only low-moderate to moderate reliabilities were obtained for both the writing indices and holistic judgments within a grade level. The strongest index overall (CWSeq, r = .56) explained nearly twice the variance as the overall weakest index (TotWd, r = .42). The two percent indices (%CWSeq, %CSWd) showed least predictive strength in Grades 6-11, and no index showed predictive strength in Grade 11. Further generalizations are also grade-specific. TotWd was reasonably strong for Grade 2 only. CSWd and CWSeq were also strong for Grade 2, but maintained moderate-level scores through Grade 8. The two percent indices showed moderate predic-

tive strength for Grades 3-5 only. %CWSeq and %CSWd were uniquely strong in Grade 4 (r = .70 and .67, respectively).

Discussion

This study examined the suitability of five countable indices of writing quality for screening-eligibility decisions, focusing on sensitivity in the low, cut-off score range. Five major findings emerged from analyses. First, using the very rough gauge of mean score increases across grade levels and across assessment periods within a school year, all five indices (TotWd, CSWd, CWSeq, %CSWd, %CWSeq) appeared suitable. Results very similar to these often are published to support the use of countable indices for screening-eligibility decisions (Shinn, 1988; Shinn, 1989).

The second major finding was more cautionary. Examination of grade-level histograms with large student samples indicated the basic unsuitability for screening purposes of three indices at certain grade levels. CSWd, CWSeq, and %CWSeq were unsuitable for use in Grade 2 because of strong positive skewness or clustering of scores at the low end of the scale. For CWSeq, marked skewness also existed in Grade 3.

The third finding emerged from more fine-grained analyses of percentile line graphs. These analyses indicated problems in using the most thoroughly researched writing index from the University of Minnesota IRLD, CWSeq, for screening-eligibility decisions. Percentile line graphs of CWSeq and its newer percent-based counterpart, %CWSeq, clearly demonstrated the weakness of the former for differentiating among low scores. While %CWSeq showed a definite ceiling effect, it showed greater measurement sensitivity in the lower half of the score distribution. Although not presented in this paper, parallel findings emerged from analyses of the other rate and percent pair, CSWd and %CSWd. CBM indices which are rate-based often avoid undesirable ceiling and floor effects (Shinn, Ysseldyke, Deno, & Tindal, 1986). However, in this case, rate-based indices showed least sensitivity in that region of the score scale where sensitivity was most critical.

The fourth finding, derived from SEm bands on percentile line graphs, was that neither CWSeq nor %CWSeq could reliably distinguish among neighboring percentile scores near the bottom of the score scale. CWSeq lacked the sensitivity to differentiate scores 30 percentile points apart (10th to 40th), and %CWSeq was insensitive to a 20 percentile point spread (10th to 30th). In the upper grades (Grade 6, 8, 11) measurement precision at the low end of the score scale was even weaker. This lack of precision poses a major problem for screening. Using CWSeq and a screening cut-off at the 35th percentile, we are in danger of excluding from our screening net large numbers of low achieving students, including those with true scores as low as the 10th percentile. We are also in danger of including in our screening net large numbers of high achieving

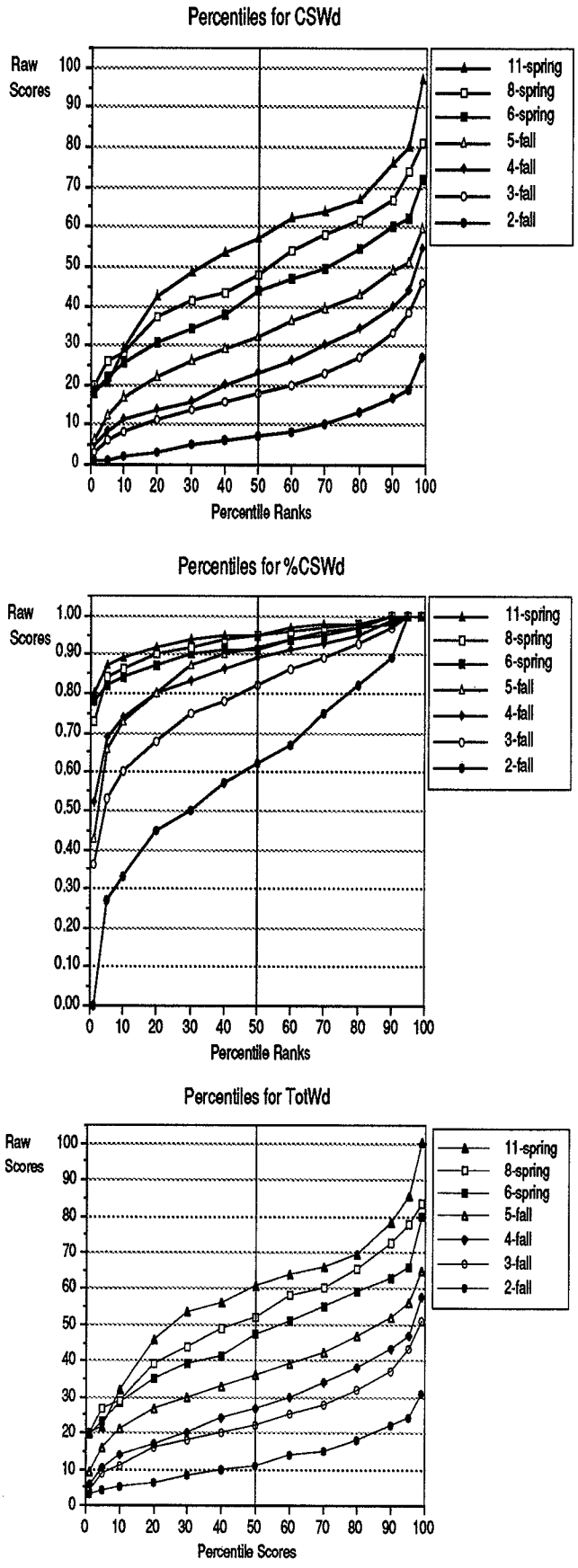

Figure 3. Percentile Graphs for CSWd, %CSWd, and TotWd.

Table 2. Correlations Among Five Countable Feature of Writing Samples and Holistic Ratings of the Same Writing Samples at Seven Grade Levels.

	Countable Indices*				
	TotWd	CSW	CWSeq	%CSWd	%CWSeq
Gr. 2 (N=449)	.49	.64	.60	.48	.43
Gr. 3 (N=575)	.40	.54	.58	.49	.52
Gr. 4 (N=447)	.36	.49	.58	.67	.70
Gr. 5 (N=446)	.44	.56	.61	.55	.55
Gr. 6 (N=92)	.41	.48	.52	.46	.41
Gr. 8 (N=91)	.43	.52	.56	.34	.36
Gr. 11 (N=64)	.39	.43	.48	.46	.42
Average	.42	.52	.56	.49	.48

*All indices significant at p<.01

students, including those with true scores above the 70th percentile. This amount of error defeats the purpose of a screening test.

The fifth finding relates not to measurement sensitivity, but to the agreement between countable indices and teacher holistic judgments. The five indices agreed moderately with teacher judgments, but the strongest index overall (CWSeq) demonstrated nearly twice the explanatory power of the weakest index (TotWd). The size of validity coefficients varied considerably from one grade to the next, notably for Grades 4 and 5. Qualitative changes in writing skill development may exist at those grades.

Which of the five indices should be used for screening-eligibility decisions, i.e., for demarcating within a particular grade level the lowest performing 30-50% of students for more intensive individual writing assessment? For all grades, 2 through 11, the percent of words spelled correctly (%CSWd) can be recommended. %CSWd was reasonably well validated by teachers' holistic judgments, and displayed suitable distribution in the lower score range. If Grade 2 is eliminated, then the percent of correct word sequences (%CWSeq) can also be recommended as an alternative (though more time-consuming) scoring method. Although the number of correct word sequences produced (CWSeq) bore the closest relationship to teachers' holistic judgments, this index showed such lack of sensitivity in the lower score range as to render it unusable for screening-eligibility decision-making. Although TotWd's score distribution was suitable, it produced the lowest validity coefficients, and so is not recommended.

Even the most robust index, %CWSeq, demonstrated a 20 percentile point range of uncertainty, making it only moderately efficient as a screening tool. Therefore, we make the above recommendations with a caveat: Other efficient methods of classroom-based writing assessment should be utilized in addition to these.

This investigation does not address the suitability of any of the five indices for program evaluation, progress monitoring, or skill diagnosis. Each of these purposes entails a separate set of measurement requirements. Progress monitoring in Special Education requires greater measurement precision than screening-eligibility, and, to date, there is not sufficient empirical evidence supporting that use (Parker, Tindal, & Hasbrouck, in press).

A limitation of this investigation is the use of teachers' holistic judgments as the sole validation criterion. The reliability of holistic judgments (and the retest reliability of the five indices) was not as high as we would wish for a validation study. Still, we have argued for the importance of this criterion, given the need for teacher acceptability of classroom-based measures. Results using other validation criteria have been summarized elsewhere (Marston, 1989; Shinn, Tindal, & Stein, 1988).

Although teachers use test scores for a variety of purposes, a single test often is not equally suited to all of these purposes. The suitability of a test for each type of decision must be assayed separately. Compared to other uses, screening-eligibility decision-making does not place high measurement demands on a test. However, given their generally low reliability and undesirable distribution shapes, neither of the rate-based IRLD writing indices, CWSeq or CSWd, can be recommended as screening tools. The third of the three original IRLD writing indices, TotWd, produced low agreement with teacher holistic judgments, and can be recommended for Grade 2 only. The two percent-based indices, %CSWd and %CWSeq offer barely sufficient validity and measurement sensitivity to serve as gross screening measures for written expression. Both of these two indices can be expected to result in a moderate number of mis-classifications. The search for efficient, psychometrically sound classroom-based writing screening tools needs to continue.

References

Barenbaum, E., Newcomer, P., & Nodine, B. (1987). Children's ability to write stories as a function of variation in task, age, and developmental level. *Learning Disability Quarterly, 10,* 175-188.

Brown, F. G. (1983). *Principles of educational and psychological testing* (3rd ed.) New York: Holt, Rinehart & Winston.

Burry, J., Catterall, J., Choppin, B., & Dorr-Bremme, D. (1982). *Testing in the nation's schools and districts: How much? What kinds? To what costs?* CSE Report No. 194, Center for the study of Evaluation, Graduate School of Education, UCLA.

Cleveland, W. S. (1985). *The elements of graphing data.* Monterey, CA: Wadsworth Advanced Books and Software.

Cronbach, L. J. (1984). *Essentials of educational and psychological testing* (3rd. ed.). New York: Holt, Rinehart, & Winston.

Deno, S., Marston, D., & Mirkin, P. (1982). Valid measurement procedures for continuous evaluation of written expression. *Exceptional Children, 48,* 368-371.

Deno, S., Marston, D., & Mirkin, P. (1980). *Relationships among simple measures of written expression and performance on standardized achievement tests* (Research Report No. 22). Minneapolis: University of Minnesota, Institute for Research on Learning Disabilities.

Elliot, S. N., & Bretzing, B. H. (1980). Using and updating local norms. *Psychology in the Schools, 17,* 196-201.

Englert, C. S. & Thomas, C. C. (1987). Sensitivity to text structure in reading and writing: A comparison between learning disabled and non-learning disabled students. *Learning Disability Quarterly, 10,* 93-105.

Englert, C. S., Raphael, T. E., Fear, K. L., & Anderson, L. M. (1988). Students' metacognitive knowledge about how to write informational texts. *Learning Disability Quarterly, 11,* 18-46.

Feldt, L. S., Steffen M., & Gupta, N. C. (1985). A comparison of five methods of estimating the standard error of measurement. *Applied Psychological Measurement, 9,* 351-361.

Freedman, D., Pisani, & Purves, (1980). *Statistics.* New York: W. W. Norton & Co.

Gardner, E. F., Rudman, H. C., Karlsen, B., Merwin, J. C. (1984). *Stanford achievement test* (7th ed.). Cleveland: Psychological Corporation.

Helton, G. B., Workman, E. A., & Matuszek, P. A. (1982). *Psychoeducational assessment: Integrating concepts and techniques.* New York: Grune & Stratton, Inc.

Hammill, D., & Larsen. S. (1983). *Test of Written Language.* Austin, TX: Pro-Ed.

Hasbrouck, J. (1987). *Training manual for direct, objective scoring of writing samples.* Resource Consultant Training Program Module No. 2. Eugene, OR: University of Oregon.

Isaacson, S. (1985). Assessing written language skills. In C. S. Simon (Ed.), *Communication skills and classroom success: Assessment methodologies for language-learning disabled students* (pp. 403-424). San Diego: College-Hill Press.

Kamphaus, R. W. & Lozano, R. (1984). Developing local norms for individually administered tests. *School Psychology Review, 13,* 491-498.

Leinhardt, G., Zigmond, N., & Cooley, W. W. (April, 1980). *Reading instruction and its effects.* Paper presented at the annual meeting of the American Educational Research Association, Boston, MA.

Lord, F. M. (1984). Standard errors of measurement at different ability levels. *Journal of Educational Measurement, 21,* 239-243.

Marston, D., & Deno, S. (1981). *The reliability of simple, direct measures of written expression* (Research Report No. 50). Minneapolis: University of Minnesota Institute for Research on Learning Disabilities.

Marston, D., Lowry, L., Deno, S., & Mirkin, P. (1981). *An analysis of learning trends in simple measures of reading, spelling, and written expression: A longitudinal study* (Research Report No. 49). Minneapolis: University of Minnesota Institute for Research on Learning Disabilities.

Marston, D. (1982). *The technical adequacy of direct, repeated measurement of academic skills in low achieving elementary students.* Unpublished doctoral dissertation, University of Minnesota, Minneapolis.

Marston, D. (1989). A curriculum-based measurement approach to assessing academic performance: What it is and why to do it. In M. B. Shinn (Ed.), *Curriculum-based measurement: Assessing special children.* (pp. 18-78). New York: Guilford Press.

Moran, M.R. (1987). Options for written language assessment. *Focus on Exceptional Children, 19*(5), 1-10.

Nodine, B. F., Barenbaum, E., & Newcomer, P. (1985). Story composition by learning disabled, reading disabled, and normal children. *Learning Disability Quarterly, 8*, 167-179.

Parker, R., Tindal, G., & Hasbrouck, J. (in press). Using direct, countable measures of writing performance for progress monitoring with middle school students in special education. *Exceptional Children.*

Phelps-Gunn, T., & Phelps-Terasaki, D. (1982). *Written language instruction: Theory and remediation.* Rockville, MD: Aspen Systems.

Sabers, D. L., Feldt, L. S., & Reschly, D. J. (1988). Appropriate and inappropriate use of estimated true scores for normative comparisons. *The Journal of Special Education, 22*(3), 358-366.

Salmon-Cox, L. (1981). Teachers and standardized achievement tests: What's really happening? *Phi Delta Kappan, 62*, 631-634.

Salvia, J. A., & Ysseldyke, J. E. (1988). *Assessment in special and remedial education* (4th ed.). Boston: Houghton Mifflin.

Schenck, S. J. (1981). The diagnostic/instructional link in individualized education programs. *Journal of Special Education, 14*(3), 337-345.

Shinn, M. R. (1981). *A comparison of psychometric and functional differences between students labeled learning disabled and low achieving.* Unpublished doctoral dissertation, University of Minnesota, Minneapolis.

Shinn, M. R. (1988). Development of Curriculum-based local norms for use in special education decision-making. *School Psychology Review, 17*(1), 61-80.

Shinn, M. R. (Ed.) (1989). *Curriculum-based measurement: Assessing special children.* New York: Guilford Press.

Shinn, M. R., Tindal, G. A., & Stein, S. (1988). Curriculum-based measurement and the identification of mildly handicapped students: A research review. *Professional School Psychology, 3*(1), 69-85.

Shinn, M. R., Ysseldyke, J. E., Deno, S. L., & Tindal, G. A. (1986). A comparison of differences between students labeled learning disabled and low achieving on measures of classroom performance. *Journal of Learning Disabilities, 19*(9), 545-552.

Sproull, L., & Zubrow, D. (1981). Standardized testing from the administrative perspective. *Phi Delta Kappan*, 628-631.

Thomas, C. C., Englert, C. S., & Gregg, S. (1987), An analysis of errors and strategies in the expository writing of learning disabled students. *Remedial and Special Education, 8*, 21-30.

Thurlow, M. L. & Ysseldyke, J. E. (1982). Instructional planning: Information collected by school psychologists vs. information considered useful by teachers. *Journal of School Psychology. 20*, 3-10.

Tindal, G. (1989). Evaluating the effectiveness of educational programs at the system level using CBM. In M. Shinn (Ed.), *Applications of curriculum-based measurement to the development of programs for mildly handicapped students.* New York: Guilford Press.

Tindal, G. & Parker, R. (1989). Assessment of written expression for students in compensatory and special education programs. *Journal of Special Education, 23*(2),169-184.

Tindal, G., Germann, G. & Deno, S. L. (1983). *Descriptive research on the Pine County Norms: A compilation of findings.* Minneapolis: University of Minnesota Institute for Research on Learning Disabilities.

Videen, J., Deno, S., and Marston, D. (1982). *Correct word sequences: A valid indicator of proficiency in written expression* (Research Report No. 84). Minneapolis: University of Minnesota Institute for Research on Learning Disabilities.