Florida Journal of Educational Research Fall 1987, Vol. 29, No. 1, Pp. 39-54.

The Effect of Test Speededness on the Validity of Reading Comprehension Scores

S. J. Jolly, R. Johnson, B. J. Jones, J. Abalos, and G. W. Gramenz
Palm Beach County, Florida, Schools

ABSTRACT. This study examined irregularities in reading comprehension (RC) test results at both the individual and aggregate level focusing on the effect of test speededness on test scores. Individual student and classroom score profiles were grouped on the basis of RC subtest completion rates. Groups that completed the RC subtest scored lower on other battery subtests (e.g., Language) than did those that did not complete the RC measure. Given these results, the test publisher agreed to generate norms based on the first 40 items of the 60-item RC subtest, using the Rasch model. Rescoring the RC subtest with the 40-item norms substantially corrected score irregularities. The implications of these findings for test score validity are discussed.

Davis (1972) identified two dimensions of reading ability: speed, the rate of comprehension of relatively easy material; and power, the ability to comprehend and apply rather difficult textual material in generous time limits. Depending on the limits established by the test publisher, standardized tests of reading comprehension may or may not involve both speed and power. In the case of a highly-speeded test, the reading comprehension score will reflect reading rate as well as power of comprehension.

After reviewing test speededness studies dating back to 1941, Rindler (1979) concluded that the relationship between speed and power was neither strong nor consistent. Using carefully-developed instrumentation, Bloomers and Lindquist (1944) found the correlation between speed and power of comprehension to be only .30. Moreover, Davidson and Carroll (1945) presented evidence that speed scores were linearly independent of power scores. They also found that scores obtained from timed tests were factorially complex, having loadings on both the speed and power dimensions of ability.

The standardized norm-referenced test batteries typically administered in school systems, however, do not have separate measures of speed and power. They usually provide only one index of reading comprehension, which may or may not involve speed, depending on the restrictiveness of the time limits. The test used in this study--Stanford Achievement Test, Seventh Edition (SAT/7)--measures reading comprehension in terms of the type of material read and the type of question asked (Gardner, et. al., 1984). Measurement of rate is not mentioned in any of the published materials (Karlsen, 1982a and 1982b; Gardner, et. al., 1981a and 1981b).

Myers (1960) surmised that publishers use time limits to ensure the financial success of standardized tests by maximizing the usefulness of a class period. Similarly, Morrison (1960) concluded that time limits are used for practical, rather than empirical, reasons. Helmstadtler and Ortmeyer (1953) recommended that evaluation of a test should include precise knowledge of the relative contributions of speed and power to test scores, and Stafford (1971) called for publishers to report speededness quotients. Morrison (1960, p. 232) argued that "time limits cannot be treated casually because they may produce a change in the determination of test scores so that something different from that intended by the test constructor is being measured." And the Committee to Develop Standards for Educational and Psychological Testing

(1985, p. 28) recommended that

> For tests that impose strict time limits, test development research should examine the degree to which scores include a speed component and evaluate the appropriateness of that component, given the constructs or content the test is designed to measure. [Standard 3.13]

It generally is agreed that restrictive time limits may cause random marking of answers, with guessing increasing as time runs out (Kendall, 1964). Such random marking increases the error of measurement present in test scores (Lord, 1964). In a study of the effect of time limits on the behavior of test takers, Mollenkopf (1960) found that subjects who realized they would not finish a test often marked answers randomly, with chance making a significant difference in scores.

Swineford (1956) considered a test to be unspeeded if virtually all subjects attempted 75 percent of the items and at least 80 percent responded to the last item. The SAT/7 Intermediate I-Form E Reading Comprehension subtest (a 60-item measure with a 30-minute time limit) does not meet these criteria. Only 71 percent of the fourth grade national sample reached item 45, and only 54 percent responded to the last item (J. M. Lenke, personal communication, September 1983). Moreover, in a study of the effect of exceeding prescribed standardized achievement test time limits, Rudman and Raudenbush (1986, p. 15) questioned "... whether the [SAT/7] Reading Comprehension subtest [allowed] sufficient time for completion."

Purpose of Study

The SAT/7 was administered to approximately 50,000 Palm Beach County, Florida, students in grades two through nine during April 1983. Local performance approximated the national average for all subtests except Reading Comprehension, which evidenced an

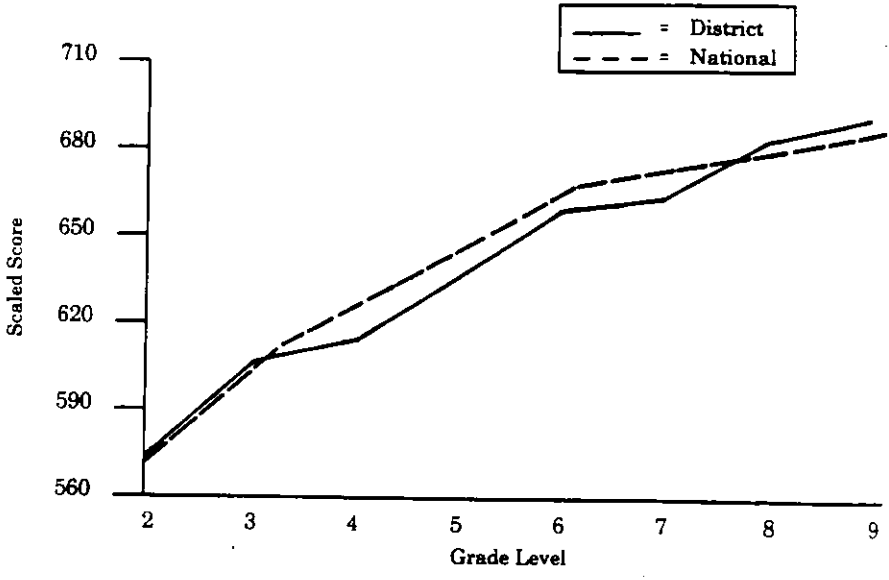

Figure 1. District vs. National Reading Comprehension Mean Scaled Scores. by Grade Level

irregular growth curve (Figure 1). There was an obvious dip from grade three (Primary 3-Form E) to grade four (Intermediate I-Form E), with depressed performance continuing through grade seven, then approaching the national average for grades eight and nine. This atypical growth pattern occasioned an investigation of Reading Comprehension scores at grade four, where the dip was first apparent.

A district item analysis revealed that 61 percent of the students did not answer the last item; that is, only 39 percent of the students completed the subtest (Table 1). In contrast, no more than 20 percent of the students omitted the final item on any other SAT/7 subtest.

Classroom analyses revealed an unexpected pattern. Some classrooms with low percentile rank scores on other SAT/7 subtests had unusually high Reading Comprehension completion rates (i.e., percentage of students marking item 60) and scores (Figure 2), whereas the opposite was true of several classrooms with high scores on the other subtests (Figure 3). These analyses revealed an apparent relationship between deviant performance on Reading Comprehension, relative to other subtests, and unusually low or high completion rates.

The apparent relationship between deviant performance and completion rate suggested that an extraneous factor was affecting Reading Comprehension scores. This study was initiated to determine whether the speeded nature of the SAT/7 Reading Comprehension subtest (Intermediate I-Form E) was related to the inconsistencies noted in Palm Beach County's classroom and district SAT/7 score profiles.

Procedures and Results

Group Comparisons Based on Completion Rates

To examine the effect of speededness, the performance on other SAT/7 subtests of students who completed

Table 1
Grade 4 District Item Analysis: SAT/7 Reading Comprehension, Intermediate I-Form E

Item No.	P-Value	Omit Percent
1	82	0
2	78	0
3	78	1
4	65	0
5	68	0
6	58	0
7	82	0
8	76	0
9	78	0
10	76	0
11	75	0
12	48	1
13	38	1
14	89	1
15	91	1
16	87	1
17	69	1
18	69	1
19	91	1
20	58	2
21	88	3
22	71	3
23	61	4
24	65	4
25	34	5
26	65	9
27	55	10
28	64	12
29	55	14
30	44	15
31	50	16
32	33	17
33	62	22
34	64	23
35	62	24
36	61	25
37	63	27
38	57	29
39	56	30
40	26	30
41	44	39
42	42	40
43	39	43
44	40	43
45	32	45
46	29	46
47	34	47
48	26	48
49	24	48
50	38	53
51	34	54
52	35	56
53	29	56
54	29	57
55	25	58
56	27	60
57	29	60
58	48	61
59	23	61
60	19	61

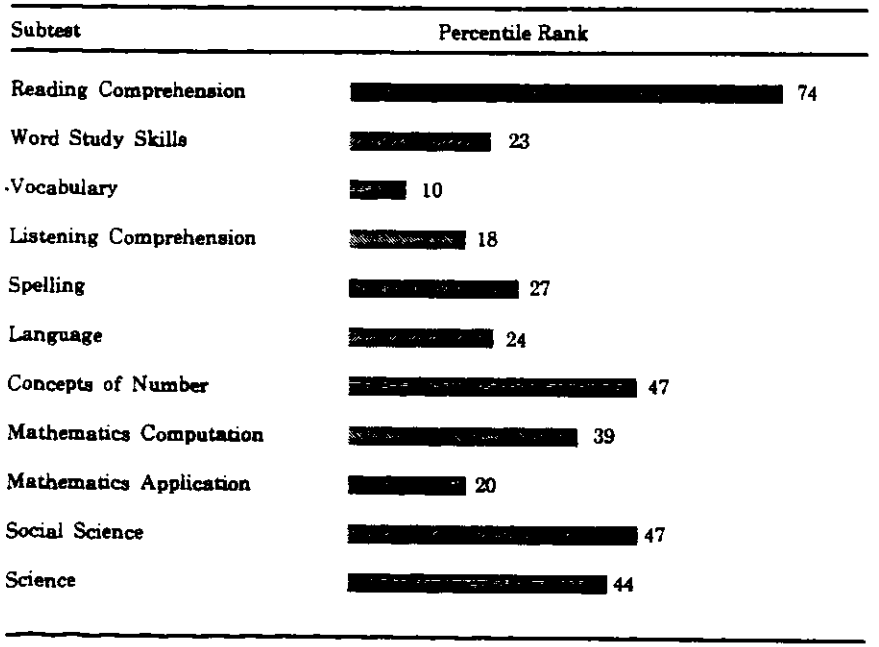

Figure 2. Low-Scoring Class with High Reading Comprehension Score and High Completion Rate (100%)

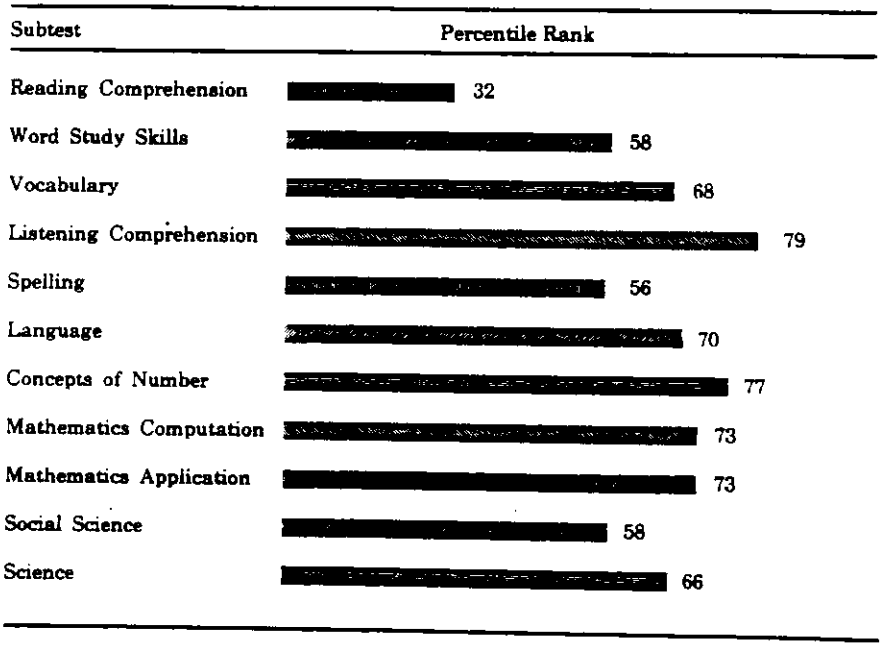

Figure 3. High-Scoring Class with Low Reading Comprehension Score and Low Completion Rate (16%)

Reading Comprehension was compared to that of students who did not finish the subtest. Students were sorted into four groups: those who completed the 60-item subtest (Group 1) and those whose last item marked was 45-59 (Group 2), 34-44 (Group 3), and 1-33 (Group 4). The results of this analysis revealed that Group 1 students had lower mean scaled scores on the other SAT/7 subtests than did students in Group 2 (Table 2). For example, the Language subtest mean of Group 2 was 4.3 scaled score points higher than the corresponding Group 1 mean. Overall, Group 1 students scored 1.9 to 6.7 scaled score points lower on the other subtests than did Group 2 students.

This analysis was repeated at the classroom level by dividing classes into four approximately equal-sized groups: those in which the percentage of students marking the last item was 80-100 (Group 1), 65-79 (Group 2), 51-64 (Group 3), and 0-50 (Group 4). The results revealed a pattern similar to that found in the student analysis (Table 3). Classes with the highest percentages of students completing the Reading Comprehension subtest scored lower on the other subtests than did those with lower completion rates. For example, Group 1 classes scored 3.3 scaled score points lower on Language than did Group 2 classes. Overall, Group 1 classes scored 3.2 to 8.8 scaled score points lower on the other subtests than did Group 2 classes.

Rescoring the Reading Comprehension Subtest

The intermediate I-Form E Reading Comprehension subtest was then rescored to minimize the effect of speededness. Because the SAT/7 had been calibrated with the Rasch model, it was possible to rescore the subtest based on a reduced number of items (J. J. Fremer, personal communication, July 20, 1984). The Psychological Corporation, publisher of the SAT/7, provided the district with norms based on the first 40 items. With an omit rate of 30 percent at item 40 (vs. 61 percent at item 60), the 40-item measure clearly was less speeded than the 60-item subtest for Palm

Table 2
Mean Reading Comprehension Scaled Scores, by Student Completion Group

Subtest	Group 1 (N=2015)	Group 2 (N=1083)	Diff. (1-2)	Group 3 (N=1064)	Diff. (2-3)	Group 4 (N=1033)	Diff. (3-4)
Reading Comprehension	640.4	618.2	22.2	597.1	21.1	567.5	29.6
Word Study Skills	617.9	622.9	-5.0	619.8	3.1	608.3	11.5
Vocabulary	630.0	632.8	-2.8	626.0	6.8	612.1	13.9
Listening Comprehension	637.0	639.0	-2.0	632.9	6.1	621.7	11.2
Total Listening	632.7	635.2	-2.5	628.8	6.4	616.8	12.0
Spelling	634.3	636.5	-2.2	629.8	6.7	607.3	22.5
Language	632.2	636.5	-4.3	632.7	3.8	618.4	14.3
Total Language	631.9	635.6	-3.7	630.4	5.2	613.8	16.6
Concepts of Number	630.3	635.2	-4.9	629.0	6.2	607.9	21.1
Mathematics Computation	626.6	628.5	-1.9	624.1	4.4	610.5	13.6
Mathematics Applications	624.7	631.4	-6.7	627.4	4.0	606.7	20.7
Total Mathematics	626.0	630.0	-4.0	625.2	4.8	608.3	16.9
Social Science	619.7	621.7	-2.0	614.0	7.7	593.0	21.0
Science	623.9	626.4	-2.5	621.0	5.4	603.0	18.0

Table 3
Mean Reading Comprehension Scaled Scores, by Classroom Completion Group

Subtest	Group 1 (N=57)	Group 2 (N=46)	Diff. (1-2)	Group 3 (N=49)	Diff. (2-3)	Group 4 (N=50)	Diff. (3-4)
Reading Comprehension	625.1	617.9	7.2	610.1	7.8	595.3	14.8
Word Study Skills	614.8	621.9	-7.1	619.0	2.9	614.8	4.2
Vocabulary	624.4	631.8	-7.4	627.6	4.2	622.1	5.5
Listening Comprehension	633.2	637.0	-3.8	633.0	4.0	630.8	2.2
Total Listening	628.2	633.6	-5.4	629.8	3.8	625.9	3.9
Spelling	625.7	631.9	-6.2	631.4	.5	624.5	6.9
Language	628.5	631.8	-3.3	632.7	-.9	628.5	4.2
Total Language	626.6	630.8	-4.2	631.3	-.5	626.0	5.3
Concepts of Number	624.9	632.8	-7.9	628.5	4.3	621.1	7.4
Mathematics Computation	622.5	625.9	-3.4	621.1	4.8	622.3	-1.2
Mathematics Applications	619.0	627.8	-8.8	626.9	.9	618.9	8.0
Total Mathematics	621.1	627.0	-5.9	624.2	2.8	620.0	4.2
Social Science	612.9	617.8	-4.9	615.2	2.6	608.5	6.7
Science	619.2	622.4	-3.2	621.8	.6	614.9	6.9

Beach County students.

The similarity of the district score distributions of the 40- and 60-item versions of the subtest alleviated concerns about ceiling effects (Figure 4). Although the internal consistency estimate (Cronbach's alpha) for the 40-item test (.91) was slightly lower than that for the 60-item test (.95), the latter coefficient could be considered spuriously high because of the highly-speeded nature of the 60 item subtest (Stanley, 1971).

Application of the 40-item Norms at the Student and Classroom Levels

To investigate the relationship of the two Reading Comprehension measures to the other SAT/7 subtests, correlations of the 40- and 60-item scores with other subtest scores were performed at the student and classroom group levels. Student 40-item scores correlated more highly with the other subtests than did the corresponding 60-item scores. Correlations with the 60-item test centered in the mid .60's, whereas those with the 40-item test clustered around .70 (Table 4).

The classroom-level correlations evidenced a similar pattern. Forty-item scores correlated more highly with the other SAT/7 subtests than did the 60-item scores. Correlations with the 60-item test centered in the low .70's, whereas those with the 40-item test clustered around .85 (Table 4).

Application of the 40-item norms resulted in both student and classroom score changes. At the student level, use of the 40-item norms produced a mean gain of 5.7 scaled score points. At the classroom level, they produced percentile rank scores consistent with those on the other subtests, particularly for the deviant classrooms found in the initial analysis. For example, the class referenced in Figure 3 had a low 60-item Reading Comprehension score (32nd percentile). Use of the 40-item norms resulted in a percentile rank

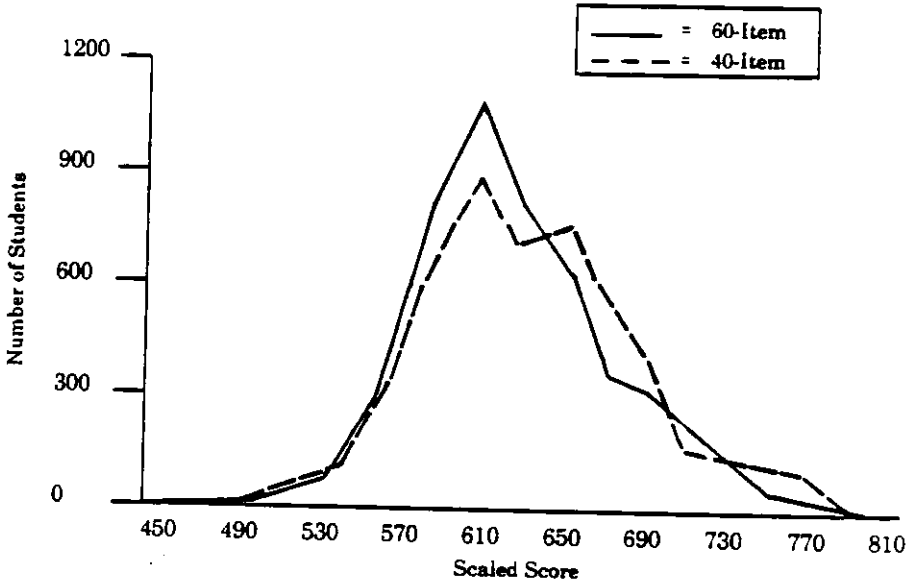

Figure 4. 40-Item vs. 60-Item Reading Comprehension Scaled Score Distributions

Table 4
Student and Classroom Group Correlations of 40-Item and 60-Item Reading Comprehension Scores With Other Subtests

Subtest	Student RC Scores (N=5100)		Classroom RC Scores (N=200)	
	60-Item	40-Item	60-Item	40-Item
Reading Comprehension (60)	—	.89	—	.92
Reading Comprehension (40)	.89	—	.92	—
Word Study Skills	.58	.66	.66	.80
Vocabulary	.62	.68	.68	.81
Listening Comprehension	.61	.70	.71	.82
Total Listening	.65	.72	.72	.84
Spelling	.64	.68	.72	.85
Language	.64	.71	.71	.82
Total Language	.69	.75	.74	.86
Concepts of Number	.59	.65	.71	.83
Mathematics Computation	.49	.53	.60	.69
Mathematics Applications	.61	.69	.72	.86
Total Mathematics	.63	.69	.71	.84
Social Science	.73	.77	.80	.90
Science	.70	.75	.79	.88

of 51, which was more in line with the rest of the class scores. Similar results were obtained for deviant classes with high 60-item Reading Comprehension scores. For example, the percentile rank of the class referenced in Figure 2 dropped from 74 to 53.

Discussion

As a group, fourth grade students who completed the Reading Comprehension subtest scored higher on this measure and lower on the other SAT/7 subtests than those (Group 2) who completed over 75 percent of the Reading Comprehension items but did not complete the subtest. Classes with the highest percentages of students completing the Reading Comprehension subtest scored higher on this measure and lower on the other SAT/7 subtests than did those with lower completion rates. Correlations of the 40-item Reading Comprehension scores with other SAT/7 subtests were higher than the correlations of the 60-item scores with these subtests for both students and classrooms. These results bring into question the validity of scores produced by the 60-item version of the Intermediate I-Form E Reading Comprehension subtest.

The Committee to Develop Standards for Educational and Psychological Testing (1985, p. 10) maintains that "validating inferences about a construct also requires paying careful attention to aspects of measurement such as test format, administration conditions, or language level, that may affect test meaning and interpretation materially." This study suggests that test speededness also is an aspect of measurement that can affect test meaning and interpretation materially.

References

Bloomers, P., & Lindquist, E. F. (1944). Rate of comprehension of reading: Its measurement and its relation to comprehension. Journal of Educational Psychology, 35(8), 449-473.

Committee to Develop Standards for Educational and Psychological Testing of the American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. (1985). Standards for Educational and Psychological Testing. Washington, D.C.: American Psychological Association.

Davidson, W. M. N., & Carrol, J. B. (1945). Speed in level components in time limit scores: A factor analysis. Educational and Psychological Measurement, 5, 411-427.

Davis, F. B. (1972). Psychometric research on comprehension in reading. Reading Research Quarterly, 7(4), 628-678.

Gardner, E. F., Rudman, H. C., Karlsen, B., & Merwin, J. C. (1981a). Stanford Achievement Test Directions for Administering (Standardization ed.). USA: Psychological Corporation.

Gardner, E. F., Rudman, H. C., Karlsen, B., & Merwin, J. C. (1981b). Stanford Achievement Test, Intermediate I, Form E (Standardization ed.). USA: Harcourt.

Gardner, E. F., Madden, R. Rudman, H. C., Karlsen, B., Merwin, J. C., Collins, R., & Collins, C. S. (1984). Stanford achievement test series: A handbook of instructional strategies. USA: Harcourt.

Helmstadtler, G. C., & Ortmeyer, D. H. (1953). Some techniques for determining the relative magnitude for speed and power components of a test. Educational and Psychological Measurement, 13(2), 280-287.

Karlsen, B. (1982a). Diagnosing comprehension: How well can students read what? USA: Psychological Corporation.

Karlsen, B. (1982b). A breakthrough in reading assessment: Stanford. USA: Psychological Corporation.

Kendall, L. M. (1964). The effects of varying time limits on test validity. Educational and Psychological Measurement, 24(4), 789-800.

Lord, F. M. (1964). The effect of random guessing on test validity. Educational and Psychological Measurement, 24(4), 745-747.

Mollenkopf, W. G. (1960). Time limits and the behavior of test takers. Educational and Psychological Measurement, 20(2), 223-230.

Morrison, E. J. (1960). On test variance and the dimensions of the measurement situation. Educational and Psychological Measurement, 20(2), 231-250.

Myers, C. T. (1960). Introduction. Educational and Psychological Measurement 20(2), 221-222.

Rindler, S. E. (1979). Pitfalls in assessing test speededness. Journal of Educational Measurement, 16(4), 261-270.

Rudman, H. C., & Raudenbush, S. W. (1986). The effect of exceeding prescribed time limits in the administration of standardized achievement tests. Paper presented at the joint meeting of the National Council on Measurement in Education and the American Educational Research Association, San Francisco.

Stafford, R. E. (1971). The speededness quotient: A new descriptive statistic for tests. Journal of Educational Measurement, 8(4), 275-277.

Stanley, J. C. (1971). Reliability. In R. L. Thorndike (Ed.), Educational Measurement (2nd ed.) (pp. 356-442). Washington, D.C.: American Council on Education.

Swineford, F. (1956). Technical manual for users of test analyses (Statistical Report 56-42). Princeton: Educational Testing Service.

G. W. Gramenz, Testing and Evaluation, Palm Beach County Schools, 3323 Belvedere Rd., West Palm Beach, Florida 33402