DOCUMENT RESUME

ED 390 879 TM 023 944

AUTHOR Rudman, Herbert C.; Raudenbush, Stephen W.
TITLE The Effect of Exceeding Prescribed Time Limits in the Administration of Standardized Achievement Tests: An Abstract. Research Series. CEPSE/No. 5.
INSTITUTION Michigan State Univ., East Lansing. Dept. of Counseling, Educational Psychology, and Special Education.
PUB DATE May 86
NOTE 30p.; Paper presented at the Annual Meeting of the American Educational Research Association and the National Council on Measurement in Education (April 1986). For related studies, see TM 023 945-946.
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Achievement Tests; Analysis of Covariance; Decision Making; Demography; *Elementary School Students; Grade 5; Intermediate Grades; Reading Comprehension; Scores; *Standardized Tests; Test Results; *Time; *Timed Tests
IDENTIFIERS *Stanford Achievement Tests

ABSTRACT

The effect on achievement when the prescribed maximum time limits of a standardized achievement test are exceeded was studied using students nested in block-by-treatment combinations. The analysis was a hierarchical (nested) randomized blocks analysis of covariance involving 408 fifth graders taking the Stanford Achievement Test in Lansing, Michigan. Results indicated that excess testing time had a significant positive linear effect on reading comprehension and total reading scores and a significant nonlinear effect on the same outcomes. The benefit of excess time was most pronounced between maximum time and 5 minutes more and between 10 and 15 minutes excess time. No significant interactions were found between excess time and selected demographic variables. It was clear that decisions made at the classroom level were substantially more sensitive to the effects of excess time than were student level decisions. Results highlight the importance of excess time at various levels of aggregation. (Contains 9 tables and 24 references.) (SLD)

THE EFFECT OF EXCEEDING PRESCRIBED TIME LIMITS IN THE ADMINISTRATION OF STANDARDIZED ACHIEVEMENT TESTS: AN ABSTRACT

Herbert C. Rudman
Stephen W. Raudenbush

OERSE/No. 5 May 1986

RESEARCH SERIES

Department of
Counseling
Educational Psychology
and
Special Education

College of Education — Michigan State University

THE EFFECT OF EXCEEDING PRESCRIBED TIME LIMITS IN THE ADMINISTRATION OF STANDARDIZED ACHIEVEMENT TESTS: AN ABSTRACT

Herbert C. Rudman
Stephen W. Raudenbush

CEPSE/No. 5 May 1986

The Effect of Exceeding Prescribed Time Limits in the Administration of Standardized Achievement Tests: An Abstract

Herbert C. Rudman and Stephen W. Raudenbush
Michigan State University

Standardized achievement tests are being used to make "high-stake" decisions, and the ways in which they are being used seem far afield from the purposes for which these tests have been designed. These types of decisions have brought new importance to the procedures used in test administration. One such procedure is the adherence to maximum time-limits established for each of the subtests used. The purpose of this study was to examine the effect upon achievement when the prescribed maximum time limits of a standardized achievement test are exceeded. The basic design involved pupils nested within block-by-treatment combinations. Covariates were utilized to increase statistical power of the analysis. The analysis employed was a hierarchical (nested) randomized blocks analysis of covariance. Results indicate that (1) excess testing time has a significant, positive, linear effect on Reading Comprehension and Total Reading scores, and (2) a significant non-linear effect on the same outcomes. The benefit of excess time is most pronounced "early" --between the maximum time and five minutes more, and "late" -- between 10 and 15 minutes excess time. No significant interactions were found between excess time and selected demographic variables. Four practical considerations were examined: the impact of excess time on published norms; the impact on decisions about student placement; the impact on decisions about teacher effectiveness; and the relationship of the specific characteristics of the standardized test used in this study and treatment effects obtained. Decisions made at the classroom level are more sensitive to the effects of excess time than are student level decisions.

The Effect of Exceeding Prescribed Time Limits in the Administration of Standardized Achievement Tests [1]

Herbert C. Rudman and Stephen W. Raudenbush
Michigan State University

The purpose of this study is to examine the effect upon achievement when the prescribed maximum time limits of a standardized achievement test are exceeded. While there may be a differential impact of exceeding testing-time limits depending upon the content tested, this study is limited to an analysis of achievement in Reading. Specifically, this investigation focuses on three Reading scores measured by the Stanford Achievement Test; Reading Comprehension, Word Study Skills, and Total Reading.

Although a study of the effect of increased test-taking time on achievement scores is not new (Boag and Neild, 1962; Daly and Stahmann, 1968; Lord, 1956), the need to reexamine this effect has been made acute by the recent emphasis on high-stakes use of standardized achievement test results. Standardized achievement tests are being used today in ways that were not seriously contemplated by test authors and developers as recently as fifteen years ago. The influence of test scores on administrative decisions seems, at times, to strain the psychometric limits of the tests (Rudman, 1985a; Rudman, 1985b; Hoover, 1984; Lewis, 1985; Traub, 1983). Some school districts have reported using gain scores derived from repeated use of these tests to award school personnel cash bonuses, (Florida Statute, 1985) and others are placing great

---

1. A paper presented at a joint meeting of the National Council on Measurement in Education, and the American Educational Research Association, April 19, 1986. The authors wish to acknowledge the assistance of Mr. Rafa Kasim, Drs. Grace Iverson, Pat Petersen, Mrs. Jane Faulds and the participating teachers and administrators of the Lansing, Michigan School District.

emphasis upon teacher evaluation (Elliott and Hall, 1985; Turlington, 1985; Tirozzi, et al., 1985). These types of decisions have brought new importance to the time-limits established for test administration. Cautions have been raised about maintaining a strict adherence to maximum time limits to avoid any possibility of undue advantage being gained when these time limits are not adhered to. If faculty and other staff personnel are to be awarded cash bonuses based upon annual increases in the test scores of students, then any factor in the administration of these tests which can influence a test score becomes a target for close scrutiny.

This kind of use of standardized achievement tests raises some interesting questions. Is the time allotted for test administration truly an important variable in the score a student makes? Should we anticipate a rise in an achievement score if a teacher exceeds the time limits stipulated in the accompanying manuals of a standardized test? If we can anticipate a difference in the number of items answered correctly, how much excess time is needed to observe an increase in test scores? How are these time allotments determined? If exceeding time limits does increase test scores, how likely are the increases to influence administrative decisions in important ways?

Nunnally (1978), in a discussion of time limits speaks of a "comfortable time limit" which is defined as "the amount of time required for 90 percent of the persons to complete a test under power conditions" (1978, p. 632). It has been common practice during a national item analysis of standardized achievement tests to note when 90 percent of the students in the item analysis sample have completed the subtest of interest. Test authors and developers have assumed that this proportion of students have answered these items on the basis of knowledge or through informed guessing. The remaining 10 percent would resort to random guessing and this would not seriously affect the final test score.

While there is good reason to believe that this assumption is an accurate one, one might suspect that some subtests (particularly Reading Comprehension) may have an element of speededness in them. Such speededness may stem from curriculum anomalies within particular school districts or from the particular subtest itself.

When tests are used for such critical purposes as determining the merit of teachers and programs, then fairness becomes an important consideration in test use. If test data are to be used to award bonuses to teachers for their students' gains in achievement from one year to the next, then the conditions under which the testing took place should be comparable from one testing environment to another. A lack of fairness may be introduced by allowing extra time for students to complete a standardized test with fixed time limits. If one class is allowed to spend thirty minutes on a reading test, and another is allowed thirty-five or forty-five minutes, then comparability is lacking, and one might assume that students can benefit from extra time taken. But does exceeding the maximum time limit really make a difference?

Several studies have investigated the effect of increased test-taking time on test scores (Evans and Reilly, 1972; Evans and Reilly, 1973; Wild and Durso, 1979; Wild, Durso, and Rubin, 1982). A common finding among them is that an increase in the allotted testing time does not result in a significant or meaningful change in test scores.

Increased test-taking time can be viewed as increasing the time for each item, or increasing the time for the total subtest. In the former instance an increase of item time gives students of varying ability uniformly increased time per item. This would permit detailed study of the impact of differential student characteristics (e.g. sex, race, socio-economic status, and ability) on item performance. The latter approach, increasing total subtest time, reflects

how tests are empirically developed and administered. Theoretically, if total testing time is increased, one could assume that time available per item is uniformly affected. However, achievement tests are sequentially constructed with the most difficult items — as determined through prior item analysis research — placed at the end of the test. Students of lesser ability may simply not know the concepts measured in the latter part of a test, and increasing the time of the total test ought not to result in a higher score (Nunnally, 1978; Mueller and Wasser, 1971; Morrison, 1960).

A recent investigation addressed the question of increased time per item, by examining its effect on two demographic variables of interest to this present study, race and sex (Wild, Durso, and Rubin, 1982). A major concern of this study was the degree to which testing time would impact selected subgroups. The primary question pursued was "...whether increased time allotments for the Graduate Record Examination's verbal and quantitative sections would differentially affect the scores of subgroups of interest" (p. 24).

No significant differences were found by sex or race. The authors concluded that:

1. ... extra time allows examinees who have completed the test to review their answers.

2. ... lower scoring examinees find the test more difficult and tend not to complete the test. Given more time to answer questions, lower scoring examinees would be expected to answer fewer of the additional questions, especially since the more difficult test items appear at the end.

PROCEDURES

Test Content

The Stanford Achievement Test, Intermediate 1, Form F was used to collect student achievement data on two subtests, Reading Comprehension and Word Study Skills. The scores obtained from these subtests were summed and expressed as a

third measure, Total Reading. Achievement test results for the previous year were used as an ability covariate and utilized the Primary 3, Form E battery of the same test series.

Both the Reading Comprehension and the Word Study Skills subtests consist of 60 items. The Word Study Skills subtest has two parts; one emphasizes structural analysis and the other phonetic analysis. The maximum time limit for the structural analysis part is 15 minutes, and for phonetic analysis is 20 minutes. The total maximum time for the Word Study Skills test is 35 minutes. The Reading Comprehension subtest is administered without any break in test content, and its maximum time allotment is 30 minutes. The authors of the Stanford Achievement Test caution that a maximum time limit is never to be extended (Gardner, Rudman, Karlsen, and Merwin, 1982).

Sample and Design

Lansing, Michigan is an urban school district which consists of 33 elementary schools. The faculty from twelve of these elementary schools volunteered to serve as participants in this study. Nineteen fifth-grade classrooms supplied usable data for 408 pupils. Table 1 describes the sample on key demographic variables and test scores.

The 19 classrooms were assigned to one of five blocks to ensure adequate matching of the schools in the sample. Classrooms were matched on two characteristics, (1) the previous year's achievement in Reading, and (2) the proportion of families receiving aid for dependent children. The blocking assured that no two classrooms within the same school would experience the same treatment. Four treatment groups representing testing-time allotments were established. Within each of the five blocks, classrooms were assigned at random to one of the four treatments. If the sample had included 20 classrooms, the design would have been a balanced randomized block design (Kirk, 1982, Chapter

6). Since the sample included 19 classrooms, one block included only three classrooms. As a result, Treatment Group 2 included only four classrooms; all other treatment groups included five classrooms.

Table 2 represents the testing-time used within the four treatment groups. Treatment Group 1 used the normal maximum time limits stipulated in the regular manuals which accompany the Stanford Achievement Test. Treatment Groups 2,3, and 4 were incremented in 10 minute intervals for the total testing-time allocated across all parts of the two subtests. Unadjusted posttest means and sample sizes are provided in Table 3.

Analysis

The basic design involved pupils nested within block-by-treatment combinations. To increase the statistical power of the analysis, covariates were utilized. Thus, the analysis employed was a hierarchical (nested) randomized blocks analysis of covariance. Analytic issues involved (a) choosing covariates; (b) handling unequal sample sizes; and (c) choosing an appropriate error term for hypothesis testing. We review key analytic decisions below.

Choosing covariates. For the Reading Comprehension subtest, the best single covariate proved to be the Total Reading pretest, r = .77. For the Word Study Skills subtest, the Word Study Skills pretest proved to be the best covariate, r = .80. For Total Reading, the Total Reading pretest proved best, r = .82. Only one covariate per outcome was needed since other likely covariates (ethnicity, native language, sex, and family configuration) were not significantly related to the outcome after adjusting for the effect of the "best" covariate.

Contending with unequal sample sizes. Because classrooms, treatments, and blocks were mildly unbalanced (see Table 3), a sequential analysis of covariance was employed via multiple regression. For each of the three outcomes, effects

were entered in the following order: the covariate; then the block effects; then the linear, quadratic, and cubic effects of the treatment; and finally, two-way interactions between treatment and the demographic variables mentioned above. None of these two-way interactions were statistically significant. Because intercorrelations among the effects were weak, the results are insensitive to ordering effects.

To facilitate a partition of the total variation (adjusted for the covariate) into between-and within-classroom components, a one factor analysis of covariance was computed for each outcome, with the nineteen classrooms serving as levels of the factor. This second ANCOVA, in combination with the regression approach mentioned above, supplied all the sources of variation needed for the analysis. The between-classroom regressions were similar to the pooled within-classroom regressions, justifying the use of the pooled, within-class regressions to adjust for the effect of the covariate.

Table 4 displays the models estimated, and for each model, the associated degrees of freedom. It also shows how each analysis of variance table was constructed. For each model estimated, a sum of squared residuals was computed. The sum of squares associated with each effect is the reduction in the residual sum of squares ($SS_{reduction}$) associated with adding that effect, and the degrees of freedom for an effect is the reduction in residual degrees of freedom ($df_{reduction}$) associated with adding that effect. Thus, the mean square associated with each effect is given by

$$MS_{reduction} = SS_{reduction} / df_{reduction}$$

Choosing the appropriate F-test. In nested designs of this type, the appropriate F test for the treatment contrasts typically uses the unexplained variation between classes as the mean square error. In such an analysis, the residual effects of classrooms are viewed as random effects. When the null

hypothesis of no residual variance between classrooms is retained, an alternative error term is available, and typically provides a more powerful F-test. The alternative is to pool the residual between-class variation with the residual within-class variation, yielding a dramatic increase in the degrees of freedom associated with error (Hopkins, 1982). Unfortunately, the hypothesis of no residual variation was rejected for Word Study Skills, $F(11, 389) = 2.36$, $p < .01$, and for Total Reading, $F(11, 389) = 1.94$, $p < .05$. For Reading Comprehension, the null hypothesis of no residual between-classroom variance was retained, $F(11, 389) = 1.38$, $.10 < p < .25$. However, for this subtest the results of hypothesis testing were insensitive to choice of the error term. Consequently, the more conservative F-test is utilized for this subtest, and therefore, Table 5 utilizes the residual between-classroom mean square for all treatment contrasts for all dependent variables.

RESULTS

The key finding of the study is a highly significant, positive, linear effect of excess time on Reading Comprehension test scores, $F(1,11) = 23.33$, $p < .01$. There was no significant linear effect for Word Study Skills and, in fact, there were no significant differences among treatment means for that subtest, $F(3,11) = 2.92$.

Largely because of the linear effect of excess time on Reading Comprehension, a significant, positive, linear effect of excess time was also manifest for Total Reading, $F(1,11) = 12.98$, $p < .01$. These results are reported in Table 5.

Estimation of the magnitude of the linear trend depends on two factors: choice of outcome measure and choice of level of aggregation. The partial Pearson product-moment correlation between excess time and the unadjusted Reading Comprehension scores is $r = .17$; between excess time and the adjusted

posttest score, r = .26 (see Table 5). However, at the classroom level, the partial correlation between excess time and the adjusted posttest classroom mean is r = .67, so that 45 percent of the adjusted between-classroom variation is attributable to this linear trend. The partial correlation between excess time and Total Reading is r = .13 for the unadjusted posttest; r = .23 for the adjusted posttest, and r = .59 for adjusted classroom means, so that about 35 percent of the adjusted between-classroom variation in Total Reading is attributable to the linear effect of excess time. Thus, the practical effect of excess time may be more pronounced for decisions about classrooms than for decisions about students. This issue is addressed further in the discussion section.

A second but unanticipated finding was a significant non-linear effect of excess time for both Reading Comprehension and for Total Reading. In each case, the cubic trend was significant: $F(1,11) = 9.75$, $p < .01$ for Reading Comprehension; and $F(1,11) = 10.86$, $p < .01$ for Total Reading. Examination of the adjusted posttest means (see Table 8) suggests that in both cases, the benefit of excess time is most pronounced "early" (between zero and five minutes) and "late" (between 10 and 15 minutes), with no effect in the middle. The significant cubic trend may also be explainable in part by an apparent outlier in the third treatment group, Block three (see Table 3). However, this unusually low scoring class also scored unusually low on the pretest, suggesting that this low value may be legitimate. The effect of removing this possible outlier would be to weaken the cubic trend and to strengthen the linear trend.

There were no significant interactions between treatment and demographic variables. The latter included sex, ethnicity, native language, and family configuration.

DISCUSSION

The Results

This study would indicate that excess testing time has a significant affect on achievement in Reading Comprehension, but not in Word Study Skills. Since the two subtests are summed into a Total Reading Score which, in turn, is sometimes used as the single indicant of achievement in Reading, it was important to also consider the Total Reading score. The effect of time on Reading Comprehension was strong enough to overcome the lack of effect on Word Study Skills, so that the Total Reading Score showed a similar linear trend of higher achievement as time of testing increased. This relationship between time and achievement was particularly emphasized in terms of scaled scores and grade equivalents (Table 6), probabilities of scoring higher than the median (Table 7) and in terms of expected gain scores (Table 8)

The cubic trend, while significant, reflected a puzzling phenomenon which indicated a decrease in the effect of excess testing-time on achievement in Treatment group three (five to ten minutes).

In part, the cubic trend reflected the apparent influence of an "outlier" classroom.. When this classroom was experimentally eliminated from the analysis data, its removal weakened the cubic trend, but amplified the result of excess testing time on achievement.

Practical Implications of the Study

This discussion will center on three questions that have practical significance for testing practice in the schools: (1) Suppose high-stakes use are to be made of standardized test data. Can the user legitimately use norms that accompany the test if departures have been made from established time limits? (2) What impact would increased testing time have on the probabilities that students would receive benefits in the form of placement in advanced classes, promotion to a higher grade, and the like? (3) What impact would

increased testing time have on the probabilities that teachers would receive awards, bonuses, or other recognition?

Impact of excess time on norms. The data as reported in this study are expressed as raw scores (number of items scored correctly). When they are rounded to integers and are transformed into scaled scores and grade equivalent scores the relationships noted earlier add still another dimension to the practical meaning of these trends (see Table 6).

It can be seen from Table 6 that five additional minutes of testing time beyond the recommended time can yield an additional .9 grade equivalent score difference between the recommended time and five minutes more, and as much as a 1.4 grade equivalent score after fifteen minutes. The results for the Total Reading test showed similar increases over time. In contrast, Word Study Skills yields only a .4 grade equivalent score difference after five additional minutes, and a .6 grade equivalent score difference after 15 minutes additional time. These differences noted in the Word Study Skills subtest were not statistically significant. With grade equivalent differences such as these, it is unwise to attempt to use published norms when test-taking time is at such great variance from the established time limit.

Impact of excess time on probability of success. Some administrators, as alluded to earlier, are faced with the need to make decisions of a high-stake nature based upon test data. If standardized achievement tests are used, and if teachers do not adhere to the prescribed time limits, decisions may be reached which may lack a degree of fairness (see Table 7). Some decisions might affect students; e.g., assignment to an accelerated instructional group or promotion from one grade to another. Other decisions could affect teachers, for example, the provision of rewards, bonuses, or recognition of merit for classroom gains as measured by a standardized test. These are examples of "high stakes" use of

tests. What impact could excess test-taking time have on such decisions?

Consider first a hypothetical situation in which decisions are to be made about students. Suppose that any student scoring above the median is to be assigned to the "high" reading program, and students scoring below the median are to be assigned to the "low" reading program. How would excess time influence the probability of "success" (high placement) based on results from this study? Rosenthal and Rubin's (1982) "Binomial effect size display" enables one to translate experimental effect sizes on a continuous variable into probabilities of success on a binary variable. Table 7a shows how excess testing time influences the probability of student success as defined above. Without excess time, a randomly selected student would have a probability of "success" of .50. With excess time that probability would increase to .64 based on 15 minutes extra time if Reading Comprehension were the test employed, and to .69 if the Total Reading test were used.

Now consider a second hypothetical situation in which teachers are to be rewarded for gains their classrooms exhibit in Reading achievement. Table 7b shows how excess time would influence the probability of teachers' success. Now the effect sizes are measured in units of the standard deviation of classroom gains based in our study. Without the benefit of excess time, the probability of success for a randomly selected teacher is .50; with 15 minutes of excess time that probability increases to .80 (for Reading Comprehension) and .86 (for Total Reading).

It is abundantly clear from these admittedly simplified hypothetical examples that decisions made at the classroom level are substantially more sensitive to the effects of excess time than are student level decisions. Consequently, discussions about the effect of irregularities in test administration on test validity must take into account the level of aggregation

at which the test results are to be used in decision making.

Another way of examining the effect of excess time on achievement can be seen in Table 8. The expected standardized gain is expressed as a ratio of the difference in the pre and posttest means divided by the standard deviation of the gain. The Total Reading standardized gain would show an increase of approximately one-third of a standard deviation and increase to .83 of a standard deviation after 15 additional minutes.

These results suggest that administrators concerned about fairness would then need to assure careful monitoring of the testing process. It is not our intention to assume that teachers would not adhere to established procedures. But it is important to recognize the consequences of not following the established testing-time restriction.

The Relationship of Test Characteristics to Treatment Effects

Much of what has been reported in this study has been influenced by the characteristics of the Reading Comprehension subtest of the Stanford Achievement Test Primary 3, Form E, and Intermediate 1, Form F. To the extent that these characteristics carry over to other subtests measured by the Stanford Achievement Test, the conclusions drawn in this section also apply to other subtests.

One indicant of the commonality of the subtests is the intercorrelation to be found between the Reading Comprehension subtest and the other ten subtests which comprise the test at the Primary 3 and Intermediate 1 levels (see Tables 9a and 9b). As one would expect, there is a moderate relationship between Reading Comprehension and the other subtests with the bulk of the relationships approximating .6 to .7 at both levels, and a median intercorrelation of .66 at Primary 3 and .69 at Intermediate 1. An important task of future research is to evaluate the effect of excess time on these subtests.

Another factor to consider when making a judgment about the generalizability of these findings has to do with the possibility that the Reading Comprehension subtest has time limits which simply are too short for the number of items to be tested. Both Reading Comprehension and Word Study Skills subtests consisted of 60 items at Intermediate 1, yet the Word Study Skills test allowed five additional minutes testing time. The data in Table 6 indicate that the Reading Comprehension subtest is more difficult than the Word Study Skills subtest (Table 6). However, five minutes of excess time increased the mean on Reading Comprehension so that it was about as difficult as the Word Study Skills subtest. While this is not a tested conclusion, the observation does lead us to ponder whether the Reading Comprehension subtest did not allow sufficient time for completion.

A third characteristic to consider is the nature of the concepts and skills tested. A description of the content of the Reading Comprehension subtest indicates that the concepts and skills embedded within the Reading Comprehension subtest measure the ability of the pupil to comprehend, among other types of reading materials, that which is to be found in typical textbook passages across a variety of subjects taught. There does not appear to be any particular anomoly in the content tested by the Reading Comprehension test when compared to other subtests. It is constructed in much the same way as all other subtests in these batteries. Items progress from the easiest to the most difficult, and the psychometric characteristics are similar to those other subtests which comprise the test (Gardner, Rudman, Karlsen, and Merwin, 1983). Additional study with other measures of Reading Comprehension are needed. If other standardized tests of Reading Comprehension indicate a similar effect of excess time on achievement then a similar construct of Reading Comprehension is being measured among them. The effect of excess time on achievement may thus be verified as stated in our

study. If the time effect cannot be replicated with other measures then a different construct of Reading Comprehension may have been measured with the instrument used in our investigation. This difference in construct, should it exist, may have been an unanticipated effect on our results.

We have no way, at this point, of knowing whether a similar study utilizing other standardized tests would yield similar results. Clearly, this subtest, in these measures, shows that increased testing time does result in increased test scores in Reading Comprehension and Total Reading. If the Reading Comprehension subtest has an element of speededness in it, then further investigations need to be done to determine whether increased time across all measures and subjects tested in the schools will result in significant gains in achievement. It is especially important to evaluate the likely consequences of any effects excess time has on decisions made at various levels of aggregation.

REFERENCES

Boag, A.K. & Neild, M. (1962). The influence of the time factor on the scores of the Triggs Diagnostic Reading Test as reflected in the performance of secondary school pupils. Journal of Educational Research, 55, 181-183.

Daly, J.L., & Stahmann, R.F. (1968). The effect of time limits on a university placement test. Journal of Educational Research, 62, 103-104.

District Quality Instruction Incentive Program, Sec. 231.532 (3) (F), Florida Statute, (1985).

Elliott, E.J. & Hall, R. (1985). Indicators of performance: measuring the educators. Educational Measurement: Issues and Practice, 4, 6-9.

Evans, F.R. & Reilly, R.R. (1972). A study of speededness as a source of test bias. Journal of Educational Measurement, 9, 123-131.

Evans, F.R. & Reilly, R.R. (1973). A study of test speededness as a potential source of bias in the quantitative score of the Admissions Test for Graduate Study in Business. Research in Higher Education, 1, 173-183.

Gardner, E.F., Rudman, H.C., Karlsen, B., & Merwin, J.C. (1982). Stanford achievement test: directions for administering Primary 3, Forms E/F and Intermediate 1, Forms E/F. Cleveland, OH: The Psychological Corporation.

Gardner, E.F., Rudman, H.C., Karlsen, B., & Merwin, J.C. (1982). Stanford achievement test. Primary 3, Form E and Intermediate 1, Form F. Cleveland, OH: The Psychological Corporation.

Hoover, H.D. (1984). The most appropriate scores for measuring educational development in the elementary schools: ge's. Educational Measurement: Issues and Practices, 3, 8-14.

Hopkins, K.D. (1982). The unit of analysis: group means vs individual observations. American Educational Research Journal, 19, 5-18.

Kirk, R.E. (1982). Experimental Design: Procedures for the Behavioral Sciences. Belmont, CA: Brook-Cole Publishers.

Lewis, A.C. (1985). Test scores test scores on the wall ... . Phi Delta Kappan, 66, 387-388.

Lord, F.M. (1956). A study of speed factors in tests and academic grades. Psychometrika, 21, 31-50.

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

Mueller, D.J. & Wasser, V. (1977). Implications of changing answers on objective test items. Journal of Educational Measurement, 14, 9-14.

Nunnally, J. (1978). Psychometric Theory. New York: McGraw-Hill Book Company.

Rosenthal, R. & Rubin, D.B. (1982) A simple, general purpose display of magnitude of experimental effect. Journal of Educational Psychology, 74, 166-169.

Rudman, H.C. (1985). Responsible and ethical test use: it cuts both ways. Paper presented at a meeting of the National Council on Measurement and the American Educational Research Association. (a).

Rudman, H.C. (1985). Testing beyond minimums. Occasional Paper No. 5. Springfield, IL: Association of State Assessment Programs. (b).

Tirozzi, G.N. et al. (1985). How testing is changing education in connecticut. Educational Measurement in Education: Issues and Practice, 4, 12-16.

Traub, R.E. (1983). A priori considerations in choosing an item response model. in Ronald K. Hambleton (ed.) Applications of Item Response Theory. Vancouver, B.C.: Educational Research Institute of British Columbia.

Turlington, R.E. (1985). How testing is changing education in florida. Educational Measurement in Education: Issues and Practice, 4, 9-11.

Wild, C.L & Durso, R. (1979). Effect of increased test-taking time on test scores by ethnic group, age, and sex. (GRE Board Research Report No. 76-6R). Princeton, NJ: Educational Testing Service.

Wild, C.L., Durso, R. & Rubin, D.B. (1982). Effect of increased test-taking time on

test scores by ethnic group, years out of school, and sex. Journal of Educational Measurement, 19, 19-28.

Table 1
Description of the Sample

Demographic Variables

Variable	Level	Frequency	Relative Frequency
1. Sex	Female	222	.543
	Male	187	.457
2. Ethnicity	Black (non-Hispanic)	75	.183
	Hispanic	68	.166
	Asian	8	.020
	Native American	13	.032
	White (non-Hispanic)	245	.599
3. Family Configuration	Two-parent	266	.650
	Single Parent	143	.350
4. Native Language	English	256	.626
	Other	67	.164
	Missing	86	.210

Test Data

Subtest	Pretest		Posttest	
	m	sd	m	sd
Reading Comprehension	36.70	13.80	41.80	11.71
Word Study Skills	40.65	10.94	43.13	10.53
Total Reading	77.35	22.28	84.93	20.32

Table 2
Testing Time by Treatment and Subtest

Treatment	Reading Comprehension	Word Study Skills			Total Reading
		I	II	Total	
1	30	15	20	35	65
2	35	17	23	40	75
3	40	19	26	45	85
4	45	21	29	50	95

Table 3
Outcome Data

Subtest/Blocks	None	Excess Time: Five Minutes	Excess Time: Ten Minutes	Excess Time: Fifteen Minutes	Overall
Unadjusted classroom means and sample sizes					
Reading Comprehension					
Block 1	32.40	39.43	28.41	44.38	36.16
Block 2	37.61	40.60	37.04	44.38	39.11
Block 3	40.57	40.88	42.72	40.86	41.30
Block 4	41.75	———	42.23	48.95	44.26
Block 5	41.24	50.36	51.04	45.77	47.38
Overall	38.78	43.22	40.95	44.38	41.80
Word Study Skills					
Block 1	37.20	47.09	28.47	47.12	40.33
Block 2	42.06	39.00	41.59	39.26	40.48
Block 3	43.14	42.23	44.20	44.86	43.56
Block 4	47.10	———	42.00	47.95	45.42
Block 5	42.24	48.21	47.14	43.86	45.58
Overall	42.36	44.44	41.37	44.62	43.13
Sample Sizes					
Block 1	20	23	17	16	76
Block 2	18	20	22	19	79
Block 3	21	26	25	21	93
Block 4	20	—	26	22	68
Block 5	21	28	22	22	93
Overall	100	97	112	100	409
Adjusted Treatment Means (Unstandardized)					
Reading Comprehension	38.89	42.80	41.09	44.91	
Word Study Skills	42.74	43.95	41.20	44.65	
Total Reading	80.31	86.89	82.52	90.01	
Adjusted Treatment Means (Standardized)					
Reading Comprehension	-.29	.08	-.06	.27	
Word Study Skills	-.04	.08	-.18	.14	
Total Reading	-.23	.10	-.12	.25	

Table 4
Analytic Method

Model/Source	df residual	Sum of squares residual
Constant	408	(1)
Constant, covariate	407	(2)
Constant, covariate, blocks	403	(3)
Constant, covariate, blocks, linear trend	402	(4)
Constant, covariate, blocks, Linear, Quadratic trends	401	(5)
Constant, covariate, blocks, linear, quadratic, cubic trends	400	(6)
Constant, Covariate, classrooms	389	(7)

Partitioning of variation

	df reduction	Sum of Squares reduction
Covariate	1	(1) - (2)
Between Classes (adjusted)	18	(2) - (7)
Blocks	4	(2) - (3)
Treatments	3	(3) - (6)
Linear	1	(3) - (4)
Quadratic	1	(4) - (5)
Cubic	1	(5) - (6)
Residual (blocks by treatments)	11	(6) - (7)
Within Classes (adjusted)	389	(7)
Total	408	(1)

Table 5
ANCOVA source tables for a) Reading comprehension, b) Word Study Skills, and c) Total Reading.

Source	df	Sum of Squares	Mean Square	F	ETA_a	ETA_b
Reading Comprehension						
Covariate	1	33010.977	33010.977		.77	—
Between Classes	18	3559.434	197.746		.25	.39
Blocks	4	529.954	132.488		.10	.15
Treatments	3	2273.811	757.937	11.03***	.20	.31
Linear	1	1602.809	1602.809	23.33***	.17	.26
Quadratic	1	1.025	1.025	.01	.00	.01
Cubic	1	669.977	669.977	9.75***	.11	.17
Residual	11	755.669	68.697	1.38*	.12	.18
Within Classes	389	19367.745	49.79			
Total	408	55938.156	137.103			
Word Study Skills						
Covariate	1	29658.803	29618.803		.81	—
Between Classes	18	1828.834	101.602		.20	.34
Blocks	4	205.673	51.418		.07	.11
Treatments	3	719.729	239.910	2.92	.13	.21
Linear	1	41.564	41.564	.51	.03	.05
Quadratic	1	166.765	166.765	.03	.26	.10
Cubic	1	511.400	511.400		.11	.18
Residual	11	903.432	82.130	2.36**	.14	.24
Within Classes	389	13755.967	34.825			
Total	408	45243.604	109.187			
Total Reading						
Covariate	1	112433.764	112433.764		.82	—
Between Classes	18	8696.347	483.130		.23	.39
Blocks	4	394.744	98.686		.05	.08
Treatments	3	5701.979	1900.660	8.04***	.18	.32
Linear	1	3067.600	3067.600	12.98***	.13	.23
Quadratic	1	68.567	68.567	.29	.02	.03
Cubic	1	2565.812	2565.812	10 86***	.12	.21
Residual	11	2599.624	236.329	1.94***	.12	.21
Within Classes	389	47342.032	121.702			
Total	408	168472.083	406.937			

* .10<P<.25
** p<.05
*** p<.01
a based on raw posttest score
b based on covariance adjusted posttest score

Table 6
A Comparison of Mean Raw Scores and Derived Scaled Scores and Grade Equivalents By Treatment Level

Subject	Treatments	Mean Score	Scaled Scores [b]	Grade Equival
Reading Comprehension	1	39	632	5.0
	2	43	645	5.9
	3	41	638	5.4
	4	45	652	6.4
Word Study Skills	1	43	624	5.1
	2	44	632	5.4
	3	41	623	4.6
	4	45	635	5.6
Total Reading	1	80	628	4.9
	2	87	639	5.7
	3	83	632	5.2
	4	90	644	6.1

[b] Madden R. et al. (1983). Stanford Achievement Test Series: Multi-Level Norms Booklet National. Cleveland, OH: The Psychological Corporation. P. 38, 164. Modified Table reprinted by permission of the publisher.

Table 7
Probability of "success" (scoring higher then the median)
as a function of excess time

Subtests	5 minutes	10 minutes	15 minutes
(a) Students as unit of analyses			
Reading Comprehension	.60	.56	.64
Total Reading	.58	.52	.69
(b) Classes as unit of analysis			
Reading Comprehension	.73	.65	.80
Total Reading	.79	.62	.86

Table 8
Expected Gain Scores Resulting From Excess Time:
Difference Between Adjusted Treatment Mean and Adjusted Control Mean

Subtests	Excess Time		
	5 minutes	10 minutes	15 minutes
Expected Raw Gains			
Reading Comprehension	4.41	2.70	6.52
Total Reading	6.58	2.21	9.70
Expected Standardized Gain			
Reading Comprehension	.38	.23	.56
Total Reading	.32	.11	.83
Expected Gain, Standardized by Standard Deviation of Classroom Mean Gains			
Reading Comprehension	1.03	.03	1.52
Total Reading	1.42	.48	2.10

Table 9a
Intercorrelations Among Stanford Tests For Primary 3,
Form E At the Beginning Of Grade 4
(N=5011)

Test Name	Variable	2	3	4	5	6	7	8	9	10	1
Reading Comprehension	1	.68	.61	.61	.73	.69	.74	.80	.77	.62	.6
Word Study Skills	2		.60	.60	.72	.70	.75	.72	.71	.57	.5
Concepts of Number	3			.66	.72	.52	.68	.62	.61	.60	.6
Mathematics Computation	4				.70	.59	.68	.63	.60	.49	.4
Mathematics Applications	5					.64	.78	.80	.77	.66	.6
Spelling	6						.72	.69	.66	.46	.4
Language	7							.79	.77	.63	.6
Social Science	8								.86	.69	.6
Science	9									.69	.7
Vocabulary	10										.7
Listening Comprehension	11										

Table 9b
Intercorrelations Among Stanford Tests For Intermediate 1,
Form F Of The Beginning of Grade 5
(N=5858)

Test Name	Variable	2	3	4	5	6	7	8	9	10	11
Reading Comprehension	1	.64	.66	.57	.68	.70	.69	.76	.73	.66	.62
Word Study Skills	2		.70	.61	.71	.73	.75	.70	.69	.57	.56
Concepts of Number	3			.70	.81	.64	.72	.72	.68	.59	.59
Mathematics Computation	4				.69	.62	.66	.62	.59	.42	.44
Mathematics Applications	5					.66	.74	.77	.72	.62	.62
Spelling	6						.75	.72	.68	.50	.49
Language	7							.79	.75	.59	.61
Social Science	8								.84	.69	.67
Science	9									.66	.64
Vocabulary	10										.76
Listening Comprehension	11										

Source: Reprinted with permission of The Psychological Corporation. These tables have been slightly modified.