DOCUMENT RESUME

ED 390 880 TM 023 945

AUTHOR Rudman, Herbert C.; Raudenbush, Stephen W.
TITLE The Effect of Exceeding Prescribed Time Limits in the Administration of a Standardized Test of Reading Comprehension and Mathematics Applications. Research Series. CEPSE/No. 6.
INSTITUTION Michigan State Univ., East Lansing. Dept. of Counseling, Educational Psychology, and Special Education.
PUB DATE May 87
NOTE 33p.; For related studies, see TM 023 944-946.
PUB TYPE Reports - Research/Technical (143)

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

ABSTRACT

This study is the second in a series designed to explore the probable consequences of exceeding the prescribed time limits in the administration of standardized achievement tests. This study considered whether the test user could apply norms that accompany the test if departures were made from established time limits, and whether increased testing time would affect student benefits in the form of placement or teacher benefits resulting from improved student achievement results. The Reading Comprehension and Mathematics Applications subtests of the Stanford Achievement Test were used for 471 Lansing, Michigan fifth graders nested in block-by-treatment combinations. There was a highly significant positive linear effect of excess time on Reading Comprehension test scores, repeating the effect in the previous study, but no significant effect on mathematics scores. Results suggested that norms lost their significance when time was increased and that the Reading Comprehension test is time sensitive. Excess time increased the probability of teacher "success" if measured by student achievement, and of student benefit, in terms of placement or ranking. Recommendations for careful test monitoring are presented. (Contains 3 figures, 10 tables, and 25 references.) (SLD)

THE EFFECT OF EXCEEDING PRESCRIBED TIME LIMITS IN THE ADMINISTRATION OF A STANDARDIZED TEST OF READING COMPREHENSION AND MATHEMATICS APPLICATIONS

Herbert C. Rudman
Stephen W. Raudenbush

CEPSE No. 6 May 1987

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 A STANDARDIZED TEST OF READING COMPREHENSION AND MATHEMATICS APPLICATIONS

Herbert C. Rudman
Stephen W. Raudenbush

CEPSE/No. 6 May 1987

THE EFFECT OF EXCEEDING PRESCRIBED TIME LIMITS IN THE ADMINISTRATION OF A STANDARDIZED TEST OF READING COMPREHENSION AND MATHEMATICS APPLICATIONS

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

This study is the second in a series designed to explore the probable consequences of exceeding the prescribed time limits in the administration of standardized achievement tests. The procedures used to establish time limits for standardized achievement tests have long been familiar (Boag & Neild, 1962; Daly & Stahmann, 1968; Lord, 1956; Nunnally, 1978). As we pointed out in our first study (Rudman & Raudenbush, 1986), one issue is whether these procedures are adequate. A second issue is that the high-stakes uses to which tests have been put make it crucial to examine the consequences of violating test administration procedures (Elliott & Hall, 1985; Tirozzi, et al., 1985; District Quality Instruction Incentive Program, 1985). If test results are sensitive to moderate departures from those procedures, decision-makers must either avoid high-stakes uses or take strong measures to standardize test procedures.

A previous study (Rudman & Raudenbush, 1986) used the Stanford Achievement Test to examine the effects of excess testing time on achievement in Reading Comprehension and Word Study Skills. No significant time effect was noted in the Word Study Skills subtest, but a significant linear effect was noted in Reading Comprehension. This led the investigators to consider three questions related to the violation of prescribed time limits: (1) Can a test 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 because of their class' higher test scores?

Impact of excess time on norms. Once procedures for establishing the norms are violated, the norms lose their practical significance. Our previous study (1986) indicated that when testing Reading Comprehension, five additional minutes of testing time beyond the recommended time can yield an additional 0.9 grade equivalent score difference and as much as a 1.4 grade equivalent score after fifteen minutes. The data from our second study showed a similar relationship between additional testing time and grade equivalent scores (See Table 6). It clearly would be unwise to use published norms when test taking time is at variance from the established time limit.

Impact of excess time on probability of success. There are a number of different circumstances in which test data could play an important role in determining the difference between "success" and "failure". The question of probability of success was viewed from two perspectives; time and the unit of analysis (the individual or the group).

We posed two hypothetical situations. On the one hand, we envisioned a student being assigned to a high ability reading program on the basis of the data we collected. Using Rosenthal and Rubin's "Binomial effect size display" (1982), our earlier study indicated that without excess testing time a randomly selected student would have a probability of "success" of 0.50. With excess time that probability would increase to .64 if fifteen extra minutes were allowed for the Reading Comprehension subtest.

On the other hand, we hypothesized that a teacher would be given some reward if his or her classroom achieved some particular gain score from previous testing. If the classroom was used as the unit of analysis instead

of an individual student, the effect size would be measured in units of the standard deviation of classroom gains. Without the benefit of excess time, the probability of success for a randomly selected classroom is .50; with fifteen additional minutes of excess time that probability would increase to 0.80 for Reading Comprehension.

These results suggested that decisions made at the classroom level were substantially more sensitive to the effects of excess time than were student level decisions. We concluded that "...discussions about the effect of irregularities in test administration on test validity must take into account the level of aggregation at which test results are used in decision making." (1986, p.12).

The Relationship of Test Characteristics to Treatment Effects

Given the previous research and theory underlying the timing of power tests, we had originally hypothesized that excess time would not seriously influence test scores. Nevertheless, we undertook the study because we felt that the high-stakes uses of tests require a rigorous validation of testing time. However, the finding of a non-trivial linear effect of excess time in the case of Reading Comprehension compels a more careful consideration of the procedures used to set time limits.

The time limits for each of the subtests of the Stanford Achievement Test have been determined by the "90% criterion." Using this criterion, testing time is the time elapsed until 90% of the examinees complete the item analysis edition of the instrument. Nunally has referred to this elapsed time as the "comfortable time limit" (Nunally, 1978, pp.632-633). The assumption underlying the procedure is that the 90% completing the test had arrived at correct answers to many of the items and had used informed guessing on the remainder. It is assumed that the remaining 10% who had not completed the test would merely employ random guessing if given more time.

Hence more time would not translate into mean test score gains.

However, this "90% criterion" is, at best, a rough approximation to optimal testing time for a power test. It supplies no empirical evidence that examinees could not make use of further time; in some instances, it may extend testing time needlessly. What we did not know from the results obtained in our previous research, nor do we know now, is whether those results constitute an artifact of the particular test used or whether the extended reading passages of the Reading Comprehension subtest present a unique timing problem.

A more rigorous approach to setting time limits for a power test would consider the functional form of the relationship between elapsed time and test scores. If a test is truly a power test, students will make good use of elapsed time up to a point. At that point, the effect of further time will diminish because the students' knowledge relevant to responding to the test items will become exhausted. Thus, the functional form of the relationship between time and test scores ought to be quadratic with negative curvature: the slope of the curve describing that relationship gradually decreases and becomes null. The optimal testing time in a power test is then the point after which further time is unhelpful (See Figure 1).

As a result of this reasoning, one goal of this series of studies is to discover the optimal testing time by estimating the functional form of the relationship between excess time and test scores for those subtests found susceptible to the effects of excess time. Tests that end before this optimal time cannot validly be considered power tests. Moreover, such tests may be sensitive to variations in test administration procedures and hence be unsuitable for high-stakes use in settings where stringent administrative control is impossible.

PROCEDURES

Test Content

The Stanford Achievement Test, Intermediate 1, Form F was used to collect student achievement data on two subtests, Reading Comprehension and Mathematics Applications. Achievement test results for the previous year were used as covariates. More specifically, these pretests resulted from the school district's routine administration of Primary 3, Form E battery of the same test series.

The Reading Comprehension subtest consists of 60 items. The Mathematics Applications subtest contains 40 items and is administered in one sitting. The maximum time limit for the Mathematics Applications subtest 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 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. Twenty-nine faculty members from 16 of these elementary schools volunteered to serve as participants in this study. However, data could be collected for only 23 of these classrooms. These 23 fifth grade classrooms supplied useable data for 471 pupils. Table 1 describes the sample on key demographic variables and test scores.

Each of the original 29 classrooms was assigned to one of seven blocks. Four treatment groups representing testing-time allotments were established. Within each of the seven blocks, classrooms were assigned at random to one of the four treatments. If the sample had included 28 classrooms, the design would have been a balanced randomized block design (Kirk, 1982, Chapter 6)

with seven blocks and four treatments. Instead, one block included five classrooms.

The blocks were constituted to be as homogeneous as possible on prior reading and mathematics test scores and on socioeconomic status as indicated by the proportion of families receiving aid to dependent children. The blocking also assured that no two classrooms within the same school would experience the same treatment. The blocking variable could be viewed as an ordinal variable, ranking the classrooms on background variables related to the outcomes. 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 5 minute intervals for the total testing-time allocated across all parts of the two subtests. Unadjusted and adjusted 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, $r = .75$. For the Mathematics Applications subtest, the Mathematics Applications pretest proved to be the best covariate, $r = .78$. For Total Reading, the Total Reading pretest proved best, $r = .82$. Only one covariate per outcome was

needed since other likely covariates (ethnicity, parent education, native language, sex, and family configuration) were not significantly related to the outcome after adjusting for the effect of the "best" covariate.

The effect of the blocking variable was studied in two ways. First, a single-degree-of-freedom linear effect was tested. Second, the non-linear effect of blocks was examined. In the case of Reading Comprehension, only the linear effect was statistically significant after controlling for the effect of the covariates (Table 5). In the case of Mathematics Applications, neither effect was significant.

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 outcome, effects were entered in the following order: the covariate; then the linear block effect; 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 23 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 covariance

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 Reading Comprehension, $F(18, 447) = 1.70$, $p < .05$, and for Mathematics Applications, $F(18, 447) = 1.85$, $p < .05$.

RESULTS

The key finding of the study was a highly significant, positive, linear effect of excess time on Reading Comprehension test scores, $F(1,18) = 8.09$, $p < .02$. There was no significant linear effect for Mathematics Applications, and, in fact, there were no significant differences among treatment means for that subtest, $F(3,18) = .62$. These results appear 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 = .11; between excess time and the adjusted posttest score, r = .17 (see Table 5). However, at the classroom level, the partial correlation between excess time and the adjusted posttest classroom mean is r = .52, so that 27 percent of the adjusted between-classroom variation is attributable to this linear trend. 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.

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

DISCUSSION

The Results

An intriguing outcome of this study is the replication of results concerning the effects of Reading Comprehension. The linear effect observed in our first study (1986) was essentially repeated. The slope of the linear effect in the first study was b = .66. In the second study the slope was b = .50. These two slopes were not significantly different, t = .73. The pooled slope was b = .60, and the pooled test of the hypothesis of no linear effect of time yielded $z = 5.58$, $p < .004 \times 10^{-7}$. The absence of any significant quadratic effect led us to believe that that point when more testing time would be unhelpful had not yet been reached in the Reading Comprehension subtest, and had probably been reached at some earlier point in Mathematics Application (See Figures 2 and 3).

In initially formulating this series of studies we assumed that the traditional approach taken to the timing of standardized achievement tests

was adequate for the development of power tests. The times established by the use of the "90%" criterion seemed appropriate. The reporting of cooperating teachers during national item tryouts served as an empirical base from which to establish maximum time limits. Certainly the data gathered in our studies (1986, 1987) would indicate that no significant effect of excess test-taking time can be found in the achievement scores of two subtests, Word Study Skills and Mathematics Application. While we are not yet prepared to say that the measurement of other subject areas will yield similar null linear effects using traditional timing procedures, we are inclined to suspect that the measurement of Reading Comprehension as reflected in the standardized test used in our investigations may present a special issue.

It is not clear yet whether our results reflect (1) a characteristic of a higher order of intellectual skills associated with Reading Comprehension, (2) special characteristics associated with the Reading Comprehension subtest of this particular test series, (3) a phenomenon related to the grade level tested, or (4) a particular characteristic of the curriculum of the school district in which the research has been conducted. We doubt that our results are peculiar to this test series or to this school district.

We suspect that a plausible explanation for our results may be that the specific skills associated with the measurement of Reading Comprehension require more time than has been allocated under the 90% criterion, and that other subtests may reflect skills which really need less time for the number of items used. Our results would indicate that the Reading Comprehension subtest is sensitive to extended test-taking time. Even after a 50% increase in total maximum time, we continue to obtain a non-trivial effect of time on achievement.

A consistent finding of both studies (1986, 1987) we have conducted is that the unit of analysis (data based on individual scores contrasted with group data) is differentially affected by excess time. 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. We have hypothesized three key decision points; the median (Table 7), the 80th percentile (Table 8), and the 20th percentile (Table 9).

Table 7a shows how excess testing time influences the probability of student success of achieving above the median using the more conservative results from the 1987 study. Without the benefit of excess time, a randomly selected student would have a probability of "success" of .50. With excess time that probability would increase to .58 if the test used was Reading Comprehension.

If the teacher were to be rewarded in some manner for her classroom's achievement, Table 7b shows how excess time would influence the probability of a teacher's "success". Now the effect sizes are measured in units of the standard deviation of classroom performance. 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 .71 when the test used is Reading Comprehension.

If we were to assume that a school district wished to place qualified students into an enrichment program and set the cut-off score at the 80th percentile Table 8a indicates how excess time would influence the probability of a student being chosen for the enrichment class. Without excess testing time a randomly selected student would have a probability of 0.20 for movement into an enrichment program if the criterion measure was the Reading Comprehension subtest used in this study. With 15 additional minutes of testing time, the probability would increase to 0.31.

However, if a school district were to award teachers for classrooms that showed an annual gain that placed them in the top 20% of all classrooms within a district, the probability of success for a randomly selected teacher would rise from .20 to .46 with 15 additional minutes of extra time for the Reading Comprehension subtest (Table 8b).

Finally, we envisioned a scenario in which students were given a standardized achievement test to screen the lowest 20% of the students for a remedial reading program. If the subtest used is sensitive to excess teaching time, and if recommended testing times are not strictly followed what would be the probabilities that students might be misclassified? A randomly selected student who received the recommended amount of testing time would have a .80 probability of scoring above the 20th percentile. The chance of that student scoring above the 20th percentile would increase to 0.88 with 15 additional minutes using the Reading Comprehension subtest (Table 9a). If the classroom were the unit of analysis, for instance if teachers were being screened for remedial supervision this probability would rise to .96 with 15 additional minutes of testing time.

We are not advocating the violation of procedures established for the administration of standardized achievement tests. We are interested, however, in the consequences of not following standardized recommendations. If testing times are not meticulously followed when scores will be used to make high-stakes or high-risk decisions, unfair advantages can accrue. These advantages are magnified when decisions are made at the classroom level. "Rule of thumb" procedures for timing tests traditionally followed in test construction may be insufficently accurate in the context of such high-stakes use, making some tests "under-timed" and others "over-timed".

Time Limits and Achievement

There has been a surprising paucity, in the past twenty years, of research dealing with the relationship between time limits and test scores. The last intensive examination of this problem was a symposium published in Educational and Psychological Measurement, 1960, 20,221-274. Test developers have generally accepted the technique of establishing time limits based on observations of some predetermined proportion of students who complete the test. Although we have referred to the convention identified as the "90% criterion" this proportion has varied somewhat between test publishers. This timing procedure has generally been satisfactory for the purposes to which standardized achievement tests have been put i.e., the monitoring of individual and class achievement, systematic analysis of curricular strength and weaknesses, orientation to new students entering the class, reporting to parents, and the like.

The "90% criterion" produced the "comfortable time" limit described by Nunnally (1978, p. 632), but it also introduced a lack of precision in the determination of optimum time limits for power tests. This imprecision has generally erred on the generous side and offered no serious consequences to the test taker, or the teacher. The context within which these tests are now used alters this relatively sanguine picture. Teachers today complain of too many tests given on too many occasions. They are concerned about the time devoted to testing as contrasted with teaching. On the one hand, test users call for maximum information from the time they devote to testing, but on the other hand, they want to get this information in less time.

Less testing time can be brought about by either shortening the test or by decreasing the maximum recommended testing time. Under present test construction practices these alternatives present disadvantages. Shorter tests generally produce lower reliability coefficients. They also present a practical limit on "maximum" information which can be derived. Decreasing

the amount of recommended maximum time can change a standardized achievement test from a timed power test to one that is speeded. If our experience with Reading Comprehension serves as a guide, the use of conventional ways of determining optimum testing time is too imprecise for the critical decisions made with standardized test data. Our results lead us to the conclusion that a more precise way of determining optimum time limits is needed. We believe that teacher use of the "90% criterion" during item analysis should be replaced with a national research program similar to the design we have here employed. This will more accurately pin-point the time when student's responses to test items become random guesses. That point will become the comfortable time limit of a power test and may very well result in a decrease in overall achievement testing time. In other words, rather than shortening a test and losing both reliability and information, we may achieve the desired effect by more effective research during a national item analysis.

RECOMMENDATIONS

1. Test designers should consider more precise methods of setting testing-time limits. Approximation to optimal times based on "rules of thumb" such as the "90% criterion", are not sufficient for the critical uses to which tests are presently put.
2. Our purpose in this work is not to evaluate the legitimacy or fairness of high-stakes uses of tests. However, tests which are sensitive to variations in the procedures of administration should not be used for high-stakes decisions, particularly for decisions about teachers, and schools, unless these procedures can be carefully monitored.

3. Comprehension of printed material is a complex of a number of skills which contribute to functional understanding of the printed word. The student is asked to read textual, recreational and informational materials and to draw certain inferences which demonstrate understanding. Can a Reading Comprehension subtest truly be a "power" test, or is there an inherently "speeded" character to reading comprehension itself? Our results suggest no diminution of effects of time even after recommended maximum time limits have been exceeded by 50%. Hence to convert a standardized test of reading comprehension into a true power test would require either eliminating many items or greatly lengthening the time limit.

However, if reading for comprehension by definition implies comprehending within a reasonable but fixed time, can it be tested by a timed power test? The answer to this question has implications both for test development and use. Items in a standardized achievement test are ordered by difficulty. We would recommend that further experimentation be done with existing tests of reading comprehension to determine whether some types of reading passages are more sensitive to testing time than others. If specific types can be identified as those which contribute to speededness ought they to represent a smaller proportion of the extended passages than they might otherwise occupy?

4. Further research is needed on other test series, battery levels, and subjects to determine if our results will be replicated. Studies using less than the recommended maximum time might help to reduce unnecessary testing time for those subject areas not yet examined by us.

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.

Bridges, K.R. (1985). Test completion speed: Its relationship to performance on three course-based objective examinations. Educational and Psychological Measurement, 45, 29-35.

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 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: 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	Relative Frequency
1. Sex	Male	45.5
	Female	54.5
2. Ethnicity	White (non-Hispanic)	60.1
	Other	39.9
3. Family Configuration	One-parent	38.1
	Two-parent	61.9
4. Parent Education	Elementary	2.0
	Junior High	3.7
	High	16.3
	High graduate	40.3
	Attended college	24.5
	College graduate	7.9
	Post graduate	5.2
4. Native Language	English	90.5
	Other	9.5

Test Data				
Subtest	Pretest		Posttest	
	m	sd	m	sd
Reading Comprehension	35.40	12.57	41.42	10.19
(Total Reading)	(76.39)	(19.83)	----[a]	----[a]
Math Application	25.80	8.58	28.07	7.58

[a] Total reading scores were available at the pretest, but not the post test.

Table 2
Testing Time by Treatment and Subtest

Treatment	Reading Comprehension	Mathematics Application
1	30	35
2	35	40
3	40	45
4	45	50

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	-----	-----	-----	43.08	43.08
Block 2	37.84	40.36	42.30	43.28	41.02
Block 3	37.96	40.91	38.40	-----	39.04
Block 4	40.71	-----	40.88	46.25	42.61
Block 5	36.50	45.23	40.57	43.52	41.42
Block 6	42.56	42.80	48.47	-----	44.12
Block 7	37.52	37.69	46.53	39.21	40.20
Overall	38.95	41.70	42.40	43.23	
Mathematics Application					
Block 1	-----	-----	-----	26.92	26.92
Block 2	27.32	25.45	29.07	27.11	27.35
Block 3	25.32	29.04	24.30	-----	26.24
Block 4	30.67	-----	27.13	29.63	29.14
Block 5	25.77	30.55	27.52	27.10	27.74
Block 6	30.52	29.50	33.13	-----	30.83
Block 7	27.14	29.92	30.29	26.11	28.14
Overall	27.80	28.79	28.31	27.45	
Sample Sizes					
Block 1	--	--	--	24	24
Block 2	19	22	27	18	86
Block 3	28	23	20	--	71
Block 4	24	--	24	24	72
Block 5	22	22	23	21	88
Block 6	25	20	15	--	60
Block 7	21	13	17	19	70
Overall	139	100	126	106	471
Adjusted Treatment Means (Unstandardized)					
Reading Comprehension	39.53	41.44	42.14	43.03	41.42
Mathematics Application	27.67	28.52	27.79	28.49	28.07
Adjusted Treatment Means (Standardized)					
Reading Comprehension	-.186	.002	.070	.158	
Mathematics Application	-.053	.059	-.038	.056	

Table 4
Analytic Method

Model/Source	df residual	Sum of squares residual
Constant	470	(1)
Constant, covariate	469	(2)
Constant, covariate, blocks	468	(3)
Constant, covariate, blocks, linear trend	467	(4)
Constant, covariate, blocks, Linear, Quadratic trends	466	(5)
Constant, covariate, blocks, linear, quadratic, cubic trends	465	(6)
Constant, Covariate, classrooms	447	(7)

Partitioning of variation

	df reduction	Sum of Squares reduction
Covariate	1	(1) - (2)
Between Classes (adjusted)	22	(2) - (7)
Blocks	1	(2) - (3)
Treatments	3	(3) - (6)
Linear	1	(3) - (4)
Quadratic	1	(4) - (5)
Cubic	1	(5) - (6)
Residual (blocks by treatments)	18	(6) - (7)
Within Classes (adjusted)	447	(7)
Total	470	(1)

Table 5:
ANCOVA source tables for a) Reading comprehension, and Mathematics Applications .

Source	df	Sum of Squares	Mean Square	F	ETAa	ETAb
Reading Comprehension						
Covariate	1	27211.561	2741.560			
Between Classes	22	2259.332	102.697			
Blocks	1	330.972	330.972			
Treatments	3	607.108	202.369	2.757*	.11	.17
Linear	1	594.006	639.003	8.09***	.11	.17
Quadratic	1	12.983	12.983	.177		
Cubic	1	.122	.122	.0017		
Residual	18	1321.252	73.403	1.699**		
Within Classes	447	19315.711	43.212			
Total	470	48786.603	103.801			
Mathematics Applications						
Covariate	1	16477.230	16477.230			
Between Classes	22	802.767	36.489			
Blocks	1	1.919	1.919			
Treatments	3	75.224	25.075	.622		
Linear	1	22.75	22.75	.56		
Quadratic	1	0.117	0.117	.0029		
Cubic	1	52.356	52.356	1.299		
Residual	18	725.623	40.312	1.851**		
Within Classes	447	9733.819	21.776			
Total	470	27013.826	57.476			

* P<.10
** p<.05
*** p<.02
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 b Scores	Grade Equivalentb
Reading Comprehension	1	40	635	5.2
	2	41	638	5.4
	3	42	642	5.7
	4	43	645	5.9
Mathematics Application	1	28	630	5.4
	2	29	635	5.6
	3	28	630	5.4
	4	29	635	5.6

b

Madden R. et al. (1983). Stanford Achievement Test Series: Multi-Level Norms Booklet National. Cleveland, OH: The Psychological Corporation. P. 38, 164-165. Modified Table reprinted by permission of the publisher.

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

Subtest	5 minutes	10 minutes	15 minutes
(a) Students as units of analyses			
Reading Comprehension	.55	.56	.58
(b) Classes as units of analysis			
Reading Comprehension	.62	.67	.71

Table 8
Probability of "success" (scoring higher than the 80th percentile)
as a function of excess time

Subtest	5 minutes	10 minutes	15 minutes
(a) Students as units of analyses			
Reading Comprehension	.25	.28	.31
(b) Classes as units of analysis			
Reading Comprehension	.37	.44	.46

Table 9
Probability of "success" (scoring higher than the 20th percentile) as a function of excess time

Subtest	5 minutes	10 minutes	15 minutes
(a) Students as units of analyses			
Reading Comprehension	.85	.86	.88
(b) Classes as units of analysis			
Reading Comprehension	.91	.94	.96

Table 10
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
(a) Expected Standardized Gain			
Reading Comprehension	.184	.252	.340
(b) Expected Gain, Standardized by Standard Deviation of Classroom Mean Gains			
Reading Comprehension	.512	.702	.947

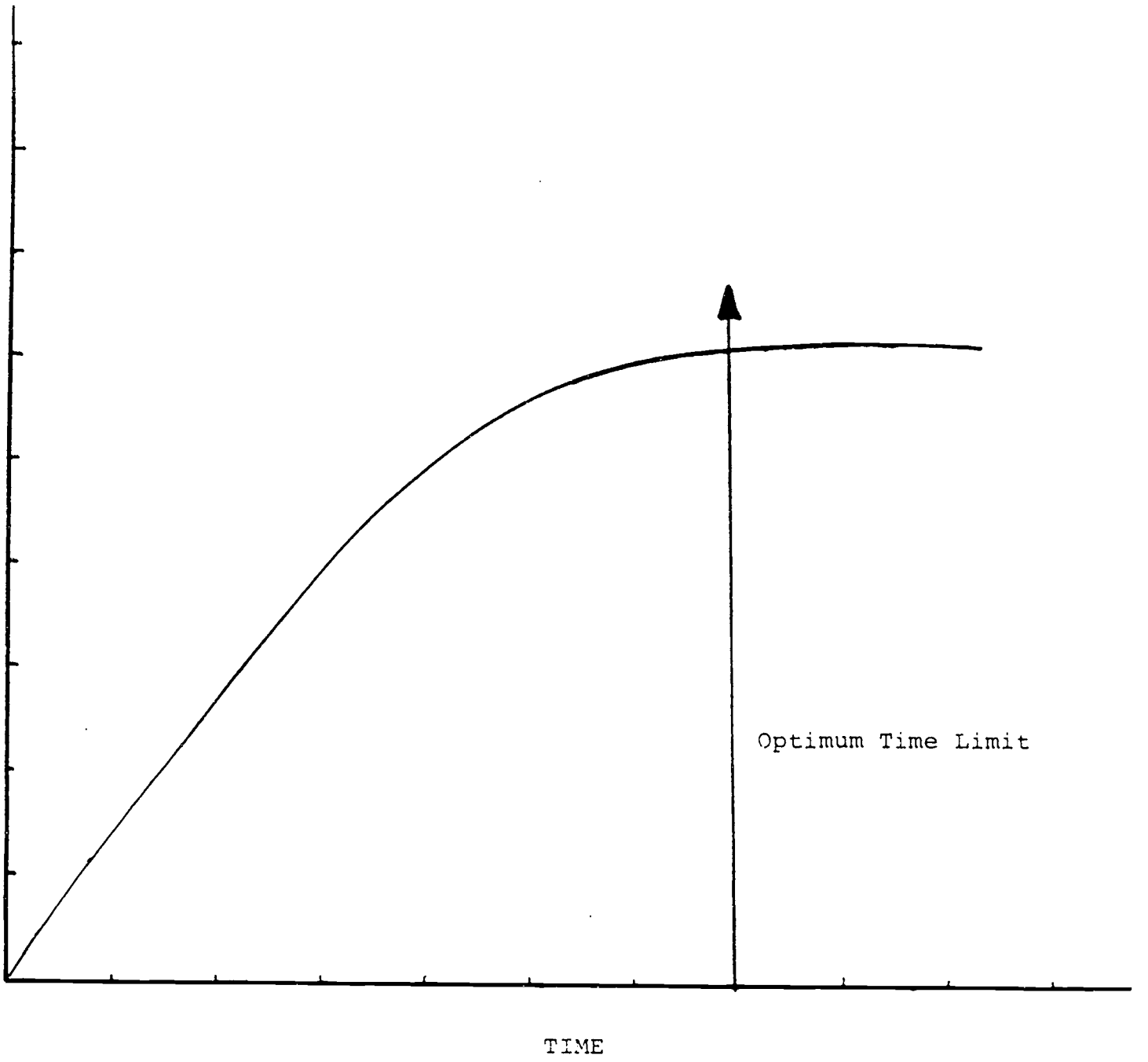

Figure 1: A Model for Determining the Optimal Time Limit of Timed Power Tests

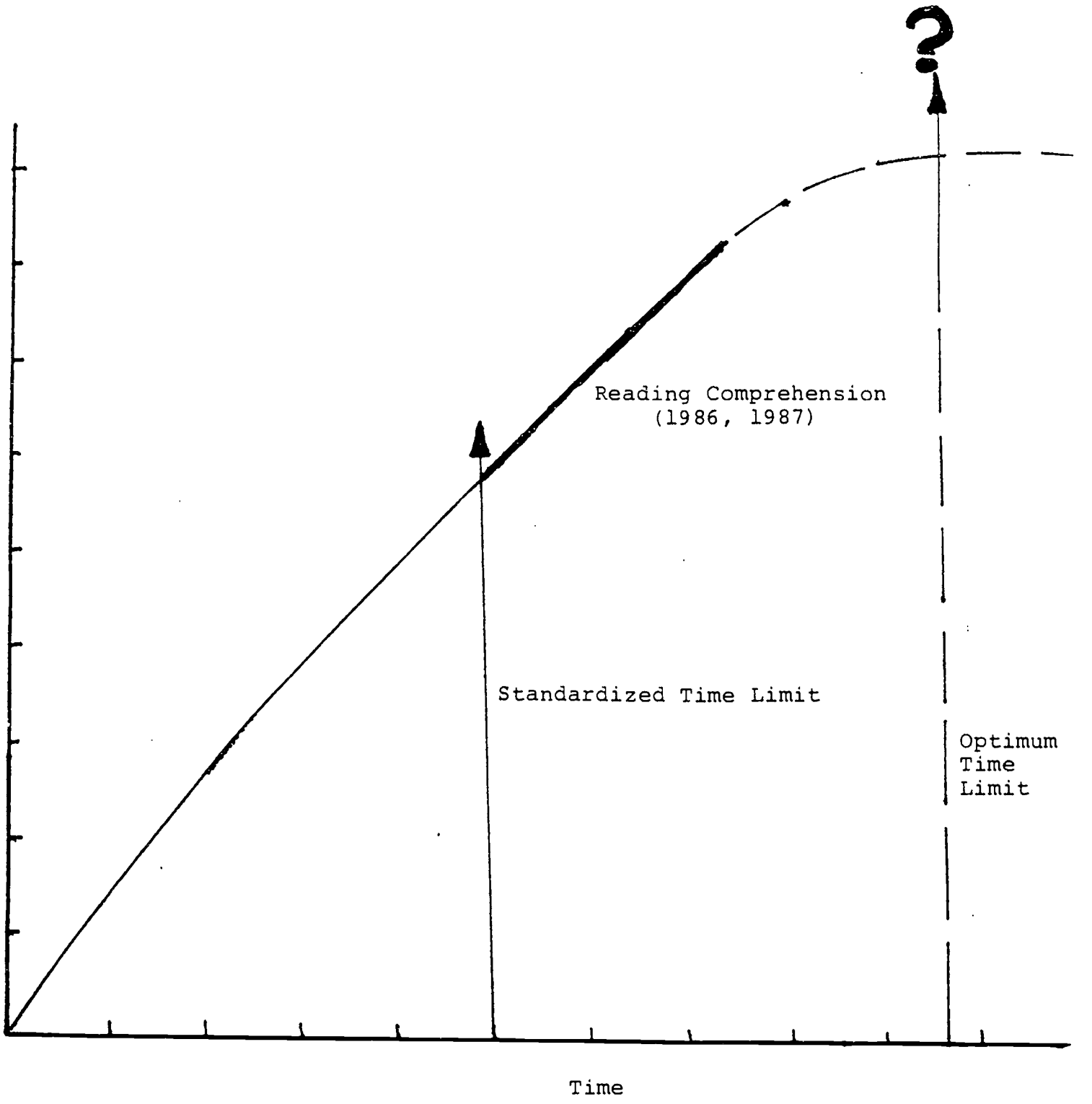

Figure 2: An Illustration of a Subtest That is Sensitive to Excess Testing Time

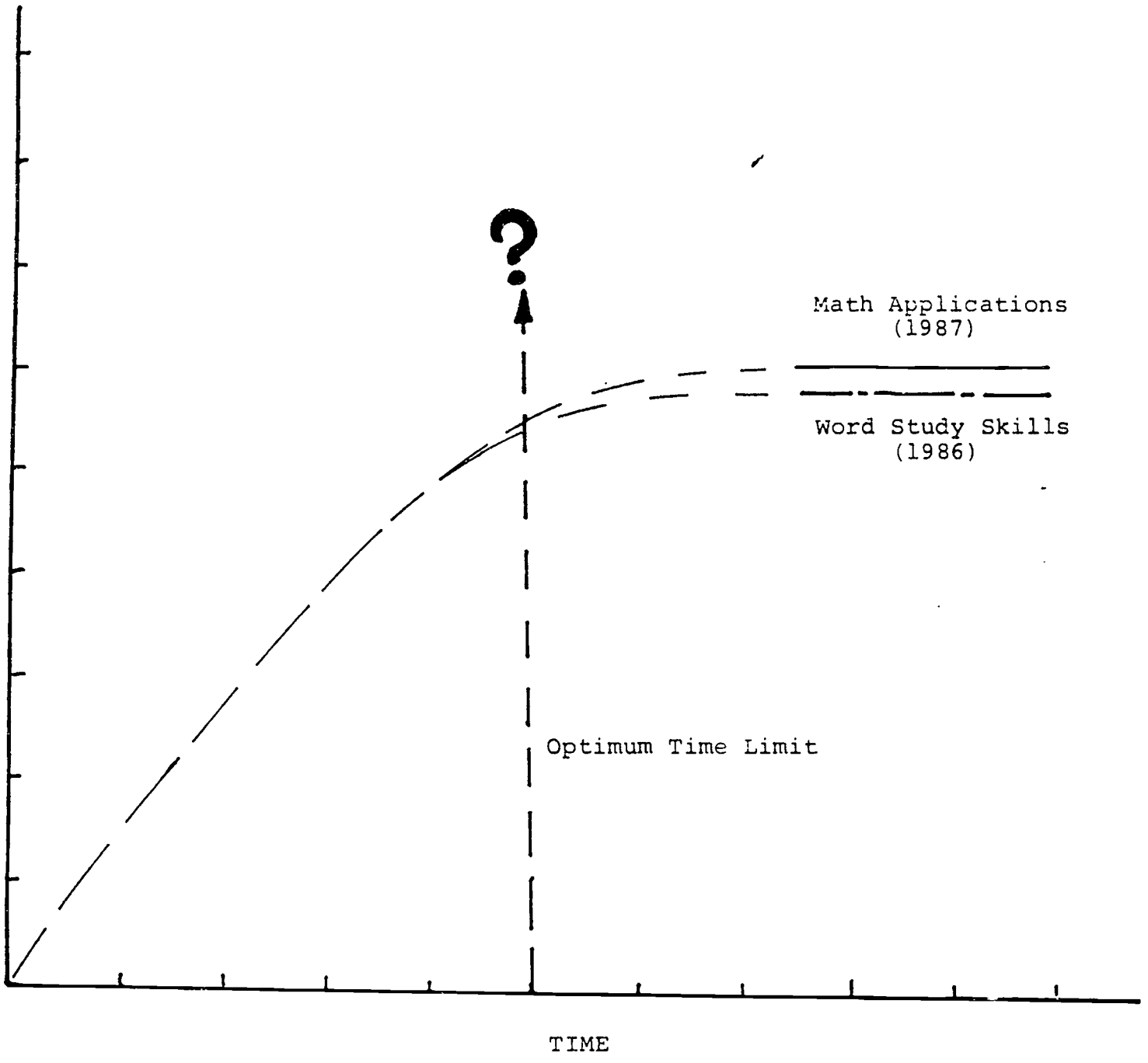

Figure 3: An Illustration ofTwo Subtests
Not Sensitive to Excess Time