USE OF SIMULATION FOR EXAMINING THE EFFECTS OF GUESSING UPON KNOWLEDGE ASSESSMENT ON STANDARDIZED TESTS

James E. Bruno

ABSTRACT

The standardized test for the assessment of cognitive skills has become the principal output variable used by economists and instructional evaluators for studying schooling. Its widespread use in instructional evaluation has raised concerns by researchers in the areas of both group and individual knowledge assessment. Because of its dominance as a measure of output of schooling, it is essential that the caveats and assumptions underlying their use be explored. The purpose of this paper is to discuss the limitations in the interpretation of standardized test results, especially as they relate to quality issues in schooling, and individual knowledge assessment. Specifically, the paper will focus particular attention on the use of a monte carlo simulation procedure for studying the effects of student guessing on the individual assessment of knowledge. For this study, test behavior was simulated with differing guess patterns, extent of guessing, and the number of choices per test item, to examine its effect upon the individual assessment of knowledge. Combining the results of the simulation study with Bayesian analysis, a calibration table containing a probability distribution of knowledge score given a person's observed score was derived.

The study indicated that guessing severely contaminates and biases upward the individual assessment of knowledge at the elementary and secondary levels. One immediate implication is the curricula building upon incomplete knowledge bases, with resulting student failure later in schooling. The paper concludes with a recommendation for establishing testing procedures for individual knowledge assessment and directs attention of educators to inner city school situations where student guessing would be expected to be more prevalent and where testing is used more extensively.

I. INTRODUCTION

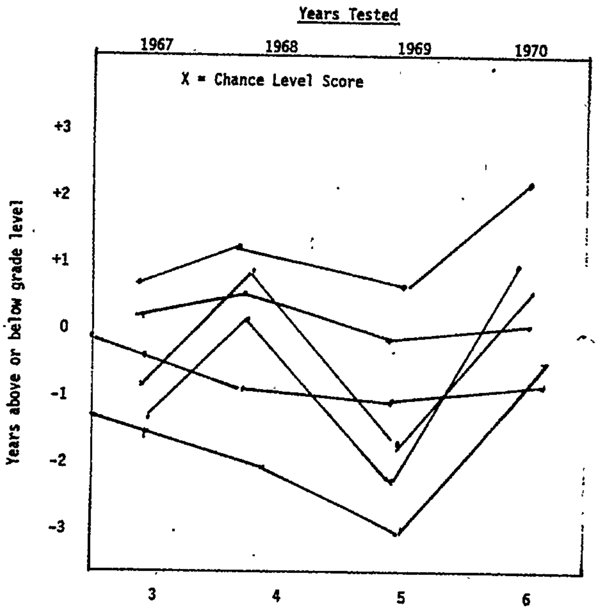
The use of monte carlo computer simulation for planning and managing in education has been applied to such important problem areas of educational forecasting (4, 11, 12), benefit-cost analysis (3, 1), and examining the assumptions of important statistical methodologies used by educators (5, 8, 16). The purpose of this paper is to examine how monte carlo computer simulation techniques are being used to determine the effects of guessing upon the assessment of knowledge on standardized tests, especially as it relates to the individual assessment of knowledge.

Many teachers have observed students taking standardized tests who guess wildly and are obviously disinterested in the testing process, due either to poor motivation or fatigue. The noise factor introduced in a testing situation has caused serious concerns among educators--especially those concerned with the individual assessment of knowledge (such as counselors and teachers) rather than overall school district performance. In fact, Rosenthal and Jacobsen (5) in Pygmalion in the Classroom, and Bruno (2), Emerging Issues in Education, focused attention on this basic point in testing situations and underscored the problems of meaningful test score interpretations for these students. This study directs attention at two issues of increasing importance for those concerned with evaluating instructional programs in the inner city: first, the "noise" factor in norm-referenced standardized testing, and second, the confidence one places in observed test scores as reflecting actual knowledge of an individual student. The "noise" factor appears dominant in precisely those areas where, due to federal intervention programs, testing is widespread and test scores are used for evaluation of programs, the design of new curriculums as well as assessing individual growth. Thus, directors of instructional programs focusing on low-achieving students, especially Title III programs in inner city schools, collect data, and perform evaluation in which an inordinate amount of measurement error exists. Quite possibly because of this noise component some Title III programs evaluated as being successful are in fact unsuccessful. More important, however, the knowledge base upon which future instruction for these students is based might be seriously overstated, hence leading to student failure later in the educational process.

Pioneering work in the area of guessing and its effect upon the assessment of knowledge has been carried out by Shuford and Massengill (12, 13, 14, 18, 19, 20). Research in this area has also been completed abroad by de Finetti (6, 7) and Toda (23).

Almost every teacher will be able to document wild changes in an individual test score over a given grade span. Figure 1 shows the grade equivalent fluctuations in sixth-grade scores for individual pupils as found by Shuford.

FIGURE 1



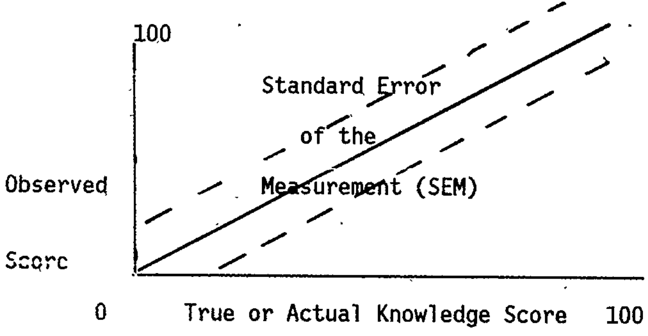
Year-to-year grade changes in sixth grade pupil's Grade Equivalent Scores compared with grade level on the Iowa Test of Arithmetic Problem Solving. Each line traces an individual pupil's progress (Shuford, 1971).

Before beginning a discussion of the monte carlo simulation analysis useo to study this problem, it is important to note some of the relationships assumed between a pupil's observed score on a standardized test and his or her actual knowledge.

Figure 2 shows the assumed relationship between actual knowledge score and observed score on a standardized test.

FIGURE 2

Assumed Relation Between True and Observed Knowledge Scores



Note from the above figure that common testing procedures and practices assume a uniform standard error of the estimate of true (SEM) knowledge score given the student observed score.

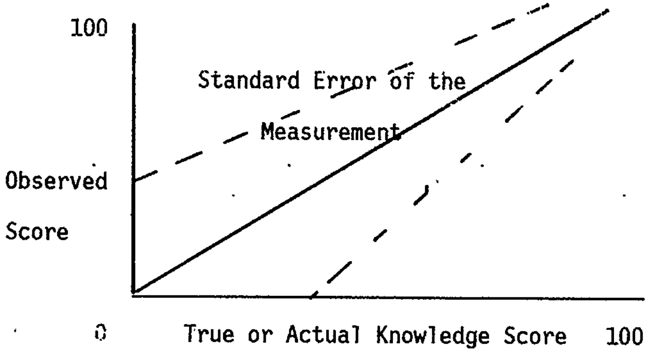
For example, if a student has an observed score of 3.0 grade equivalents and another student 9.0 GE; if the standard error or the precision for this test is reported to be .4 of a GE, then 3 ∓ .4 and 9 ∓ .4 reflect, according to test practices, the students' actual true knowledge score for the test. But is the SEM totally independent of student ability?

The standard error of the measurement in a student-observed score on a norm-referenced test has been studied by educational researchers with conflicting results. Thorndike (21) examined factors besides guessing which enter into the reliability or standard error of the measurement. The relationship between true score and standard error where only guessing is a factor was examined by Zimmerman and Williams (24). The above studies were purely theoretical in nature and did not examine explicitly the probability distribution of actual knowledge scores given a specific observed score.

Intuitively, as the achievement level increases, the amount of guessing should diminish (hence, the error), since the student actually possesses subject matter mastery and is able to distinguish between completely right and wrong responses. If the achievement level is low, however, one should expect the frequency of guessing to increase, since subject matter mastery is low and the ability to distinguish between choices on a particular question diminishes. Stated differently, the standaro error or test precision or partial knowledge should be negatively correlated with achievement level when dealing with the individual assessment of knowledge. This is cone shaped relationship demonstrated in Figure 3.

FIGURE 3

Probable Relation Between True and Observed Scores



Thorndike and Hagen (21), using the Lorge Thorndike Intelligence Test, present empirical data which support this hypothesis, namely, the standard error of the measurement at various score points grows larger with lower raw scores. Recently Dosher (9) in experimental work at UCLA has also corroborated this finding.

Thus, the test precision implications of this study, for evaluation of Title III programs whose efforts are directed specifically at the lower end of the achievement distribution, are extremely relevant for educational policy makers. An analytical inquiry, given certain guessing assumptions, of exactly how actual knowledge is distributed and distorted for each observed score with different guessing assumptions is the principal focus of the monte carlo study. The sensitivity of these probability distributions of actual knowledge to changes in these assumptions is also explored.

II. METHODS

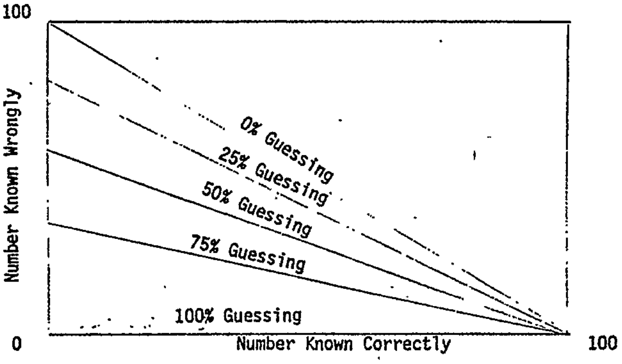
The following procedure was employed in this monte carlo study directed at examining the distribution of individual actual knowledge given various observed scores and guessing patterns. First, a 100-item, four-choice examination (different choices were later explored) was used as the basis for analysis. Second, a guessing pattern and extent of guessing* based upon the Shuford experiments (these parameters were later varied to test for sensitivity) was assumed in order to introduce noise into the observed test score. Third, a testing situation involving 10,000 students taking the above defined test was simulated on a computer and by use of Bayesian analysis and monte carlo procedures, an analytical determination for the distribution of possible actual knowledge scores given a student's observed score was obtained. Finally, the parameters such as extent of guessing, guess pattern, and number of choices were systematically varied to examine the sensitivity of these assumptions to the amount of measurement error introduced into the student observed score.

Figures 4 and 5 show the relationship between correct answers and wrong answers for various guessing extents while Figure 6 presents the mathematical relationships between guessing extent and the number of answers known correctly by the student. Row 3 of Table 1 and column 3 of Table 2 were used in this first part of the study. Note from Figures 4 and 5 that the expression "known wrongly" refers to a student who answers a question totally certain he is correct (not guessing or partial knowledge) and the answer is wrong.

Mathematically from the input data, monte carlo techniques, and the assumptions defined above, it was possible to obtain the distribution of student observed test scores (OS) for each given student actual knowledge score (TS):

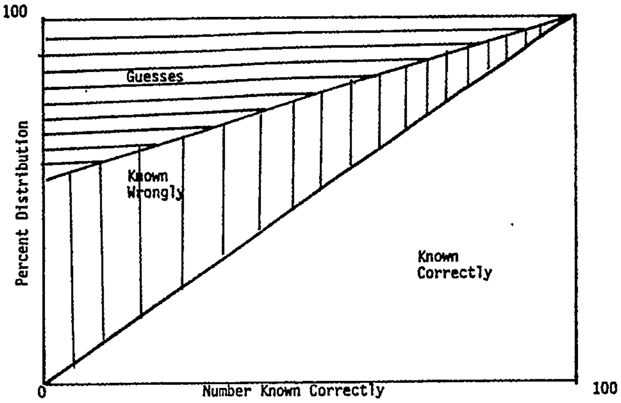
$$P(OS|TS)$$

FIGURE 4



Relation between number answered with 100% certainty and correct and the number known within 100% certainty (no guessing) and answered wrong.

FIGURE 5



Relation between correct answers, wrong answers, and guesses for 50 percent guessing extent.

If we also assume that the probability distribution of actual knowledge scores for the test is normally distributed (this assumption can also be varied to account for different student inputs), we can find the distribution of observed scores P(OS) by using the formula:

$$P(OS) = P(OS|TS)\ P(TS)$$

Finally, using Bayes' theorem we can derive the probability distribution of actual or true knowledge scores given an individual observed score

$$P(TS|OS) = \frac{(P(OS|TS)\ P(TS)}{P(TS)}$$

Monte carlo techniques were employed, of course, to mathematically simulate the correct response in a guessing situation.

*The guessing pattern refers to the partial knowledge of the student taking the test. For a particular item, can the student distinguish between the one, two, three, or more choices? Guessing extent refers to the overall number of items on the test where some guessing exists.

TABLE 1

Extent of Guessing

Situation	Number of Questions (NQ)	Number Totally Right (NR)	Extent of Guessing
1	NQ	NR	None
2	NQ	NR	Low
3	NQ	NR	Medium
4	NQ	NR	High
5	NQ	NR	Total

	Guess Factor	Number Totally Wrong (NW)	Number of Guesses (NG)

1	0	NW=1.00
		(NQ-NR)
			NG=NQ-NR-NW
			NG=0

2	25	NW=.75
		(NQ-NR)
			NG=NQ-NR-.75NQ
			+.75NR
			NG=.75NQ+.75NR

3	50	NW=.59
		(NQ-NR)
			NG=NQ-.50NQ+.50NR
			NG=.5NQ+.50NR

4	75	NW=.25
		(NQ-NR)
			NG=NQ-.75NQ
			+.75NR
			NG=.25NQ+.75NR

5	100	NW=0 (NQ-NR)	NG=NQ-NR

TABLE 2

Guess Pattern Used in Study

	2-Choice Test	3-Choice Test	4-Choice Test	5-Choice Test
Between 2 choices N_1	1.00NG	.75NG	.75NG	.75NG
Between 3 choices N_2	0	.25NG	.15NG	.15NG
Between 4 choices N_3	0	0	.10NG	.05NG
Between 5 choices N_4	0	0	0	.05NG

Between two choices generate
$0 \leq x \leq 1$
by means of a random number generator if
$x \leq .5$
answer is correct. For three choices if
$x < .33$
answer is correct; for four choices if
$x < .25$
answer is correct; for five choices if
$x \leq .20$
answer is correct.

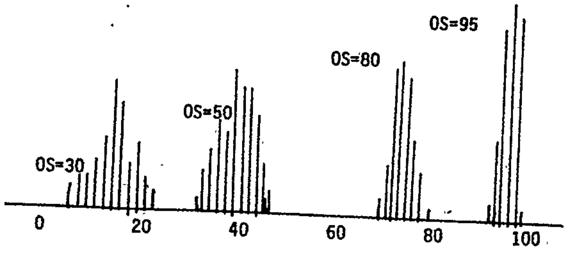
For complex binomials involving three, four, and five choices, the formula becomes extremely difficult to define. The monte carlo procedure was used to simulate the extremely complicated exact analytical formulas associated with complex binomial distributions and numerically approximate its solution. Within the accuracy of this study, the numerical monte carlo procedure closely approximates the exact analytical solution.

III. RESULTS

Notice from Figure 6 that the range in the distribution of student actual knowledge score decreases (hence cumulative probability distribution curve gets steeper) as we approach higher mastery levels numerically corroborating the results of the Thorndike-Hagen study. Common test practice, of course, would assume that the cumulative distribution curve is a line perpendicular to the X-axis at the student observed score, or that the student observed score equals his actual knowledge score. Notice how difficult, even with minor amounts of guessing as a noise factor, it would be to distinguish the actual knowledge score of a student with an observed score of 50.

FIGURE 6

Sample Histogram for Various Student-Observed Scores



It should be especially noted that the errors in individual observed test scores derived by the analytical procedure presented here tend to understate the actual error found in testing situations. This understatement is due, of course, to the relatively large number of test items (100-item test) used in this simulation. Many scales and parts of standardized tests commonly used by school districts have fewer than 20 items. The "noise" inherent in measuring individual test reliability for scales with this few items would be very high and make accurate interpretations for individual counseling virtually impossible.

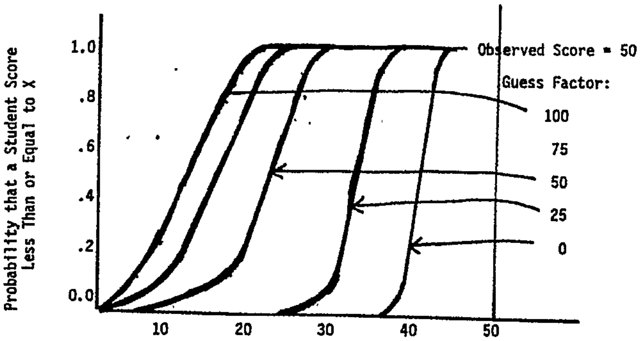
Extent of Guessing and Its Effect upon Test Precision

The extent of guessing on standardized tests is shown by Figures 7 and 8 to have a great effect upon test precision. These graphs summarize the probability distribution of true scores given a student observed score for guessing extents (the percentage of items which the student does not know correctly or wrongly) of 0, 25, 50, 75, and 100%.

Figure 7 is for an observed score of 50 and Figure 8 is for an observed score of 80. Notice the spread in the probability distribution of exact knowledge scores for an observed score P(TS|OS) varies directly as the guess factor and is inverse to subject matter mastery.

FIGURE 7

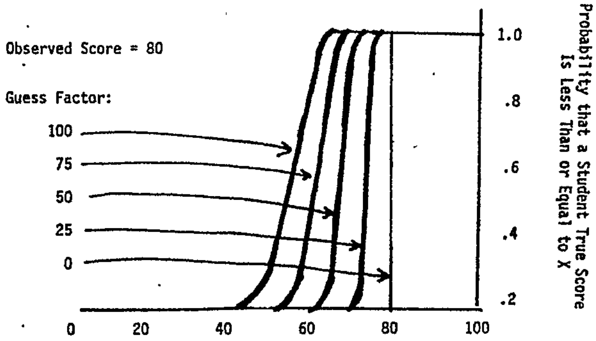
Cumulative Distribution of Actual Knowledge Scores for Different Guess Factors with a Student Observed Score of 50



Realize also, that a student's actual knowledge score is always less than or equal to his observed score and that as the guessing factor increases, students with a given observed score are more likely to have lower actual knowledge scores. This suggests that to the extent that guessing exists in standardized testing situations in school districts, students might really be achieving at a much lower level than we had thought. For example, with guessing a student scoring 80% correct certainly has not mastered 80% of the test items (see Figure 8). (Test manufacturers would agree with this statement, but provide no clue other than the SEM as to the uncertainty in the observed score; in this study the probability distribution of actual knowledge is defined.)

FIGURE 8

Cumulative Distribution of Actual Knowledge Scores with Various Guess Factors for a Student-Observed Score of 80

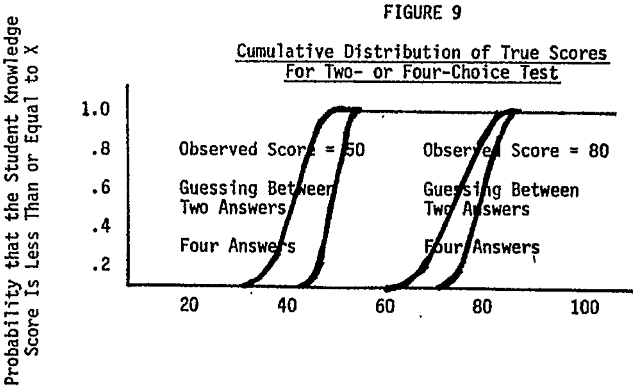


Number of Responses and Its Effect upon Test Precision

Variations which might arise from varying the guess pattern were also investigated. In this case a four-choice, 100-item exam with 50% guessing extent was examined for situations in which the guessing pattern only involved guessing between two of the responses (student has partial information), and a pattern involving guessing between all four of the responses for those items where guessing occurred. Figure 9 shows the cumulative distribution of actual knowledge scores for observed score of 50 and 80 when the student is in a guessing situation involving two choices and one involving four choices.

FIGURE 9

Cumulative Distribution of True Scores For Two- or Four-Choice Test



A situation very characteristic of inner city schools, according to many teachers, is that a student merely answers the first few items, and does precisely what is described above--namely, chooses at random one of the four responses for the rest of the examination. For a student observed score of 50 on this particular test, 39 would represent the 90% floor for the guessing between four choices and 29 would be the 90% floor if the student only guessed between two choices.* Again this finding in terms of individual test reliability has serious implications for programs aimed at educational programs in inner city schools.

In summary, the reliability of individual test scores for purposes of evaluation or of selection into special education programs is seriously eroded by guessing, (a phenomenon which, according to inner city teachers, is prevalent).

IV. SUMMARY OF CALIBRATION ANALYSES

This research was principally directed at several issues in standardized testing which are of particular importance to those concerned with the evaluation of instructional programs in the inner city. The following describes the findings for these issues.

Extent of Guessing Versus Test Precision

Test manufacturers assume that no guessing occurs or that situation, as defined in Table 3, is operative. By examining the changes in the precision for different extents of guessing, from zero guessing to total guessing or guessing factors from 0 to 100, one can roughly gauge how sensitive test precision is to this assumption. This study found for a four-choice response and fixed guessing pattern as defined in Table 4 that the extent of guessing is critical for test precision and should be investigated in further depth. Test manufacturers should direct particular attention to the extent of guessing which exists in the inner city schools. This is important for two reasons. First, one would typically find more guessing on standardized tests in these schools, and second, typically large-scale educational intervention programs are operative in these schools and subsequent evaluations are based upon the results of standardized tests.

Test Calibration

Particular attention was given in this study to developing an analytical procedure based upon a computer simulation model involving Bayesian analysis and monte carlo techniques for assessing test precision given a fixed guessing pattern and guessing extent. Tests such as CTBS and the Cooperative Primary could be calibrated and tests could be selected for schools which yield the highest precision given the achievement range of the students.

True Mastery Versus Apparent Mastery of Test Content

Because a student observed score is composed of those items to which he knows the answer and those items on which he made a lucky guess, the observed score on a test with any amount of guessing is an inflated measure of the amount of knowledge possessed by the student. The greater the degree of guessing, the greater the amount of bias becomes. This effect raises the possibility that most students are really achieving at much lower levels than we had been led to believe by taking their test scores at face value. Students may have many more gaps in their knowledge than previously expected and thus may be expected to have much greater difficulty in building upon this incomplete knowledge base.

V. SUMMARY

> The usual tests and the language habits of our culture tend to promote confusion between certainty and belief. They encourage both the vice of acting and speaking as though we were certain when we are only fairly sure and that of acting and speaking as though the opinions we do have were worthless when they are not very strong. (17)

The above quote by the famous American decision theorist Leonard J. Savage directs attention to one of the most critical problem areas in education--namely, the use of multiple response standardized examinations which require a student to force-choice his answer to a particular question. Unfortunately, these test scores are then used to make decisions regarding individual student selection to certain programs and recruitment and, more important, to assess individual knowledge at various stages of the educational process.

BIBLIOGRAPHY

1. Bruno, James E. Monte carlo techniques. In Hensley & Yates, Techniques of educational forecasting (Chapter 7). McCutchan & Co., 1974.
2. Bruno, James E. (Ed.). Emerging issues in education. Lexington, Mass.: D. C. Heath, 1972.
3. Bruno, James E. Use of monte carlo techniques for determining optimal size of substitute teacher pool in larger urban districts. Journal of Socio Economic Planning Sciences, 1970, 4, 415-428.
4. Bruno, James E. An analysis of the incentive increment program at Beverly Hills Unified School District. Technical report, 1967.
5. Danemann, E. PERT analysis using monte carlo techniques. Santa Monica, Calif.: RAND Corporation, January 1966. (RM 4881)
6. de Finetti, Bruno. La prevision: Ses lois logiques, ses sources subjectives. Annales de

*The 90% floor means that 90% of the distribution of actual knowledge scores would be above this score given the observed score.

l'Institut Henri Poincare, 1964, 7. (Translated and reprinted as Foresight: Its logical laws, its subjective sources. In Henry E. Kyburg, Jr., & Howard E. Smokler (Eds.), Studies in subjective probabilities. New York: Wiley, 1973.

7. de Finetti, Bruno. Methods for discriminating levels of partial knowledge concerning a test item. The British Journal of Mathematical and Statistical Psychology, 1965, 18, 87-123.

8. Dei Rossi, Gerald S. Exploring the effects of distorting classical linear regression assumptions. RAND Corporation P4103, August 1964.

9. Doscher, Lynn. Monte carlo analysis of guessing on standardized test reliability. Doctoral dissertation, 1978.

10. Gold, B. K. Quantitative methods for administrative decision making in junior colleges. Ed.D. dissertation, UCLA.

11. Griffin, Mary, & Schmitt, John. A monte carlo model for the prediction of public school enrollments. Paper read at AERA meeting, 1966.

12. Massengill, H. Edward, & Shuford, E. H., Jr. Decision-theoretic psychometrics: A logical analysis of guessing. Lexington, Mass.: The Shuford-Massengill Corp., 1966.

13. Massengill, H. Edward. What pupils and teachers should know about guessing. Lexington, Mass.: The Shuford-Massengill Corp., 1967.

14. Massengill, H. Edward, & Shuford, E. H., Jr. A report on the effect of degree of confidence in student testing. Lexington, Mass.: The Shuford-Massengill Corp., 1968.

15. Rosenthal, Robert, & Jacobson, Lenore. Pygmalion in the classroom, teacher expectation and pupils' intellectual development. New York: Holt, Rinehart, & Winston, 1968.

16. Rutherford, Brent M. The accuracy, robustness and relationships among correlational models for social analysis: A monte carlo simulation. CASEA, Eugene, Oregon, September 1972.

17. Savage, L. J. The foundations of statistics. New York: Wiley, 1954.

18. Shuford, E. H., Albert, A., & Massengill, H. E. Admissible probability measurement procedures. Psychometrika, 1966, 31, 125-145.

19. Shuford, E. H. The effects of guessing upon the Iowa Test of Basic Skills. Lexington, Mass.: The Shuford-Massengill Corp., 1971. See also: 1967--Individual and social justice in objective testing; The relative effectiveness of five instructional strategies; How to shorten a test and increase its reliability and validity; What pupils and teachers should know about guessing; 1966--The worth of individualizing instruction; The effect of guessing on the quality of personnel and counseling decisions; A logical analysis of guessing.

20. Shuford, E. H. Confidence testing: A new tool for measurement. Proceedings of the 11th Annual Conference of the Military Testing Association, September 15-19, 1974.

21. Thorndike, Robert Ladd (Ed.). Educational measurement. Washington, D.C.: American Council on Education, 1971.

22. Thorndike, Robert L., & Hagan, Elizabeth. Measurement and evaluation in psychology and education (3rd ed.). New York: Wiley, 1969.

23. Toda, M. Measurement of subjective probability distribution ESD-TDR, 63-407. Bedford, Mass.: Decision Sciences Laboratory, L. G. Hanson Field, 1963.

24. Zimmerman, Donald W., & Williams, Richard H. Interpretation of the standard error of measurement when true scores and error scores on mental tests are not independent. Psychological Reports, 1966, 19, 611-617.