THE CONSTRUCTION AND VALIDATION OF A SIMPLIFIED MEASUREMENT ANALYSIS MODEL FOR ITEM SELECTION AND INTERPRETATION OF SCORES

A Dissertation

Presented to

The Graduate Faculty

of the College of Education

University of Houston

In Partial Fulfillment

of the Requirements for the Degree

Doctor of Education

---

by

Raymond E. Hassett

December 1970

ACKNOWLEDGMENTS

I wish to gratefully acknowledge the chairmanship of Dr. Robert McClintock, whose broad knowledge and experience in measurement and statistics were of inestimable assistance in completing this study. The guidance of Drs. Joseph Carbonari, Richard Evans, Marvin Sterrett, and James Tinsley was also much appreciated.

For the kindness and energetic assistance of Dr. Irvin Miller and his classes in completing numerous pilot studies, I will be eternally thankful. The constructive comments of Dr. William Nesbitt were also of valuable assistance in firming the foundation of this study.

To Pam Duncan and her perseverance and skill in typing the final copy of the dissertation, I give my sincere thanks for a job well done.

To my wife Lois and two children, Lloyd and Mary Alice, I give my heartfelt thanks for their patience and understanding in standing by me in this long, but rewarding task.

Raymond Hassett

Houston, Texas
December, 1970

THE CONSTRUCTION AND VALIDATION OF A SIMPLIFIED MEASUREMENT ANALYSIS MODEL FOR ITEM SELECTION AND INTERPRETATION OF SCORES

An Abstract of a Dissertation
Presented to
The Graduate Faculty
of the College of Education
University of Houston

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Education

---

by
Raymond E. Hassett
December 1970

Hassett, Raymond E. "The Construction and Validation of a Simplified Measurement Analysis Model for Item Selection and Interpretation of Scores." Unpublished Doctor's dissertation, University of Houston, 1970.

ABSTRACT

The purpose of the study was to construct and validate a simplified Measurement Analysis Model that could be used by a teacher in analyzing teacher-made tests in the classroom. The model was developed in an attempt to give teachers the item analysis tools of the measurement specialist without requiring a high degree of computational or statistical skill.

A systematic review of the literature was made in three areas: (1) main constructs in measurement theory, (2) major components used in standardized item analysis procedures, and (3) procedures currently available for estimating major item analysis components.

Systematic sampling at specific percentile values, and the centroid of the area sampled from the extremes in terms of the normal distribution, formed the foundation for developing the Measurement Analysis Model. Eighteen sampling patterns were developed and analyzed. These sampling patterns involved sampling from the extremes of distributions which had been rank ordered, stratified sampling at specific points in a rank ordered

distribution, and combinations of the above. The sampling patterns were used with skewed populations and class sizes ranging from 20 to 60 with absolute deviations calculated between sample estimates and actual values of major item analysis components obtained by a standardized item analysis procedure.

The final Measurement Analysis Model used a systematic sampling pattern involving a constant sample size of ten for estimating the Test Mean, Test Standard Deviation, Test Variance, Test Reliability Coefficient and Standard Scores. It also used a constant sample size of ten from each extreme of a rank ordered distribution in estimating Item Discrimination Index, Item Probability Index, and Item Difficulty Index. Tables were developed to assist a teacher in estimating Test Standard Deviation, Test Variance, Test Reliability Coefficient, Item Discrimination Index, and Item Probability Index.

Fifty-five experienced teachers and forty-one inexperienced students were randomly selected to test the objectivity and reliability of the Measurement Analysis Model. An analysis of variance and a coefficient of equivalence were used to test the objectivity of the model and a split-half reliability procedure was used to test the reliability. Twenty-three teacher-made tests were analyzed and actual values obtained by using a standardized

item analysis procedure were compared with estimates obtained by using the Measurement Analysis Model as a test for validity and feasibility.

The Measurement Analysis Model was found to be objective, reliable, valid and feasible for use in the classroom. The study pinpointed areas where further refinement would increase the efficiency of the model. The major item analysis components obtained by using the model were very close approximations of actual values obtained by using a standardized item analysis procedure.

Several recommendations were made as direct results of this study: (1) in-service instruction of teachers in valid and reliable measurement and evaluation procedures, (2) further research in sampling from the extremes of rank ordered populations, (3) further research in using mean deviation scores and correction terms generated by calculating the predicted value of the true centroid intersection point on the z-axis of the normal distribution in estimating Test Variance, (4) further research in the generation of tables of point biserial values using the centroid of upper and lower portions of the normal distribution, and (5) further research in developing even more simplified measurement analysis models for use by teachers in the classroom.

TABLE OF CONTENTS

LIST OF TABLES

Table Page

Table Page

Table Page

LIST OF FIGURES

LIST OF GRAPHS

Graph Page

Chapter 1

PURPOSE

> If he (the teacher) does an inadequate job of evaluation, he not only has inferior data to communicate to authorities and to his students, but he may develop misconceptions about the effect that his teaching is having on students.
>
> Henry Clay Lindgren, 1967[1]

A current trend in the public school classroom is toward individualized instruction, focused on the needs of the individual and his potential. This predicates a capability to measure and evaluate the learning progress of an individual student. In a recent survey, George Gallup polled a nationwide sample of 1,592 adults and 299 students on the issue of teacher accountability. Sixty-seven percent of those polled favored ". . . a system that would hold teachers and administrators more accountable. . . ."[2]

The feeling that the public is ready to lose patience with current measurement procedures is summed up by Dyer in the following statement:

---

[1]Henry Clay Lindgren, Educational Psychology in the Classroom (New York: John Wiley and Sons, Inc., 1967), p. 427.

[2]George Gallup, "Second Annual Survey of the Public's Attitude Toward the Public Schools," Phi Delta Kappan, LII (October, 1970), 101.

> A comprehensive program to develop better methods for assessing achievement in all the classrooms of the country would go a long way toward taking the heat out of the controversies now plaguing us and would furnish the means for replacing the anarchy in education with new vitality and a sense of direction.[3]

RATIONALE

The following premises served as the foundation for this study:

1. A teacher is a rational decision maker.

2. The role of a rational decision maker requires diagnostic and prescriptive capabilities.

3. The measurement process requires that a teacher be able to sample typical behavior in a classroom and record this behavior in a meaningful format.

4. There are three basic continua in the measurement process--response, stimulus, and judgment.

5. A decision point involves the ability to discriminate and this implies the ability to measure just noticeable differences.

---

[3]Henry S. Dyer, "On the Assessment of Academic Achievement," in Assessing Behavior: Readings in Educational and Psychological Measurement, eds. John T. Flynn and Herbert Garber (Reading, Massachusetts: Addison-Wesley Publishing Company, 1967), p. 20.

NEED

Teachers are faced with continual pressure to produce meaningful evidence that learning has actually taken place. The pressure exists, not only in developing meaningful progress reports, but also in determining the validity and reliability of classroom procedures used to diagnose learning progress. Teachers do not currently possess adequate means by which to accomplish these tasks; consequently, many evaluations made from data obtained in a classroom are suspect. Teachers need simplified procedures which will allow them to closely approximate the item analysis components used by measurement specialists in building valid and reliable measurement instruments. A teacher's prerogative to use any type of assessment procedure in a classroom should be held inviolate; however, the validity and reliability of those procedures must be demonstrated.

SCOPE OF STUDY

It was proposed that a simplified Measurement Analysis Model be developed which would be useable by a teacher without a high degree of computational or statistical skill. The model would be designed for use with measurement procedures involving the use of decision points and would be validated by applying the basic procedures to a wide array of teacher-made measurement instruments.

The study was designed to answer the following questions:

1. Objectivity. Would teachers using the Measurement Analysis Model arrive at the same relative decisions as a standardized item analysis procedure when rating measurement items?

2. Reliability. Would teachers using the Measurement Analysis Model make consistent decisions when rating measurement items?

3. Validity. Would major item analysis components obtained by using the Measurement Analysis Model be comparable in validity to those obtained by using a standardized item analysis procedure?

4. Feasibility. Would the Measurement Analysis Model operate under actual classroom conditions and still result in valid and reliable item analysis components when compared to a standardized item analysis procedure?

DEFINITION OF TERMS

decision point--the point at which a decision must be made as to successful or nonsuccessful achievement of a standard.

evaluation--the judgment made after an analysis of information available on a subject or trait.

feedback--process whereby data are fed back into the system to modify and correct an ongoing program.

instrument--any item or group of items which have been developed to assess achievement.

item analysis--an evaluation of test results to determine which items are effective in measuring differences among subjects.

Item Difficulty Index--the ability of an item to measure differences between those who score high on the criterion and those who score low, expressed as a proportion.

Item Probability Index--the probability that an item will measure differences between those who have and those who do not have the trait being measured.

measurement--the process of assigning numerical values to traits observed or measured in terms of specific rules.

normal distribution--a theoretical distribution which can be inscribed by a symmetrical curve with a kurtosis of 3 and a skewness of 0. The curve possesses points of inflection at ± 1 standard deviation unit.

reliability--consistency or stability of a test. A test is considered to be reliable if it consistently produces similar results under similar circumstances.

standard score--a score which is expressed in standard deviation units from the mean.

Test Reliability Coefficient--a measure of the homogeneity of a test as reflected in its consistency or

stability. It rests on the basic assumption that the test is measuring one common trait or factor.

validity--the extent to which a test measures what it is supposed to measure.

OUTLINE OF CHAPTERS

Chapter 1: Purpose. Introduction to the study demonstrating the underlying premises, need for study, scope of study, definition of terms, and outline of succeeding chapters in the study.

Chapter 2: Review of the Literature. Review of current research and opinions bearing on the study; guidelines used in conducting research--evaluation model, main constructs in measurement theory, major item analysis components, and current procedures available for estimating major item analysis components; and an interpretation of these studies.

Chapter 3: Development of Model. General considerations and statistical foundations used in developing model, selection of sampling patterns, sampling procedures, and model format.

Chapter 4: Field Investigative Procedures and Findings. Procedures used in selecting subjects and collecting data and an analysis of data involving objectivity, reliability, validity, and feasibility of Measurement Analysis Model.

Chapter 5: Summary, Conclusions, and Recommendations. Summary of data collected and analyzed, conclusions reached, and recommendations as to further research needed in this area.

Chapter 2

REVIEW OF THE LITERATURE

There is a polyglot of written material on measurement and evaluation procedures in statistical, psychological, educational, and other sources. It was, therefore, determined that the best procedure to follow was a systematic sampling of existing literature bearing directly on the development of a Measurement Analysis Model. The research was intended to be comprehensive, but not exhaustive. Guidelines used in conducting the research were: the frame of reference to be used for the study, the main constructs of measurement theory to be used as the foundation for the Measurement Analysis Model, the major item analysis components to be considered for inclusion in the Measurement Analysis Model, and the procedures which were currently available for estimating major item analysis components.

FRAME OF REFERENCE

Carl Smith and Roger Farr, Indiana University, made a comprehensive study of evaluators and evaluation designs in Title I and Title III programs under a United States

Office of Education research grant.[1] They selected a continuous evaluation model developed by Guba and Stufflebeam as an example in the development of a Simulation Training Package for Evaluation.[2]

The CIPP model (context, input, process, and product) [see Figure 2, Appendix A], is not new;[3] however, it does present a very precise picture of the need to evaluate all aspects impinging on the variables being studied in the classroom. It gives a conceptual framework from which to view the instructional process in its entirety. A study conducted by T. B. Greenfield supported the need to study more than just the classroom in evaluating the instructional program.[4] He found, in his study of an educational system in Canada, that there were several sources of variation in pupil achievement. These findings are given in Table 1.[5]

---

[1]Carl B. Smith and Roger Farr, Evaluation Training Package (Bloomington: Indiana University, 1969), pp. 1-2.

[2]Ibid., pp. 4 and 10-13.

[3]Daniel L. Stufflebeam, "The Use and Abuse of Evaluation in Title III," Theory Into Practice, VI (June, 1967), 126-133.

[4]T. B. Greenfield, "Administration and Systems Analysis," The Canadian Administrator, III (April, 1964), 25-30.

[5]Ibid., p. 29.

Table 1

Sources of Variation in Classroom Achievement

Source of Variation in Achievement	Percentage of Total Variation in Achievement
Classes	19.28%
Schools	2.80
Districts	10.28
Pupils	67.64
Total	100.00

Diagnostic Measurement

Diagnostic measurement predicates an awareness of the child's needs and activities that will help him attain those needs. Mitzel, in reporting on the instruction revolution, referred to a program developed at the University of Pittsburg entitled, "individual prescribed instruction" or IPI.[6] He further stated that "achievement tests keyed to course objectives would have to be constructed and used as both diagnostic placement and end-of-course determiners."[7] McKay, research director, Maryland State Department of Education, stated that "we are looking forward to the time when a teacher will receive a printout - a profile of every

[6]Harold E. Mitzel, "The Impending Instruction Revolution," Phi Delta Kappan, LI (April, 1970), 435.

[7]Ibid.

student in her class - that she can use to fit an education program to each student's needs."[8] De Lay and Nyberg reported success with students when they were given non-evaluative feedback or ". . . information about where the student is now in relation to where he wants to be."[9]

Dorothy Wood reported that "a test to be used in this way (diagnosis) must be so arranged as to provide separate scores on the specific areas in which diagnostic interest centers."[10] In a report by the American Association of School Administrators, it was stated that "a second use of testing is to diagnose learning difficulties of an individual student or an entire class to provide information helpful in planning subsequent teaching."[11] Ralph Taylor, in writing for the Association for Supervision and Curriculum Development, stated that "one purpose of assessment of the

---

[8]Richard M. McKay, "State-wide Information System," Phi Delta Kappan, LI (November, 1969), 178.

[9]Donald H. De Lay and David Nyberg, "If Your School Stinks, CRAM It," Phi Delta Kappan, LI (February, 1970), 312.

[10]Dorothy A. Wood, Test Construction - Development and Interpretation of Achievement Tests (Colombus, Ohio: Charles E. Merrill Books, Inc., 1961), p. 6.

[11]American Association of School Administrators, National Educational Assessment: Pro and Con (Washington, D. C.: National Education Association, 1966), p. 11.

individual is to determine his readiness to pursue the next step of learning."[12] Diagnosis is also stated by Helmstadter as one of the prime purposes of measurement.[13]

Generalization

The primary need of the teacher in the classroom is an awareness of where the child is in terms of his learning needs. Research indicated two major facets which should be considered in the development of a Measurement Analysis Model: the first involved an awareness of the different forces operating in the instructional environment and the second involved the ability to diagnose individual learning strengths and weaknesses. The Measurement Analysis Model must enable teachers to select items which discriminate between those who have and those who do not have a given trait and the model must assist in analyzing a wide range of measurement procedures used in a wide range of circumstances.

Another important facet of the Measurement Analysis Model must be the assignment of values to individual items

---

[12]Ralph W. Tyler, "The Purposes of Assessment," Improving Educational Assessment & An Inventory of Measures of Affective Behavior, ed. Walcott Beatty (Washington, D.C.: Association for Supervision and Curriculum Development, 1969), p. 3.

[13]G. C. Helmstadter, Principles of Psychological Measurement (New York: Appleton-Century-Crofts, 1964), pp. 6-9.

which will allow these items to be placed in an item pool for later use in assessing individual learning progress. The resultant scores must result in a valid and reliable profile of each student and/or the instructional climate. This would assist a teacher and a student in the planning and prescribing of subsequent learning activities.

MAIN CONSTRUCTS IN MEASUREMENT THEORY

Normal Distribution

The attributes of the normal distribution are well documented in the literature and it is not intended that a full discourse be given here. A discussion of the development of the normal distribution is given by Walker[14] and its attributes by Ferguson.[15] It has been theoretically and empirically demonstrated that the frequency distribution of many physical, biological, and psychological measurements are approximations of the normal form.[16]

---

[14]H. M. Walker, "Studies in the History of Statistical Method," in Elementary Statistical Methods (New York: Henry Holt and Company, Inc., 1943), pp. 166-200.

[15]George A. Ferguson, Statistical Analysis in Psychology and Education (New York: McGraw-Hill Book Company, 1966), pp. 95-104.

[16]Ibid., p. 95.

Psychophysical Theory

Guilford discussed two basic continua in psychophysical investigations: a response continuum which is measurable in physical units and a stimulus continuum which is measurable in psychological units.[17] He portrayed these two continua as:

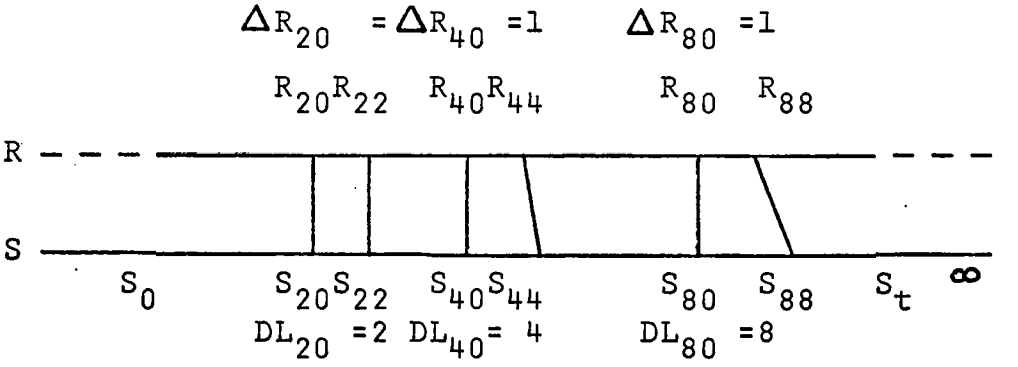


Figure 1

The Absolute Threshold and Terminal Stimulus Shown on Response Continuum R and Stimulus Continuum S

> . . . Three different limens are illustrated, showing that three increments on R, called ΔR, are equal but correspond to proportional increments on S, in conformity to Weber's Law. Some liberties have been taken with scale consistency to illustrate all these properties.[18]

The response continuum is denoted by a shorter line as there are some stimuli too weak to elicit a response and some too strong to be handled.[19] A stimulus limen may be

---

[17] J. P. Guilford, Psychometric Methods (New York: McGraw-Hill Book Company, Inc., 1954), p. 21.

[18] Ibid.

[19] Ibid.

computed statistically and ". . . defined as that low stimulus quantity that arouses a response 50 percent of the time."[20] Guilford further stated that "a true response, like a true score on a mental test, is the response this individual should give to a certain stimulus if there were no disturbing forces at the moment. Operationally, it can be defined as the central tendency of all responses the individual would give to the stimulus on a very large number of occasions."[21]

It is often impossible to observe the response continuum as an externally observable aspect of behavior and Guilford handled this problem by developing a third or judgment continuum.[23] The problem then became, ". . . what stimulus value would give judgments J = 1 with a probability of .5."[24]

Rasch also proposed a probabilistic model which involved a forced (two) choice situation and stated that ". . . it has become clear that this model is in fact the complete answer to the requirement that statements about the parameters and adequacy of a discrete probabilistic model

---

[20]Ibid., p. 22.

[21]Ibid., p. 28.

[22]Ibid., p. 29.

[23]Ibid., p. 31.

[24]Ibid.

be objective in a sense to be fully specified."[25] The model included the statement that ". . . when any person encounters any item . . . the outcome of such an encounter is governed by the product of the person and the easiness of the item and nothing more!"[26]

Forced Choice Rating Systems

The Encyclopedia of Educational Research, in a review of studies conducted on forced choice, stated that "two-choice questions have been shown to be highly efficient."[27] Zavala published a review of the 18-year history of the forced choice rating scale technique and stated that the FC (forced choice) has also been used in problems related to signal detectability and auditory thresholds."[28] He further stated that the "FC scale provides a test relatively

---

[25]Georg Rasch, "An Individualistic Approach to Item Analysis," in Readings in Mathematical Social Science, eds. Lazarsfeld and Henry (Chicago: Science Research Associates, Inc., 1966), p. 89.

[26]Benjamin Wright and Nargis Panchapakesan, "A Procedure for Sample-Free Item Analysis" (paper developed at University of Chicago, January, 1968, Chicago, Illinois), p. I-1. (Mimeographed.)

[27]William E. Coffman, "Achievement Tests," Encyclopedia of Educational Research, ed. Robert L. Ebel (4th ed.; New York: Macmillan Co., 1969), p. 11.

[28]Albert Zavala, "Development of the Forced-Choice Rating Technique," Psychological Bulletin LXIII (1965), p. 117.

free of the disadvantages of many of the traditional tests."[29] After reviewing studies on the forced choice technique, Helmstadter concluded that "in general, the forced choice technique does seem to tend to overcome some of the most serious rating errors."[30] Guilford concluded that "when the items are of specific actions observed by the rater, the check list becomes essentially an achievement or proficiency test and its score has the status that would be accorded to that type of measurement."[31] Luce also presented a comprehensive discussion on the rationale for using a probabilistic model in measurement with a forced choice format.[32]

Generalization

The research given above seemed to indicate the presence of three primary item analysis components in measurement theory: an Item Difficulty Index (stimulus), an Item Discrimination Index (discriminal dispersion), and an Item Probability Index (probability of a specific stimulus with a specific discriminal dispersion eliciting a desired response). Treatises on these indexes are readily available

---

[29]Ibid., p. 122. [30]Helmstadter, op. cit., p. 188.

[31]Guilford, op. cit., pp. 273-274.

[32]R. Duncan Luce, Individual Choice Behavior - A Theoretical Analysis (New York: John Wiley and Sons, Inc., 1959), pp. 1-92.

in most textbooks on psychometrics.[33] There was also considerable evidence to support the desirability of using a forced choice format in a probabilistic theory of measurement.

The forced choice technique should enable us to dichotomize all decision points and assign a value of "1" for successful completion and a value of "0" for unsuccessful completion by the subject encountering a decision point. All means of observation and measurement can be subsumed under the forced choice format--the prime consideration would be the ability of the rater to discriminate between those who did and those who did not demonstrate the quality or trait being measured. This ability to discriminate could be verified through an item analysis procedure and weaknesses of the rater could be pinpointed.

MAJOR ITEM ANALYSIS COMPONENTS

Test Mean

Ferguson stated that "the sum of squares of deviation about the arithmetic mean is less than the sum of squares of deviation about any other value."[34] He further stated that it is ". . . regarded as an appropriate measure of central

---

[33]Guilford, op. cit., pp. 414-469; see also Helmstadter, op. cit., pp. 157-178.

[34]Ferguson, op. cit., p. 52.

location for interval and ratio variables."[35] He also noted that "if the frequency distribution is represented graphically, the mean is a point on the horizontal axis which corresponds to the centroid, or center of gravity, of the distribution."[36] The mean is given as

$$\bar{X} = \frac{\sum_{i=1}^{N} X_i}{N}, \qquad \text{where}$$

$\bar{X}$ = arithmetic mean,

X_i = raw score, and

N = number of raw scores.

Test Standard Deviation and Test Variance

A measure of central tendency is not sufficient in describing the distribution of a population parameter. Some measure of dispersion, spread, or variability is needed.[37] A working definition of the variance is ". . . the sum of squares of the deviations of the observations from $\bar{X}$ divided by . . . the total number of observations."[38] In terms of the variance, "the standard deviation is defined as the

---

[35]Ibid., p. 57. [36]Ibid.

[37]Wilfred J. Dixon and Frank J. Massey, Jr., Introduction to Statistical Analysis (New York: McGraw-Hill Book Company, Inc., 1951), p. 19.

[38]Ibid., p. 20.

positive square root of the variance."[39] The standard deviation is given as

$$s = \sqrt{\frac{\Sigma(X - \bar{X})^2}{N}}, \qquad \text{where}$$

s = standard deviation.

Test Reliability Coefficient

Downie stated that "if our measurements are reliable, we can make our statements with confidence."[40] Flanagan stated that "it appears quite obvious that some published tests would have been improved had half of the items not been included."[41] Hoyt, in his research with the analysis of variance technique in estimating test reliability, found that the analysis of variance gave precisely the same result as the Kuder-Richardson Formula 20.[42] Lord stated that:

> The use of r_{20} (Kuder-Richardson Formula 20) is appropriate whenever one is willing to ignore any difference between the mean test score

---

[39] Ibid.

[40] N. M. Downie, Fundamentals of Measurement: Techniques and Practices (New York: Oxford University Press, 1967), p. 82.

[41] John C. Flanagan, "General Considerations in the Selection of Test Items and A Short Method of Estimating the Product-Moment Coefficient from Data at the Tails of the Distribution," Journal of Educational Psychology XXX (December, 1939), p. 674.

[42] Cyril Hoyt, "Test Reliability Estimated by Analysis of Variance," Psychometrika VI (June, 1941), p. 156.

> of the group and their mean true score, i.e., when one is concerned only with the relative rather than the absolute size of the scores of the group. On the other hand, $\underline{r}_{21}$ (Kuder-Richardson Formula 21) should be used whenever one is concerned with the actual magnitude of the errors of measurement, e.g., whenever there is a predetermined cutting score which divides the examinees into passing and failing groups.[43]

Gulliksen derived a reliability formula which was identical to the Kuder-Richardson Formula 21 and it ". . . has the advantage of being very simple to calculate, since it uses only the mean, variance, and number of items. Also, it has the advantage of being a lower bound . . ."[44] This formula can be stated as

$$\underline{r}_{xx} = \frac{K}{K - 1}\left[1 - \frac{M_x - \frac{M_x^2}{X}}{S_x^2}\right],$$

$\underline{r}_{xx}$ is the reliability of the test,

K is the number of items in the test,

M_x is the test mean, and

S_x^2 is the variance of raw scores on the test.[45]

---

[43]Frederick M. Lord, "Sampling Fluctuations Resulting from the Sampling of Test Items," Psychometrika XX (March, 1955), p. 9.

[44]Harold Gulliksen, Theory of Mental Tests (New York: John Wiley and Sons, Inc., 1950), p. 225.

[45]Ibid.

Lyerly developed a table of significant values for the Kuder-Richardson Formula 21 and found that it was affected more by the number of subjects than by the number of items.[46]

Standard Scores

Good stated that ". . . test scores would be far easier to interpret if the school made its own conversion of the raw point scores into standard scores of some sort . . ."[47] Ebel went further in discussing the "low esteem" held by test specialists for raw scores and stated that ". . . normative standard scores are currently far more popular than content standard scores."[48] This preference for conversion of raw scores to standard scores was also stated by Payne,[49]

---

[46]Samuel B. Lyerly, "Significance Levels for the Kuder-Richardson (21) Reliability Coefficient," Educational and Psychological Measurement XIX (1959), pp. 74-75.

[47]Warren R. Good, "Misconceptions About Intelligence Testing," in Assessing Behavior: Readings in Educational and Psychological Measurement, eds. John T. Flynn and Herbert Garber (Reading, Massachusetts: Addison-Wesley Publishing Company, 1967), pp. 159-160.

[48]Robert L. Ebel, "Content Standard Test Scores," Educational and Psychological Measurement XXII (1962), pp. 16-17.

[49]David A. Payne, The Specification and Measurement of Learning Outcomes (Waltham, Massachusetts: Blaisdell Publishing Company, 1968), p. 108.

Edwards and Scannell,[50] Guilford,[51] and Helmstadter.[52] Ferguson described the standard score as ". . . a deviation from the mean divided by the standard deviation . . ."[53] He further stated that "transformation of a set of scores to the normal form is a relatively simple procedure."[54] The formula can be shown as

$$z' = \bar{X}' + SD' \left(\frac{X - \bar{X}}{SD} \right), \qquad \text{where}$$

z' = new standard score,

$\bar{X}'$ = new standard score mean,

SD' = new standard score standard deviation,

X = raw score,

$\bar{X}$ = test mean, and

SD = test standard deviation.

Item Discrimination Index

Ausubel stated that "an obvious attribute of an effective test is ability to distinguish maximally between individuals who vary with respect to the trait or competence being

---

[50]Allen J. Edwards and Dale P. Scannell, Educational Psychology - The Teaching-Learning Process (Scranton, Pennsylvania: International Textbook Company, 1968), p. 547.

[51]Guilford, op. cit., p. 83.

[52]Helmstadter, op. cit., p. 51.

[53]Ferguson, op. cit., p. 255.

[54]Ibid., p. 263.

measured."[55] Research by Englehart, in the comparison of several indexes of item discrimination, supported the conclusion ". . . that the easily computed index D (the difference between the proportions of persons passing an item in the upper and lower criterion groups) is remarkably effective in identifying poor items."[56] He also stated that ". . . D is more indicative of the actual number of discriminations made by an item than are correlation type indexes. From a time-and-motion point of view, D is probably the most economical index to calculate."[57] Perry and Michael stated that ". . . corresponding to values of the point biserial coefficient the fiducial interval at a specified probability level is narrower than it is for the ordinary Pearsonian coefficient."[58] Aleamoni and Spencer found that a ". . . choice of a biserial or point biserial item discrimination index would yield the

---

[55]David P. Ausubel, Educational Psychology - A Cognitive View (New York: Holt, Rinehart and Winston, Inc., 1968), p. 582.

[56]Samuel T. Mayo, "The Methodology and Technology of Educational and Psychological Testing," Review of Educational Research XXXVIII (February, 1968), p. 93.

[57]Ibid.

[58]Norman C. Perry and William B. Michael, "The Reliability of a Point Biserial Coefficient of Correlation," Psychometrika XIX (December, 1954), p. 314.

same rank ordering of items."[59] Davis concluded that "items with discrimination indices above 20 will ordinarily be found to have sufficient discriminating power for use in most achievement and aptitude tests."[60] "The formula for point biserial $\underline{r}$ is

$$\underline{r}_{pbi} = \frac{\bar{X}_p - \bar{X}_q}{s_t}\sqrt{pq} \text{ , where}$$

s_t = standard deviation of all scores on continuous variable, defined as $\sqrt{(X - \bar{X})^2/N}$,

p and q = proportions of individuals in two categories of dichotomous variable, and

$\bar{X}_p$ and $\bar{X}_q$ = mean scores on continuous variable of individuals within the two categories."[61]

Ferguson stated that "this statistic (point biserial) can always be interpreted as a measure of the degree to which the continuous variable differentiates, or discriminates, between the two categories of the dichotomous variable."[62] A

---

[59]Lawrence M. Aleamoni and Richard E. Spencer, "A Comparison of Biserial Discrimination, Point Biserial Discrimination, and Difficulty Indices in Item Analysis Data," Educational and Psychological Measurement XXIX (1969), p. 355.

[60]Frederick B. Davis, Item-Analysis Data: Their Computation, Interpretation, and Use in Test Construction (Cambridge, Massachusetts: Howard University, 1949), p. 15.

[61]Ferguson, op. cit., p. 240.

[62]Ibid., pp. 240-241.

limiting factor of the point biserial index is that "the maximum value of r_{pbi} never reaches +1; the minimum value never reaches -1."[63]

Item Probability Index

Thorndike discussed the relativity of psychological measurement and concluded that ". . . all predictions must have an element of tentativeness about them, be couched in terms of probabilities rather than in absolutes!"[64] Lord stated the ". . . assumption that the probability that an examinee will answer an item correctly is a normal-ogive function of his test-attribute score, implying that the attribute distribution is normal."[65] Ferguson stated that ". . . we define the limen as that point measured in 'σ-units' of ability where the probability of a person of that ability either passing or failing the item is one half."[66]

---

[63]Ibid., p. 241.

[64]Robert L. Thorndike, Educational Decisions and Human Assessment," in Assessing Behavior: Readings in Educational and Psychological Measurement, eds. John T. Flynn and Herbert Garber (Reading, Massachusetts: Addison-Wesley Publishing Company, 1967), p. 230.

[65]Frederick M. Lord, "Measurement Theory," Encyclopedia of Educational Research, ed. Robert L. Ebel (4th ed.; New York: MacMillan Co., 1969), p. 789.

[66]George A. Ferguson, "Item Selection by the Constant Process," Psychometrika VII (February, 1942), p. 20.

He further stated that "the estimation of such a probability is fundamental in the reduction of mental-test method to a sound theoretical basis."[67] One of the uses of the term probability is defined as ". . . the probability of an event as the ratio of the number of favorable cases to the total number of 'equally likely cases.'"[68] In another discussion of correlation coefficients, Ferguson stated, "thus $\underline{r}^2$ can quite meaningfully be interpreted as a proportion and $\underline{r}^2$ x 100 as a percent."[69] Therefore, a correlation or discrimination index of 0.50 represents a 25 percent association, etc. It is also possible to test the hypothesis that $\underline{r}_{pbi} = 0$ through the use of a $\underline{t}$-test. One formula for the $\underline{t}$-test is

$$\underline{t} = \underline{r}_{pbi}\sqrt{\frac{N - 2}{1 - \underline{r}_{pbi}^2}} \quad , \quad \underline{df} = N - 2.$$

Davis stated that "there is every reason to suppose that a normal distribution of talent does ordinarily underlie the bimodal point distribution of scores necessarily found for items scored simply 'right' or 'wrong,' especially if about 50% of the sample answers the item correctly."[70]

---

[67] Ibid.

[68] George A. Ferguson, Statistical Analysis in Psychology and Education (New York: McGraw-Hill Book Company, 1966), p. 83.

[69] Ibid., p. 127.

[70] Davis, op. cit., p. 10

Item Difficulty Index

Flanagan stated that to achieve ". . . maximum amount of discrimination between the individuals in a particular group, a test should be composed of items all of which are of fifty percent difficulty . . ."[71] Flanagan further stated that "it has frequently been pointed out that items which either all students or no students get correct are performing no measuring function in the test."[72] Ferguson found that "the general tendency is for the correlation between any two tests to decrease with increase in the difference in difficulty between them."[73] He further stated that "if the criterion of internal consistency, as I understand it, is to be reasonably approximated, the items in a test must be homogeneous with respect to difficulty, the difficulty of an item being described by the proportion of persons in a clearly 'defined population' who pass it."[74] A hypothesis proposed by Guilford ". . . to account for this result assumed that different abilities are brought into play when items are easy as compared

---

[71]Flanagan, op. cit., p. 675. [72]Ibid.

[73]George A. Ferguson, "The Factorial Interpretation of Test Difficulty," Psychometrika VI (October, 1941), p. 328.

[74]Ibid., p. 323.

with the situation when items are difficult."[75] His factor analysis of ten sub-tests of the Seashore Test of Pitch Discrimination resulted in three factors, each loading on a different segment of the difficulty continuum.[76] Gulliksen also reached the conclusion ". . . that the reliability of a test increases . . . as the dispersion of the item difficulties decreases . . ."[77]

Generalization

A teacher in a classroom must be able to select decision points which are valid and at the mean difficulty level of an individual student. This requires the teacher to make a determination on the probability that a specific item will separate those who have and those who do not have the trait being measured. After forming a group of items into a unifactor test with homogeneous difficulty levels, the teacher must be able to compute the reliability of the scores obtained and to state these scores in a meaningful frame of reference. The above requires the computation of three primary item analysis components (Item Discrimination Index,

---

[75]J. P. Guilford, "The Difficulty of a Test and Its Factor Composition," Psychometrika VI (April, 1941), p. 67.

[76]Ibid., pp. 67-77.

[77]Harold Gulliksen, "The Relation of Item Difficulty and Inter-Item Correlation to Test Variance and Reliability," Psychometrika XX (1945), p. 89.

Item Probability Index, and Item Difficulty Index) and five secondary item analysis components (Test Mean, Test Standard Deviation, Test Variance, Test Reliability Coefficient, and Standard Scores).

ESTIMATES OF MAJOR ITEM ANALYSIS COMPONENTS

Test Mean, Test Standard Deviation, and Test Variance

The work by Mosteller in the middle 40's concentrated on economically analyzing large masses of data by sampling at specific percentile values.[78] He found that very efficient approximations could be obtained from samples of ten cases or less taken from large populations.[79] His work was supported by Yost who obtained similar results.[80] Dixon and Massey also reported that the efficiency of these estimates remained high with small samples of ten and under.[81] The percentile values used in estimating the mean were: 5, 15, 25, 35, 45, 55, 65, 75, 85, and 95.[82] The percentile values used in

---

[78]Frederick Mosteller, "On Some Useful 'Inefficient' Statistics," Annals of Mathematical Statistics XVII (1946), p. 377.

[79]Ibid., p. 388.

[80]Earl K. Yost, Jr., "Joint Estimation of Mean and Standard Deviation by Percentiles," (unpublished Master's thesis, University of Oregon, 1948), p. 8.

[81]Dixon and Massey, op. cit., p. 230.

[82]Ibid.

estimating the Standard Deviation were: 98.5, 95, 90, 84, 75, 25, 16, 10, 5, and 1.5.[83] The percentile values used in estimating both the Test Mean and the Test Standard Deviation were: 3, 10, 20, 30, 50, 50, 70, 80, 90, and 97.[84] A problem discovered in using the same percentile values in estimating both the Test Mean and the Test Standard Deviation, was a resultant loss in efficiency as certain aspects of the two parameters were not compatible. A constant multiplier was used to obtain the final values for all estimates: Test Mean (0.1000), Test Standard Deviation (0.0739), and Test Mean and Test Standard Deviation (0.1104).[85]

A method of using the upper and lower sixth of the parent population was developed by W. L. Jenkins for estimating the Test Standard Deviation and is currently being used by the Educational Testing Service.[86] This can be written as

$$\text{Standard deviation} = \frac{\text{Sum of high sixth - sum of low sixth}}{\text{Half the number of students}}$$

---

[83]Ibid., p. 231. [84]Ibid., p. 232.

[85]Ibid., pp. 230-232.

[86]Paul B. Diederich, Short-cut Statistics for Teacher-made Tests (Princeton, New Jersey: Educational Testing Service, 1960), p. 23.

Item Discrimination Index and Item Difficulty Index

The works of Kelley, first published in 1928, are considered classics in the development of tables using the upper and lower 27 percent of the extremes.[87] Flanagan,[88] Davis,[89] Fan,[90] and others[91] have also developed similar tables. The 27 percent of the upper and lower extremes is considered optimum by many measurement specialists; however, recent research has found relatively little difference between results obtained using percentages of 10, 20, 25, 27, or 33.[92] Flanagan reported that sampling from the 27 percent at the extremes did not give proper weight to values nearer the extremes and he found that double and triple weights

---

[87]Truman L. Kelley, "The Selection of Upper and Lower Groups for the Validation of Test Items," Journal of Educational Psychology XXX (1939), p. 17.

[88]Flanagan, op. cit., p. 674.

[89]Davis, op. cit., pp. 13-14.

[90]C. T. Fan, "Note on Construction of an Item Analysis Table for the High-Low-27-Percent Group Method," Psychometrika XIX (September, 1954), p. 231.

[91]Robert M. Colver, "Estimating Item Indices by Nomographs," Psychometrika XXIV (June, 1959), p. 179; see also James J. Kirkpatrick and Edward E. Cureton, "Simplified Tables for Item Analysis," Educational and Psychological Measurement XIV (1954), p. 709.

[92]Harold Gulliksen, Theory of Mental Tests (New York: John Wiley and Sons, Inc., 1950), p. 373.

tended to correct this deficiency.[93] Diederich preferred to use the upper and lower 50 percent of the extremes because he found that ". . . for items in the middle range of difficulty (that 25 to 75 percent of the students answered correctly), the biserial correlation with total test is approximately equal to three times the high low difference, expressed as a percent of the class."[94] This can be stated as

H = the number of highs who got the item right,

L = the number of lows who got the item right,

H + L = "SUCCESS" (the total number who got the item right), and

H - L = "DISCRIMINATION" or "the high-low difference" (how many more highs than lows got the item right).[95]

Test Reliability Coefficient

Diederich was the only reference found for estimating reliability which would be suitable for use by teachers in a classroom. He prepared two tables for easy and hard tests which were entered by number of items, standard deviation, and test difficulty range.[96] The tables resulted in rough

---

[93]John C. Flanagan, "The Effectiveness of Short Methods for Calculating Correlation Coefficients," Psychological Bulletin IL (March, 1952), p. 344.

[94]Diederich, op. cit., p. 8. [95]Ibid., p. 9.

[96]Ibid., p. 31.

reliability estimates which did not possess the accuracy desired in the Measurement Analysis Model.

Standard Scores

This is a substitution problem and most measurement specialists seemed to take it for granted that teachers would be able to master the "simple" formula given for converting raw scores to standard scores.

Item Probability Index

No resource was found on the probability that an item would discriminate which would be suitable for easy reference by a classroom teacher; however, most introductory statistics texts carry tables with significant values for correlation coefficients.[97]

Generalization

Many valuable procedures for estimating item analysis components are present in a number of different technical and semi-technical sources; however, with the exception of Diederich's publication from Educational Testing Service,[98]

---

[97]James L. Bruning and B. L. Kintz, Computational Handbook of Statistics (Glenview, Illinois: Scott, Foresman and Company, 1968), p. 219; see also George A. Ferguson, Statistical Analysis in Psychology and Education (New York: McGraw-Hill Book Company, 1966), p. 406.

[98]Diederich, op. cit., pp. 1-37.

there seemed to have been few attempts to make these tools accessible to a classroom teacher. There also seemed to be a reluctance by researchers to publish background research on simplified techniques, probably prompting Gulliksen in 1950 to pass over short-cut methods of estimating item analysis components as he concluded that ". . . comparisons have not been made with computing cost and statistical precision."[99]

---

[99]Harold Gulliksen, Theory of Mental Tests (New York: Wiley and Sons, Inc., 1950), p. 385.

Chapter 3

DEVELOPMENT OF MEASUREMENT ANALYSIS MODEL

The primary frame of reference used in developing the Measurement Analysis Model was the principle of parsimony, whereby the point of maximum information with a minimum of problems in calculating major item analysis components was ascertained while maintaining an acceptable degree of validity and reliability. Calculations were kept minimal with unitary sampling and programming patterns used whenever possible. The tools of the measurement specialist were simplified and made available to a classroom teacher without the statistical and mathematical skills normally required.

STATISTICAL FOUNDATIONS

The normal distribution has many useful and unique characteristics which have been supported by empirical research. The exact delineation of the normal curve inscribing the normal distribution allows us to calculate the areas of given segments as they contain a known proportion of the total area. The use of systematic sampling at specific percentile values and the use of calculus in locating the centroid of the areas sampled in terms of the normal distribution, formed the foundation for developing the Measurement Analysis Model.

Systematic Sampling

Chance, or random error, as used in many random sampling techniques was found unsuitable for use in developing the Measurement Analysis Model. Systematic, or stratified random sampling, resulted in samples being more closely representative of each segment of the parent population. The error limits were found to be related to the deviation of the parent population from basic normality assumptions.

Centroid of Normal Distribution

The centroid of the area above any point z on the z-scale of the normal distribution is

$$\dot{z} = \frac{e^{-\frac{z^2}{2}}}{A\sqrt{2\pi}} \quad \text{or} \quad \frac{1}{e^{\frac{z^2}{2}} A\sqrt{2\pi}}, \qquad \text{where}$$

$\dot{z}$ = centroid of area above z,

e = natural logarithm base,

z = point on z-scale of normal distribution,

A = area above point on z-scale, and

π = constant (3.141593).

For the special case when A = 0.50 and z = 0.0, the equation reduced to

$$\dot{z}_{0.50} = \frac{1}{e^{\frac{z^2}{2}} A\sqrt{2\pi}} = \frac{1}{e^{0.0}\,(0.50)\sqrt{6.283186}}$$

$$= \frac{1}{1.2533314} = 0.797885; \qquad \text{however,}$$

the point of inflection for the normal curve occurs at σ, or ± 1 standard deviation unit. This made it possible to derive a correction term which could be used to derive the predicted true centroid intersection point on the z-axis. This was calculated as

$$C_{\dot{z}_{0.50}} = \frac{\frac{1}{\sigma}}{A\sqrt{2\pi}\,\sigma} = \frac{1}{0.797885} = 1.2533314 \quad \text{and}$$

when applied to the derived centroid value of 0.797885 for the area above 0.0, the predicted true centroid intersection point on the z-scale was found to be

$$\dot{z}_{0.50} = \dot{z}_i\left(\frac{\frac{1}{\sigma}}{A\sqrt{2\pi}\,\sigma}\right) = \dot{z}_i(1.2533314) \qquad \text{which,}$$

for the special case given above, resulted in a new predicted centroid intersection point on the z-scale for the upper half of the normal distribution. This was found to be

$$\dot{z}_{0.50} = 0.797885(1.2533314) = 1 \qquad \text{which}$$

is the true inflection point of the normal curve.

SELECTION OF SAMPLING PATTERNS

Preliminary research resulted in the selection of eighteen sampling patterns [see Figure 3, Appendix A], which were classified into three major groups.

Proportions and Numbers

Sampling from the extremes of distributions which have been rank ordered, has been reported by several measurement specialists. Six sampling distributions, using numbers and proportions, were used in selecting cases from each end of the upper and lower extremes of the parent population of thirty cases. The number of cases and the proportion drawn from each extreme were

A. 15 (0.5000),

B. 12 (0.4000),

C. 11 (0.3810),

D. 10 (0.3333),

E. 8 (0.2703), and

F. 5 (0.1667).

Systematic

This involved stratified sampling at specific points in a rank ordered distribution, starting from each extreme. The five variations used were

G. 5 (every third case, starting with second case) [see Figure 4, Appendix A],

H. 6 (every second case),

J. 7 (every second case),

K. 7 (every second case, starting with extreme case), and

L. 8 (every second case, starting with extreme case).

Composites

Seven combinations of proportions and systematic sampling patterns were formed for sampling from each end of the upper and lower extremes. These were

M. 5 (percentile values developed for estimating both the Test Mean and Test Standard Deviation) [see Figure 5, Appendix A] ,

N. 5 (percentile values developed for estimating the Test Standard Deviation) [see Figure 6, Appendix A],

P. 8 (3 cases from extreme and 5 from every second case),

Q. 10 (empirically developed through an iterative process) [see Figure 7, Appendix A],

R. 10 (5 cases at extreme and 5 cases from every second case),

S. 10 (7 cases at extreme and 3 cases from every second case), and

T. 11 (7 cases at extreme and 4 cases from every second case).

POPULATIONS WITH DIFFERENT SKEWNESSES

Sampling Procedures

The test results of 153 Foundations of Education students at the University of Houston on a multiple choice pre-test of fifty items were used as the sample population.

Ten systematic samples of 30 students each were drawn from the parent population to achieve different levels of skewness. Table 2 presents the sampling distributions used in this portion of the study.

The different sampling patterns were evaluated by calculating the absolute deviations of estimates of major item analysis components derived by using the Measurement Analysis Model from actual values derived by using a standardized item analysis procedure.

Test Mean

Several sampling patterns resulted in close approximations of the Test Mean [see Table 15, Appendix A.] Sampling pattern "G," presented in Graph 1, was selected because it had been empirically tested and involved a constant sampling of ten cases selected at specific percentile points. Sampling pattern "J," although the most accurate of those analyzed, involved different sampling sizes for different class sizes. Sampling pattern "Q," developed empirically during the course of this study through an iterative process, involved a constant sampling size of 10; however, it had not been exhaustively tested through different class sizes.

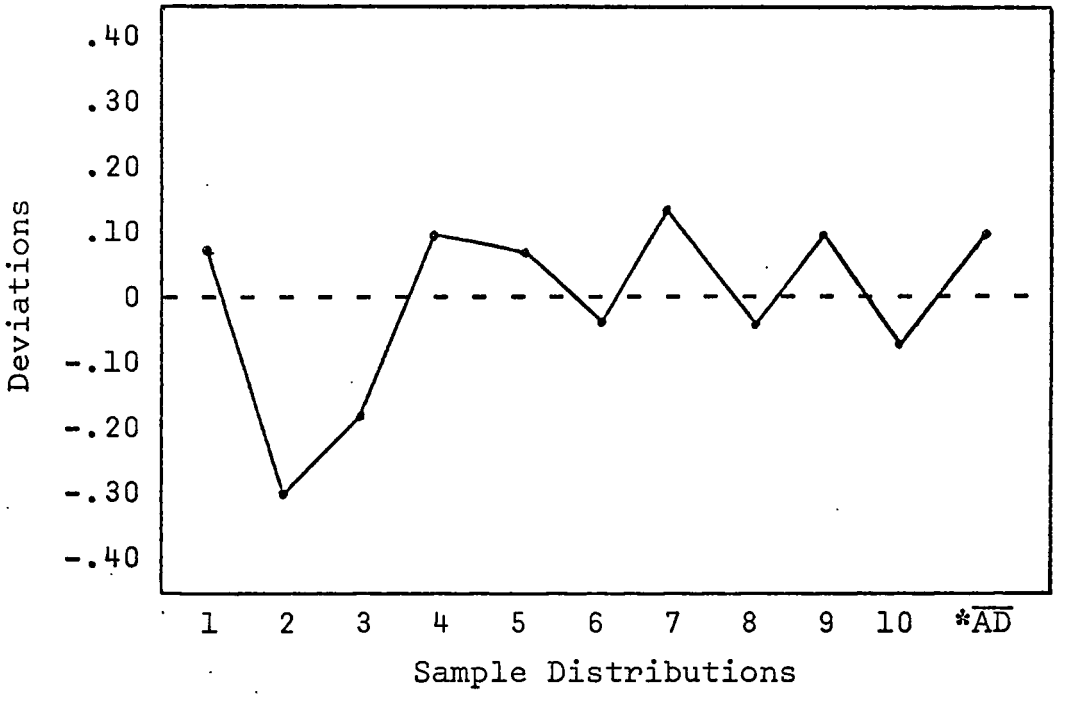
Table 2

Skewed Sampling Distributions

Distribution	Mean	S.D.	Variance	Skewness	Kurtosis	r_{xx}
1	29.33	3.42	11.68	1.20	4.00	-.04
2	32.10	3.08	9.47	1.19	3.43	-.22
3	30.97	4.11	16.93	.56	2.33	.31
4	26.10	5.62	31.61	.50	2.88	.62
5	20.53	6.00	36.05	.11	1.77	.68
6	28.83	5.99	35.87	.07	2.14	.67
7	23.87	6.50	42.19	-.08	2.42	.72
8	18.23	3.81	14.53	-.22	2.19	.21
9	19.70	3.78	14.29	-.60	2.56	.17
10	16.87	2.47	6.12	-1.09	3.71	-.84
Population	24.68	5.52	30.43	.02	3.01	.60

Graph 1

Deviation of Measurement Analysis Model Estimates of Test Mean from Actual Values Using Sampling Pattern "G" with Skewed Distributions



*mean absolute deviation.

Test Standard Deviation and Test Variance

Sampling patterns "A" through "F," using correction terms generated from the centroid of the area sampled [see Table 10, Appendix A], resulted in very close approximations; however, they involved using different sample sizes for different class sizes [see Tables 16 and 17, Appendix A.] Sampling pattern "N" had been empirically tested, used a constant sampling size, and resulted in close approximations through the use of a table [see Table 11, Appendix A.]

Graphs 2 and 3 present the deviation of sampling pattern "N" estimates of Test Standard Deviation and Test Variance from actual values.

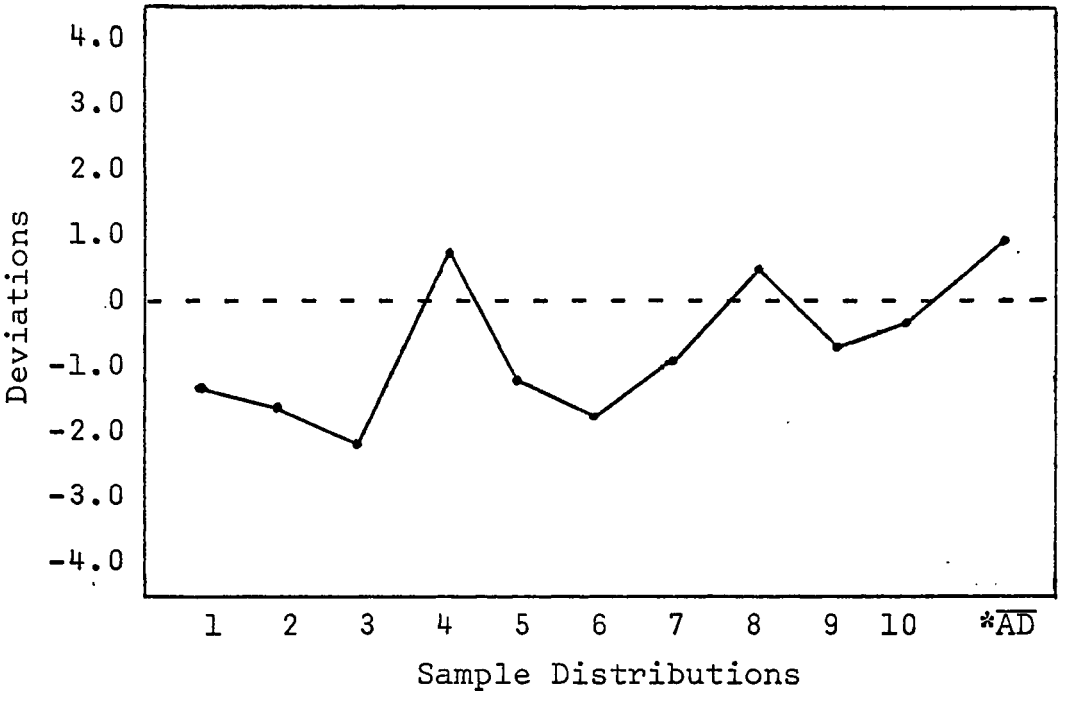
Graph 2

Deviation of Measurement Analysis Model Estimates of Test Standard Deviation from Actual Values Using Sampling Pattern "N" with Skewed Distributions

*mean absolute deviation.

Graph 3

Deviation of Measurement Analysis Model Estimates of Test Variance from Actual Values Using Sampling Pattern "N" with Skewed Distributions



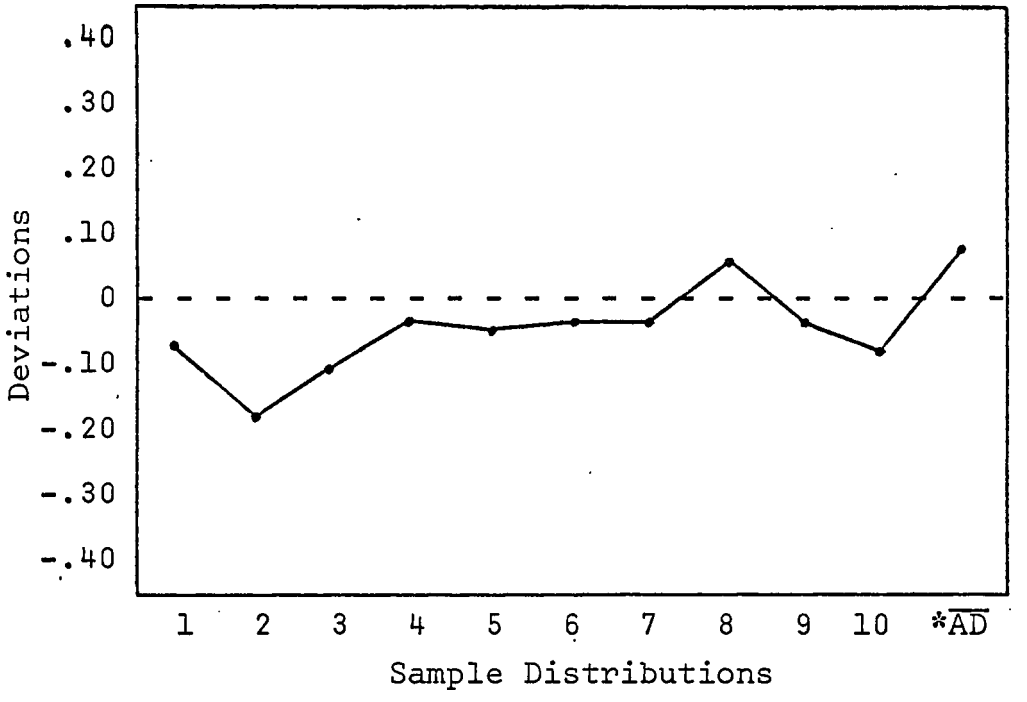
*mean absolute deviation.

Test Reliability Coefficient

Sampling patterns "G" and "N," presented in Graph 4, were selected as most suitable for estimating the Test Mean and Test Variance and also resulted in close approximations of the Test Reliability Coefficient. A table [see Table 12, Appendix A] was developed from the Kuder-Richardson Formula 21 with entry through the mean proportion and variance. Information on other sampling patterns is presented in Table 18, Appendix A.

Graph 4

Deviation of Measurement Analysis Model Estimates of Test Reliability Coefficient (KR 21) from Actual Values Using Sampling Patterns "G" and "N" with Skewed Distributions



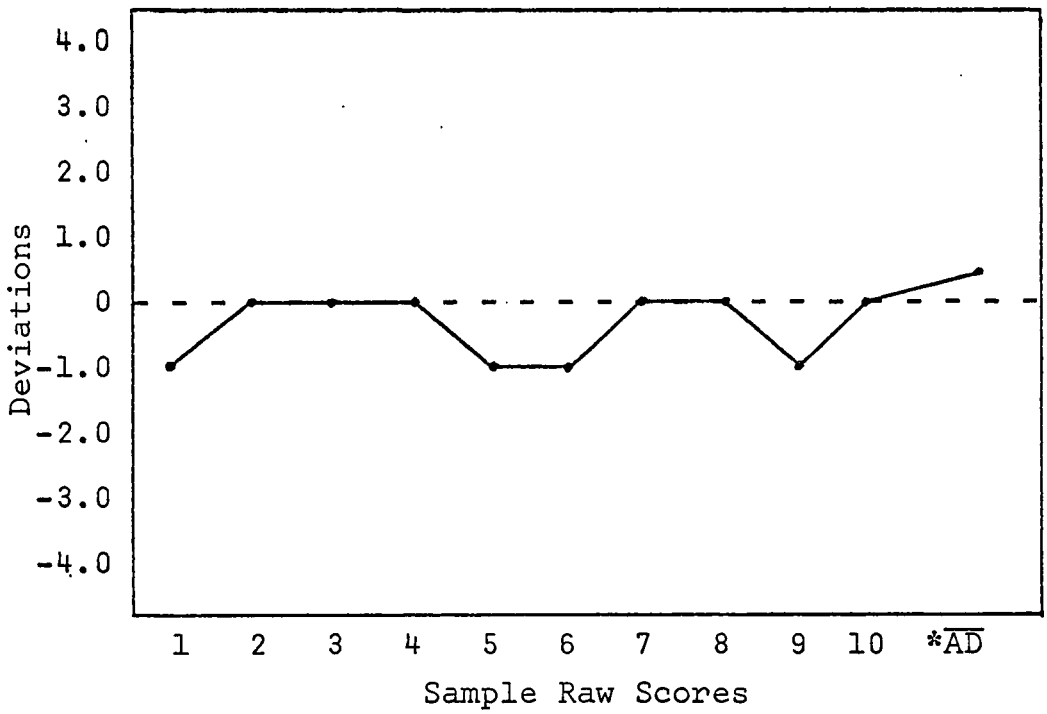
*mean absolute deviation.

Standard Scores

Sampling patterns "G" and "N," presented in Graph 5, were selected for estimating Standard Score values using a substitution technique and a preset mean of 80 and a standard deviation of 10. The above Standard Score values were selected as generating Standard Scores similar to scores most frequently found in a classroom situation. Information on other sampling patterns is presented in Table 19, Appendix A.

Graph 5

Deviation of Measurement Analysis Model Estimates of Standard Scores from Actual Values Using Sampling Patterns "G" and "N" with Skewed Distributions



*mean absolute deviation.

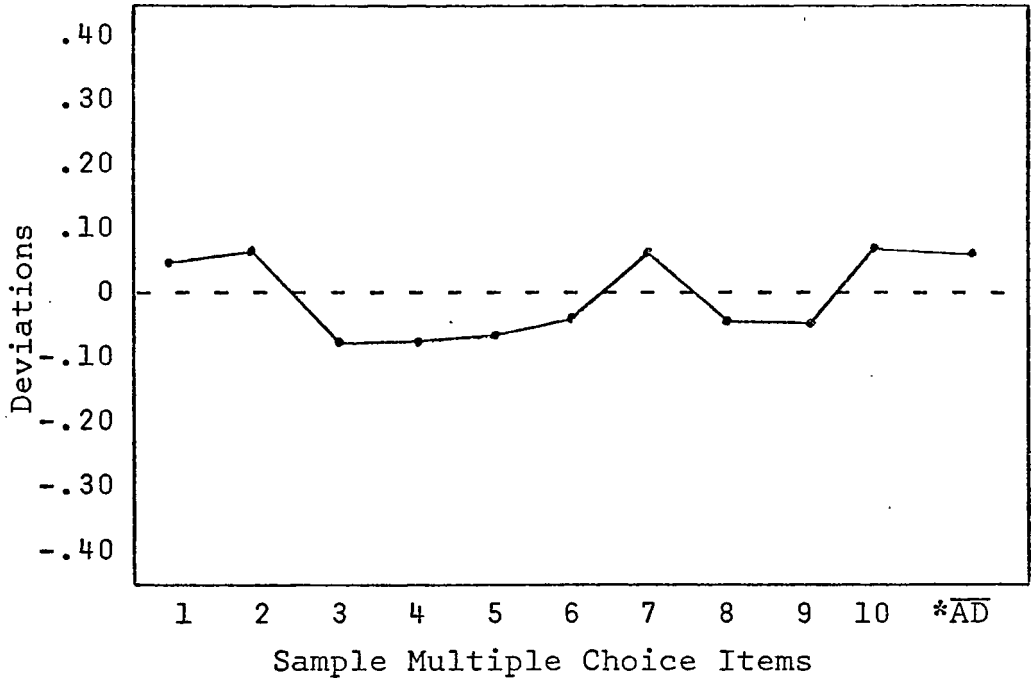
Item Discrimination Index

A point biserial correlation coefficient table [see Table 13, Appendix A] was generated using the centroid for 38.1 percent of the area in each of the upper and lower extremes of the normal distribution. Samples were taken from 100 cases in each of the upper and lower portions of the normal distribution using different proportions of subjects answering an item correctly. Entry to the table was through the number of subjects answering an item correctly from upper

and lower extremes of the population. Sampling pattern "D," presented in Graph 6, was selected as the closest approximation to 0.3810 of the extremes for an expected average classroom size of 26 [see Figure 8, Appendix A.] Although different sampling sizes for different class sizes would result in closer estimates [see Table 20, Appendix A], the additional accuracy would complicate the model.

Graph 6

Deviation of Measurement Analysis Model Estimates of Item Discrimination Index from Actual Values Using Sampling Pattern "D" with Skewed Distributions



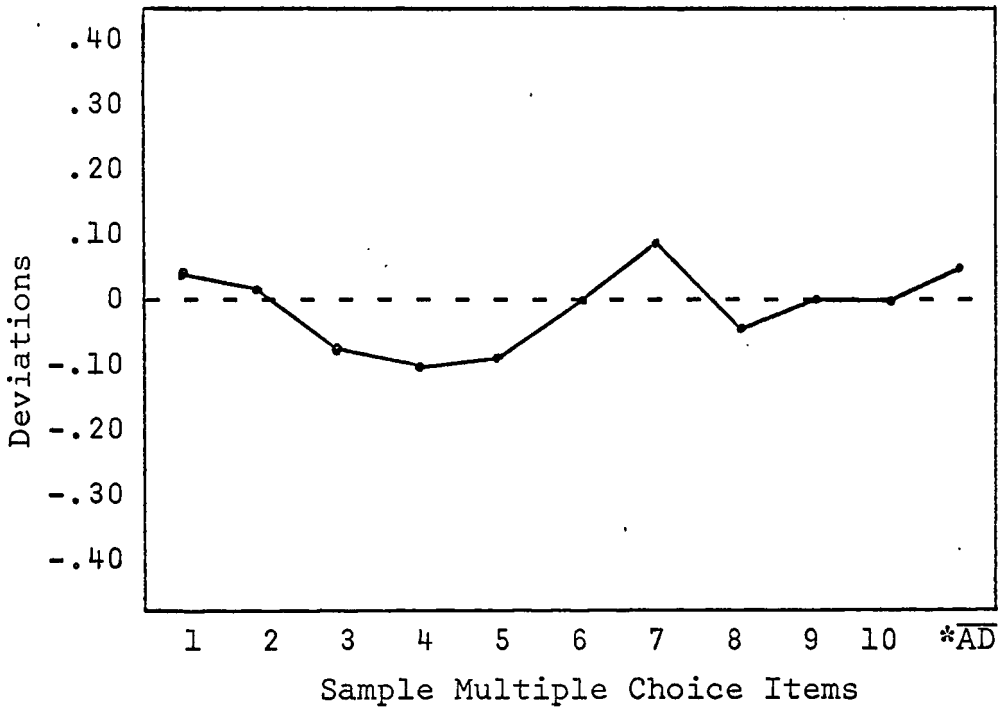
*mean absolute deviation.

Item Probability Index

A table [see Table 14, Appendix A] was generated using a standardized procedure for converting point biserial correlation coefficients with 18 degrees of freedom to probability indexes. Sampling pattern "D," presented in Graph 7 and previously selected for the Item Discrimination Index, resulted in very close approximations. Information on other sampling patterns is presented in Table 21, Appendix A.

Graph 7

Deviation of Measurement Analysis Model Estimates of Item Probability Index from Actual Values Using Sampling Pattern "D" with Skewed Distributions



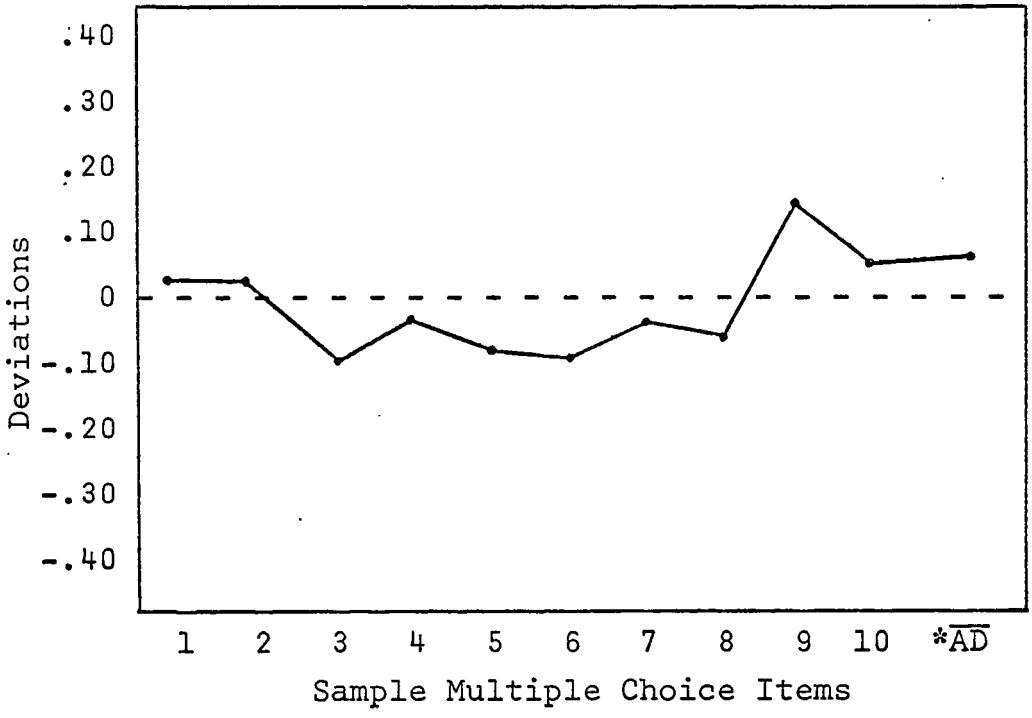
*mean absolute deviation.

Item Difficulty Index

Sampling pattern "D," presented in Graph 8 and previously selected for the Item Discrimination Index, resulted in close approximations of the Item Difficulty Index. Sampling pattern "D" also had the advantage of using a sample size of 10 which resulted in easy calculations. Information on other sampling patterns is presented in Table 22, Appendix A.

Graph 8

Deviation of Measurement Analysis Model Estimates of Item Difficulty Index from Actual Values Using Sampling Pattern "D" with Skewed Distributions



*mean absolute deviation.

POPULATIONS OF DIFFERENT CLASS SIZES

Sampling Procedures

The test results of 161 Foundations of Education students at the University of Houston on a multiple choice final test of fifty items were used as the sample population. Ten different sample sizes were drawn from the parent population using a random table of numbers. Table 3 presents the different class sizes used in this portion of the study.

The sampling patterns selected for evaluation of different skewed distributions were also evaluated over different class sizes. Estimates of major item analysis components were compared with the actual values derived from a standardized item analysis procedure and the absolute deviations calculated. Graphs 9 through 16 present the deviation from actual values using selected sampling patterns [see Tables 23-30, Appendix A for tabular presentation.]

Table 3

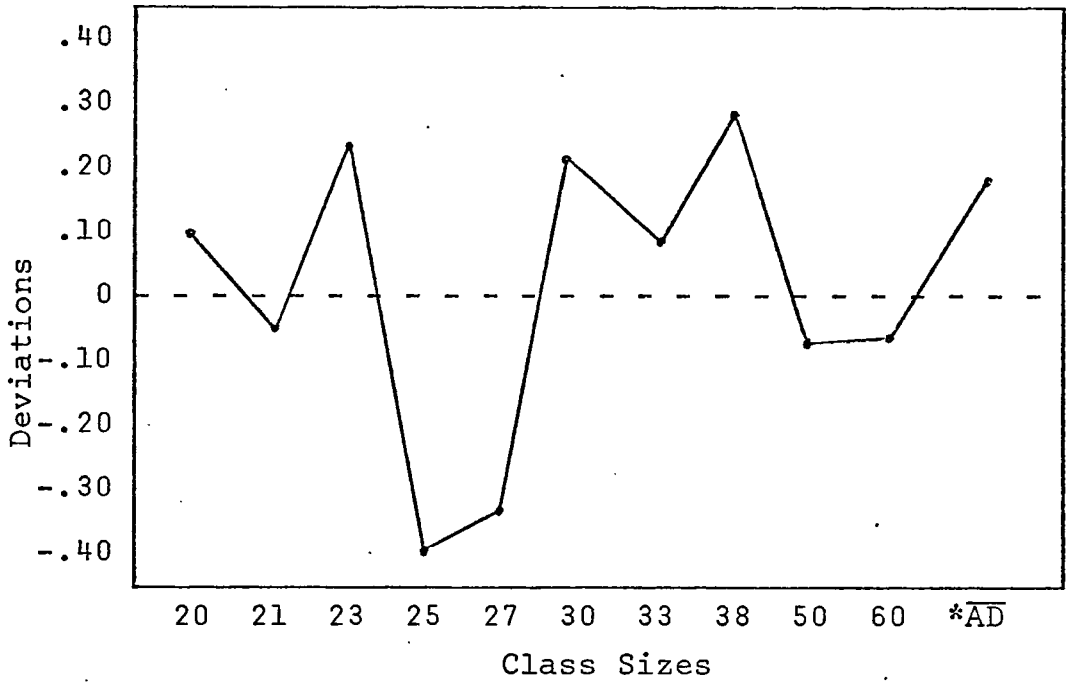
Different Class Sizes

Class Size	Mean	S.D.	Variance	Skewness	Kurtosis	r_{xx}
60	31.37	4.92	24.17	-.13	2.23	.53
50	31.78	4.50	20.26	-.72	3.66	.44
38	31.03	5.40	29.16	-.42	2.64	.61
33	31.42	4.74	22.44	.16	1.70	.49
30	30.50	4.52	20.40	.22	1.92	.43
27	30.33	4.71	22.15	-.54	3.35	.47
25	33.60	3.89	15.17	-.61	4.01	.28
23	29.57	3.87	14.98	.36	2.52	.20
21	30.14	3.81	14.53	.08	1.71	.18
20	30.90	4.15	17.25	.10	1.95	.32
Population	31.43	4.77	22.79	-.27	2.61	.50

Test Mean

Graph 9

Deviation of Measurement Analysis Model Estimates of Test Mean from Actual Values Using Sampling Pattern "G" with Different Class Sizes

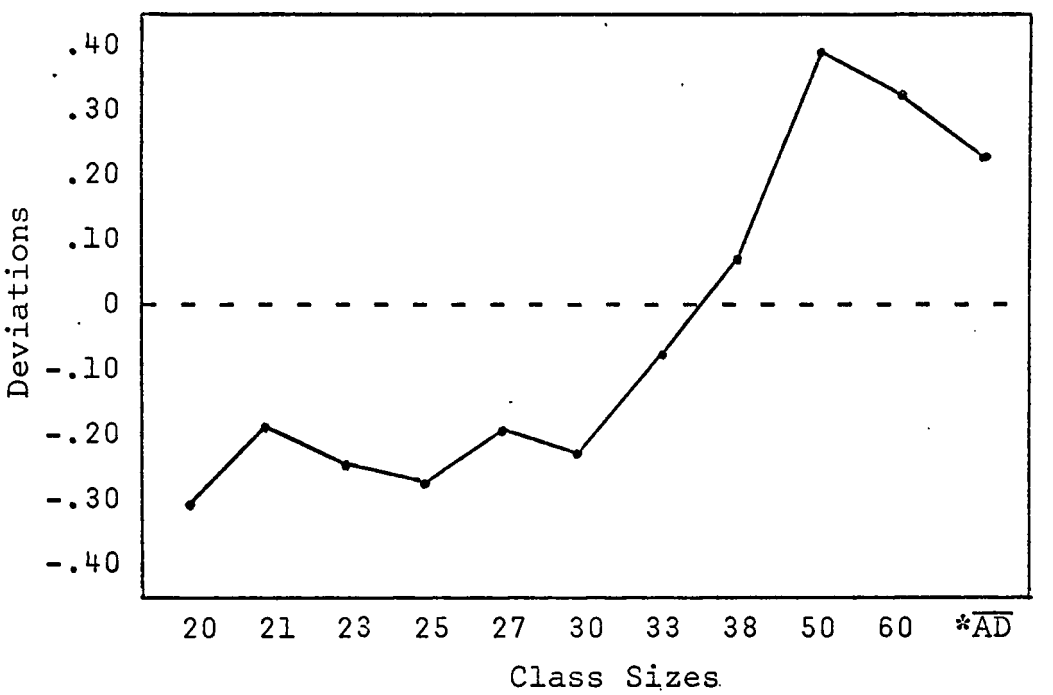


*mean absolute deviation.

Test Standard Deviation

Graph 10

Deviation of Measurement Analysis Model Estimates of Test Standard Deviation from Actual Values Using Sampling Pattern "N" with Different Class Sizes

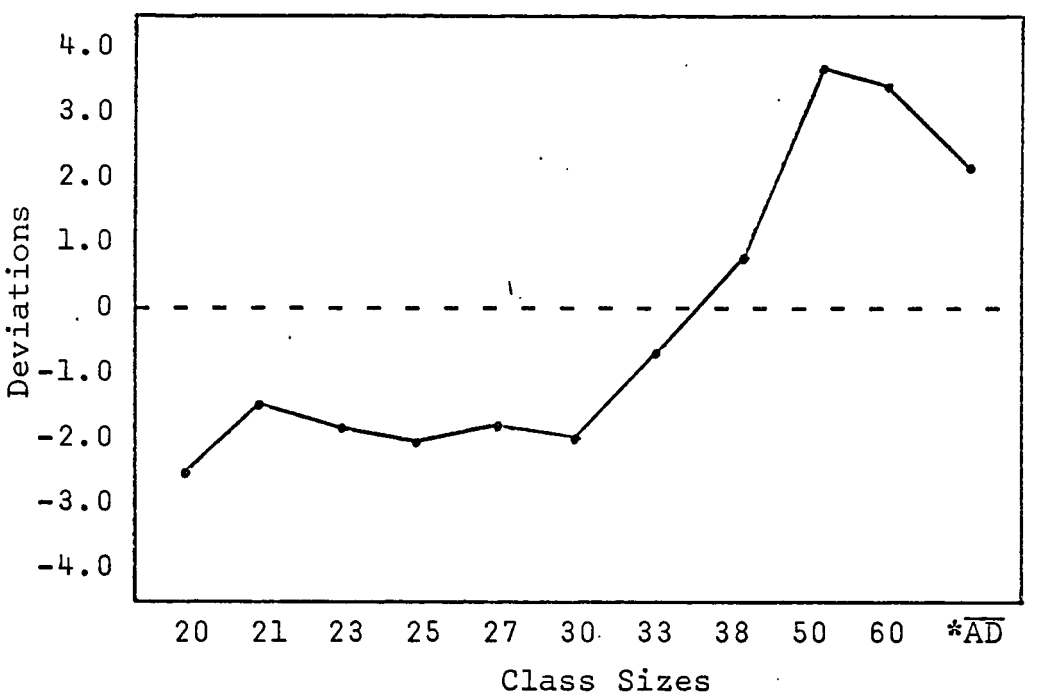


*mean absolute deviation.

Test Variance

Graph 11

Deviation of Measurement Analysis Model Estimates of Test Variance from Actual Values Using Sampling Pattern "N" with Different Class Sizes

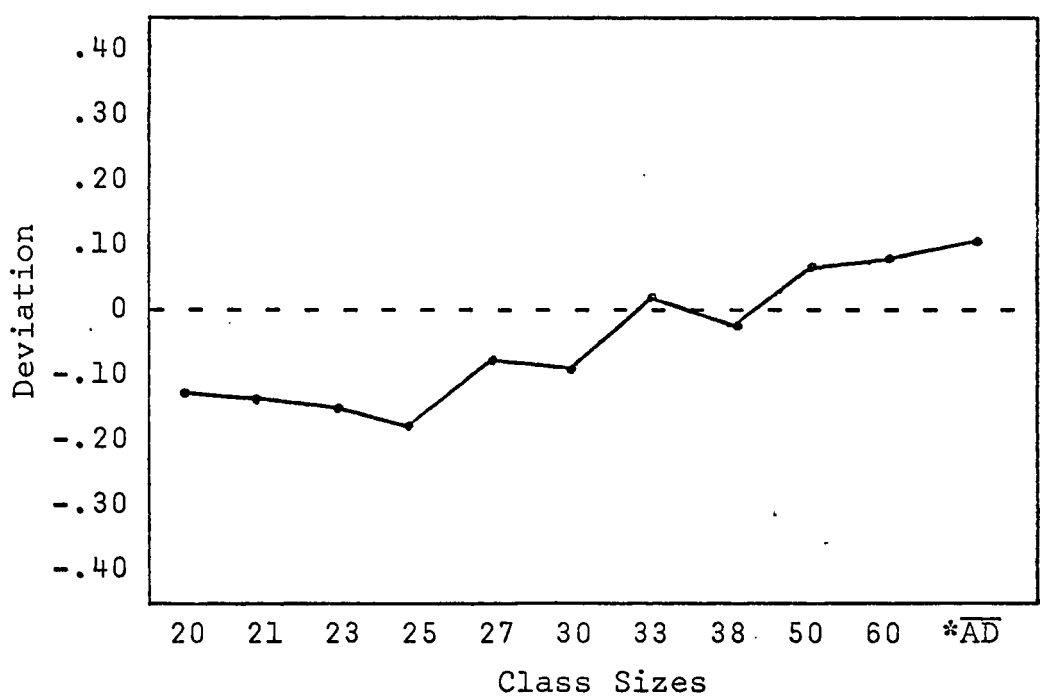


*mean absolute deviation.

Test Reliability Coefficient

Graph 12

Deviation of Measurement Analysis Model Estimates of Test Reliability Coefficient (KR 21) from Actual Values Using Sampling Patterns "G" and "N" with Different Class Sizes

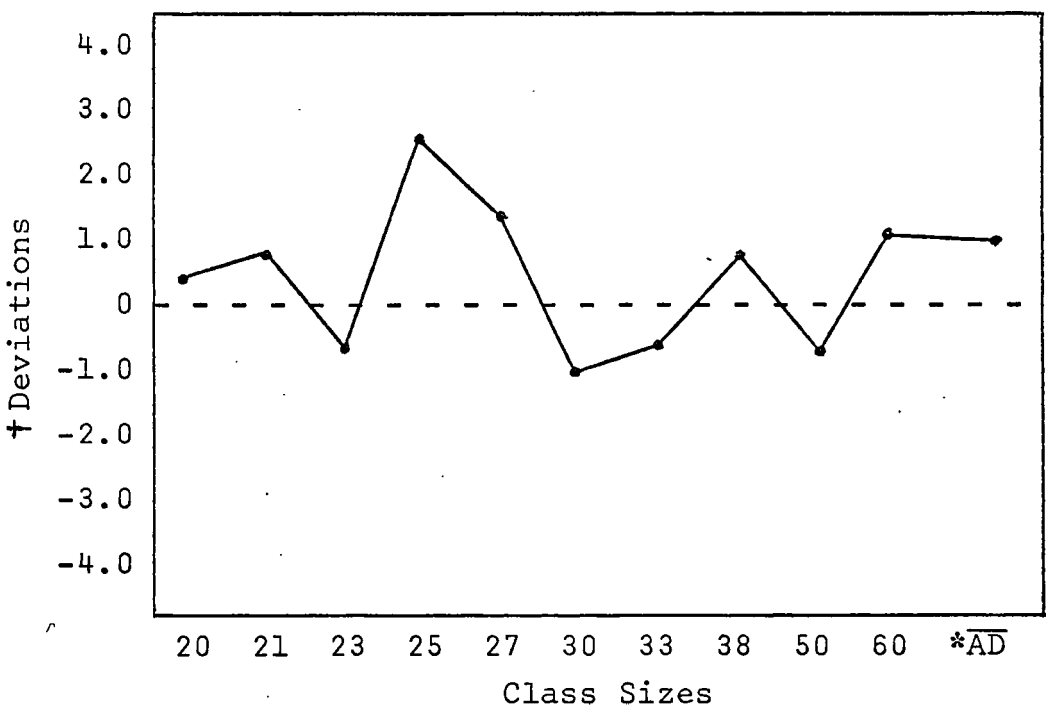


*mean absolute deviation.

Standard Scores

Graph 13

Deviation of Measurement Analysis Model Estimates of Standard Scores from Actual Values Using Sampling Patterns "G" and "N" with Different Class Sizes



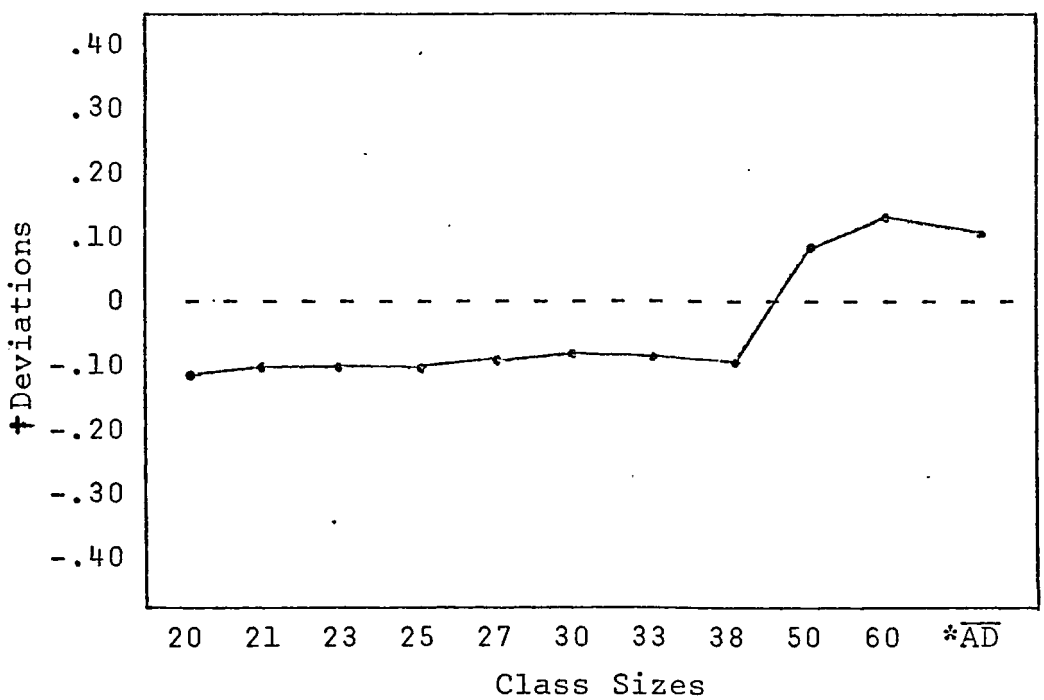
*mean absolute deviation.

†average deviation with sign retained to show direction of most consistent error.

Item Discrimination Index

Graph 14

Deviation of Measurement Analysis Model Estimates of Item Discrimination Index from Actual Values Using Sampling Pattern "D" with Different Class Sizes



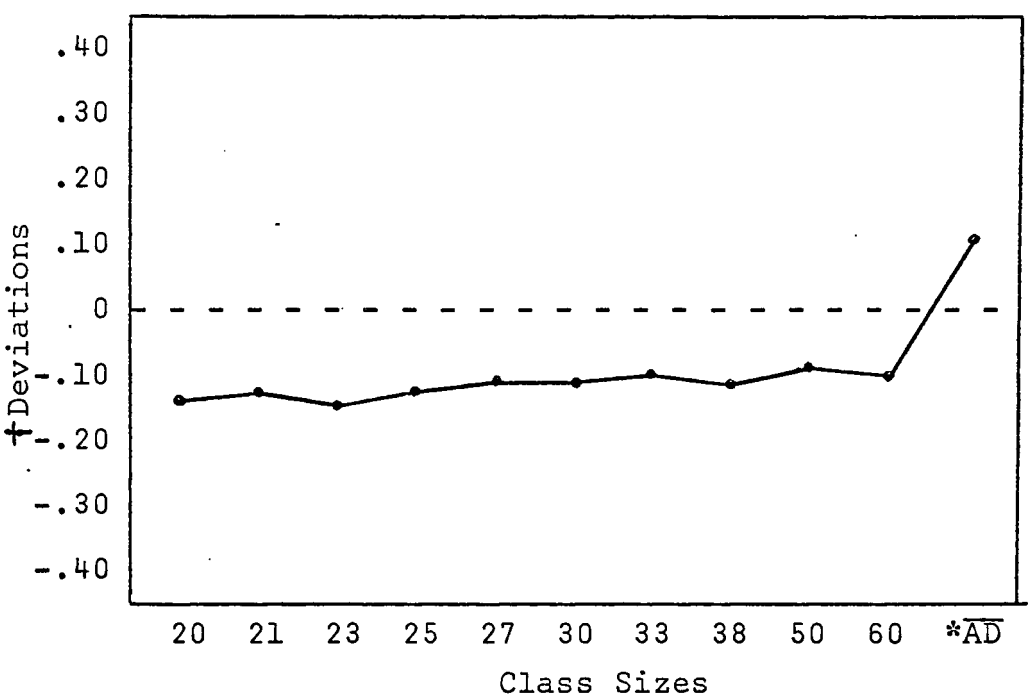
*mean absolute deviation.

†average deviation with sign retained to show direction of most consistent error.

Item Probability Index

Graph 15

Deviation of Measurement Analysis Model Estimates of Item Probability Index from Actual Values Using Sampling Pattern "D" with Different Class Sizes



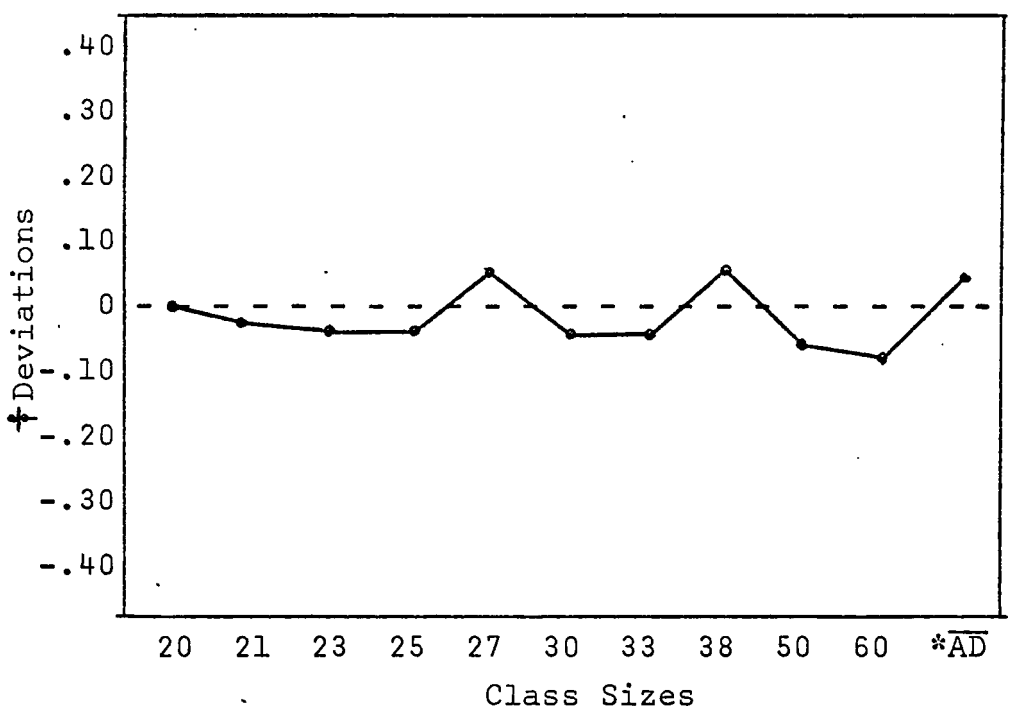
*mean absolute deviation.

†average deviation with sign retained to show direction of most consistent error.

Item Difficulty Index

Graph 16

Deviation of Measurement Analysis Model Estimates of Item Difficulty Index from Actual Values Using Sampling Pattern "D" with Different Class Sizes



*mean absolute deviation.

†average deviation with sign retained to show direction of most consistent error.

MODEL FORMAT

Pilot testing brought out a problem in using the Measurement Analysis Model--in that, teachers encountered difficulty in using simultaneously the programmed steps and tables in the appendix. It was decided that the material would be better separated into two parts. Part I consisted of short programmed steps for each major item analysis component prefaced by an introductory statement explaining its use in measurement. Part II consisted of definitions, examples, and tables. All steps in Part I were coded to match examples and tables in Part II. This arrangement resulted in the teacher being able to refer to Part II while using Part I.

GENERALIZATIONS

Although exploratory research resulted in 18 different sampling patterns which demonstrated promise in estimating major item analysis components, sampling patterns using constant sample sizes for all class sizes were selected as most feasible for classroom use. All sampling patterns selected for final inclusion in the Measurement Model resulted in very close approximations of actual values obtained by a standardized item analysis procedure. More precise sampling techniques were found; however, they would require more complex calculation procedures and/or extensive tables which would

complicate the model. The Measurement Analysis Model used in this study is given in Appendix B.

Chapter 4

FIELD INVESTIGATIVE PROCEDURES AND FINDINGS

OBJECTIVITY

Subjects

Fifty-five experienced teachers, members of the summer school staff of a local public school district, were selected to form the experienced group and forty-one students, randomly selected from an undergraduate course in teacher education at the University of Houston, formed the inexperienced group.

Null Hypothesis

It was hypothesized that there would be no difference in the decisions made on the discriminal ability of an item by experienced or inexperienced subjects using or not using the Measurement Analysis Model.

Procedures

Group I (experienced) and Group III (inexperienced) received the Measurement Analysis Model as a guide and were given instructions to analyze student responses and rate twenty sample test items taken at random from the results of a fifty-item test on the foundations of education given at

the University of Houston. Group II (experienced) and Group IV (inexperienced) received the same instructions to analyze the same student responses and rate the same twenty test items; however, they were not allowed to use the Measurement Analysis Model.

Experienced Teachers

A short, five minute briefing was given to all experienced teachers in one general meeting and each teacher was then given a separate folder containing either instructions, test results, and the Measurement Analysis Model or just instructions and test results. Each folder was numbered consecutively from 1 to 55. Odd numbers, containing the Measurement Analysis Model, formed Group I and the even numbers, containing only instructions and test results, formed Group II. All teachers were cautioned to complete the ratings on an individual basis and return the completed folders. The members of Group I received 28 folders and returned 18--two were improperly completed. The members of Group II received 27 folders and returned 19--one was improperly completed and two were received too late for inclusion in the final results.

Inexperienced Subjects

A short, five minute briefing was given to all inexperienced subjects in one general meeting and each subject was given a separate folder containing either instructions, test

results, and the Measurement Analysis Model or just instructions and test results. Each folder was numbered consecutively from 101 to 141. Odd numbers, containing the Measurement Analysis Model, formed Group III and the even numbers, containing only instructions and test results, formed Group IV. All subjects were cautioned to complete the ratings on an individual basis and return the completed folders. The members of Group III received 21 folders and returned 18--one was improperly completed and one was received too late for inclusion in the final results. The members of Group IV received 20 folders and returned 18--two were improperly completed.

Statistical Techniques

An analysis of variance, presented in Table 4, was used to analyze the data [see Tables 31 through 34, Appendix C]; a Shapiro-Wilk analysis of variance test[1] as a check on the normality of the distribution within each cell; an F-maximum test for homogeneity of variance among cells; and a Scheffé comparison of means test for critical differences, presented in Table 5. A coefficient of equivalence was also calculated as a measure of the inherent objectivity of the Measurement Analysis Model. All tests were conducted at the .05 level of significance.

---

[1]S. S. Shapiro and M. B. Wilk, "An Analysis of Variance Test for Normality (Complete Samples)," Biometrika LII (December, 1965), pp. 591-611.

Findings

Table 4

Analysis of Variance Using Data Received from Experienced and Inexperienced Subjects Rating Test Items With and Without the Measurement Analysis Model

Source	SS	df	ms	F
Total	312.58	63		
Model	253.17	1	253.17	259.59*
Experience	0.34	1	0.34	0.35
Model x Experience	0.56	1	0.56	0.57
Error	58.52	60	0.98	

*$p < .05$.

Groups I and III did not meet the requirements of the Shapiro-Wilk test for normality of distribution at the .05 level of significance.

The results of the F-maximum test for homogeneity of variances used to check the assumption of equality between cells showed the variance for Group III (0.191) as significantly different from the variances for Groups I (1.075), II (0.988), and IV (1.646) at the .05 level of significance.

Table 5

Scheffé Comparison of Means Test
for Critical Differences

Group	II	III	IV	Mean
	F-value	F-value	F-value	
I	5.416*	0.476	5.474	2.785
II		5.891*	0.059	6.576
III			5.950*	2.452
IV				6.617

*$p < .05$.

Using the actual values [see Table 35, Appendix C] obtained from a standardized item analysis procedure as the true score and the model values as the estimated score, a Spearman's rho correlation coefficient of 0.824 was obtained which was significant at the .05 level.

RELIABILITY

The data collected to investigate the objectivity of the Measurement Analysis Model was also used to determine the reliability.

Null Hypothesis

It was hypothesized that there would be no relationship between decisions made on the discriminal ability of an

item by the same experienced or inexperienced subject using or not using the Measurement Analysis Model.

Procedures

A split-half reliability procedure, presented in Table 6, was used to compare responses to odd and even test items for each group and the Spearman-Brown Prophecy formula was used to estimate the full reliability of the test if restored to its original length. A t-test was made at the .05 level of significance.

Findings

Table 6

Split-half Reliability Coefficients for Data Received from Experienced and Inexperienced Subjects Rating Test Items Using or Not Using the Measurement Analysis Model

Group	Split-half Value	Spearman Brown Value
I	0.45*	0.62*
II	0.17	0.29
III	0.29	0.45*
IV	0.43*	0.60*

*$p < .05$.

VALIDATION AND FEASIBILITY

Subjects

Teacher-made tests were selected from tests given during summer school at a local school district. Tests received were varied in format--essays, true-false, short answer, and multiple choice. They also varied in scope--unit tests, final tests, term projects, and charts. Areas tested varied from first grade reading and arithmetic to twelfth grade civics.

Sixty-eight tests were received and twenty-three met the requirements for a partial item analysis. The primary reason for rejecting most of the tests was an insufficient number of students taking each test. Only one of the tests received was suitable for a complete item analysis. The primary reason for rejecting the other tests was differential weighting of individual items.

Null Hypotheses

1. Validation. It was hypothesized that there would be no relationship between major item analysis components obtained by using the Measurement Analysis Model and actual values obtained by using a standardized item analysis procedure.

2. Feasibility. It was hypothesized that the Measurement Analysis Model would not operate under actual classroom

conditions and result in valid and reliable item analysis components.

Procedures

The Measurement Analysis Model was used to obtain major item analysis components from teacher-made tests under actual classroom conditions [see Tables 36 through 38, Appendix A.] The components were then compared with those obtained by a standardized item analysis procedure using appropriate correlational techniques [see Tables 7, 8, and 9.] A *t*-test was made and the Shapiro-Wilk test was used to verify normality of distribution. Both tests were made at the .05 level of significance.

Findings

Table 7

Comparison of Measurement Analysis Model Estimates of Test Mean, Test Standard Deviation, Test Variance, and Test Reliability Coefficient with Actual Values Obtained from a Standardized Item Analysis Procedure

Subject	Test Mean	Test Standard Deviation	Test Variance	Test Reliability Coefficient
Correlation Coefficient	0.988*	0.975	0.972	0.960

*Spearman's rho used for Test Mean; Pearson's *r* used for all others. All correlation coefficients were significant at the .05 level.

Table 8

Comparison of Measurement Analysis Model Estimates of Standard Scores with Actual Values Obtained from a Standardized Item Analysis Procedure

Subject	Teacher-made Tests				
	A	B	C	D	E
Correlation Coefficient	0.999	1.000*	1.000*	1.000*	1.000

Subject	Teacher-made Tests				
	F	G	H	J	K
Correlation Coefficient	1.000	1.000*	1.000*	1.000*	1.000*

Subject	Teacher-made Tests				
	L	M	N	P	Q
Correlation Coefficient	1.000*	1.000*	1.000	1.000*	1.000*

Subject	Teacher-made Tests				
	R	S	T	U	V
Correlation Coefficient	0.999	1.000*	1.000*	1.000	1.000

*Spearman's rho used for these tests; Pearson's r used for all others. All correlation coefficients were significant at the .05 level.

Table 9

Comparison of Measurement Analysis Model Estimates of Item Discrimination Index, Item Probability Index, and Item Difficulty Index with Actual Values Obtained from a Standardized Item Analysis Procedure

Subject	Item Discrimination Index	Item Probability Index	Item Difficulty Index
Correlation Coefficient	0.800*	0.830*	1.000*

*Spearman's rho used for all indexes. All correlation coefficients were significant at the .05 level.

Chapter 5

SUMMARY, CONCLUSIONS, AND RECOMMENDATIONS

SUMMARY

Objectivity

The null hypothesis that there would be no difference in the decisions made on the discriminal ability of an item by experienced or inexperienced subjects using or not using the Measurement Analysis Model was rejected. It was concluded that subjects using the Measurement Analysis Model did make the same relative decisions on the discriminal ability of sample items while subjects without the model were unable to do so. Experience did not seem to be a factor. Using a standardized item analysis procedure to ascertain the true score, the results of a coefficient of equivalence test indicated that the Measurement Analysis Model was an inherently objective instrument.

Reliability

The null hypothesis that there would be no relationship between decisions made on the discriminal ability of an item by the same experienced or inexperienced subject using or not using the Measurement Analysis Model was rejected. It was concluded that subjects using the Measurement Analysis

Model were able to make consistent decisions when rating a measurement item. Group IV, without the model, also made consistent decisions; however, an inspection of the data indicated that subjects based their decisions on the Item Difficulty Index.

Validation

The null hypothesis that there would be no relationship between major item analysis components obtained by using the Measurement Analysis Model and actual values obtained by using a standardized item analysis procedure was rejected. It was concluded that there was a significant relationship between estimates of major item analysis components obtained by using the Measurement Analysis Model and actual values obtained by using a standardized item analysis procedure.

Feasibility

The null hypothesis that the Measurement Analysis Model would not operate under actual classroom conditions and result in valid and reliable item analysis components was rejected. It was concluded that the Measurement Analysis Model did operate under actual classroom conditions and did obtain estimates of major item analysis components which were found to be significantly related to the actual values obtained by using a standardized item analysis procedure. This

was true of all classroom tests which met the basic requirements for an item analysis.

CONCLUSIONS

The study demonstrated that a Measurement Analysis Model could be developed which would yield valid and reliable item analysis components, even when used by relatively unskilled subjects. An analysis of the data revealed arithmetic operations (particularly the use of negative numbers) which interfered with the effectiveness of the model. The study also pinpointed areas where modifications may be made to build a more efficient Measurement Analysis Model.

Discussion

An unresolved question is whether teachers would accept any valid and reliable psychometric procedure as a useful tool in evaluating measurement procedures used in the classroom. There is ample theoretical and empirical evidence that accurate item analysis procedures, some requiring a minimum of computational skill, are available for use in the classroom. This study did not include provisions for the development of a positive frame of reference toward measurement by each classroom teacher. It also failed to consider evidence that teachers have shown little interest in increasing their measurement and evaluation skills through formal training.

The inability of many subjects to handle negative numbers demonstrated a lack of basic computation skills; however, a reassessment of the Measurement Analysis Model indicated that it was possible to go directly to the Item Probability Index without computing the Item Discrimination Index. Not only would this simplify the model, it would also remove a major stumbling block for many teachers.

The development of more tables for sampling from the extremes could also reduce the sampling patterns to two. This would not only streamline the model, but would increase the accuracy and allow sampling of fewer cases.

The table of point biserial values developed, using the centroid of the area sampled, has a definite advantage over most tables currently in use--it does not generate a consistent error when estimating the Item Discrimination Index for items ranging in difficulty from 0.30 to 0.80. In fact, there was evidence that the values generated were closer estimates of the true discriminal power of an item than the actual point biserial values, especially for items labeled "hard" or "easy."

The use of the normal distribution as the foundation for the Measurement Analysis Model had a tendency to correct the variance estimates obtained in skewed samples from the parent population. There seemed to be some evidence that the variance estimates generated by the model were closer to the

values of the population parameters than the actual values obtained from a standardized item analysis procedure.

RECOMMENDATIONS

In-Service Instruction

A paramount need pointed out by this study is for in-service instruction for classroom teachers in valid and reliable measurement and evaluation procedures. The study demonstrated that many teachers were unable to evaluate current measurement procedures used in the classroom and were also unable to identify valid and reliable decision points which would allow them to determine who does and who does not possess the trait they were attempting to measure. It is urgently recommended that a course in measurement and evaluation be a basic requirement for teacher certification.

Sampling from Extremes of Populations

Empirical evidence gathered in developing the Measurement Analysis Model indicated that extremely small samples could be taken from the extremes of rank ordered distributions to obtain very close estimates of major item analysis components. Preliminary evidence indicated that samples varying from 0.1667 to 0.5000, taken from each tail of a distribution, would result in close approximations to actual values obtained by using a standardized item analysis procedure; however, more definitive research is needed to

determine whether there is an optimum area to sample from the extremes.

Estimating Test Variance

The use of mean deviation scores, using the correction terms generated by the predicted value of the true centroid intersection point on the z-axis of the normal distribution, resulted in very close approximations of the Test Standard Deviation and the Test Variance. Further research seems justified to determine whether 0.3810 or 0.2703 represents the optimum area to be sampled from the extremes when estimating the variance. It also seems possible that a quick and simple test for checking the normality of a distribution can be developed using this procedure.

Table of Point Biserial Values

Preliminary research indicated the validity of generating a table of point biserial values from the centroid of the normal distribution. Cursory examination of evidence generated during a pilot study indicated that many different z-scale values generate accurate tables; however, a value of 0.797885 generates limits for the point biserial which are closest to its theoretical values. Further research is needed to gather empirical evidence on the optimum z-scale value to use in generating a table of point biserial values.

Measurement Analysis Model

The results of this study indicate a need for further exploratory research in developing a valid and reliable Measurement Analysis Model for use in the classroom. Although the model developed for this study could be used effectively in a classroom, there needs to be more research with even more simplified techniques to assist a teacher in validating measurement procedures used in a classroom.

APPENDIX A: Development of Measurement Analysis Model

Figure 2

Developing Evaluation Designs

Focusing the Evaluation*

1. Define the decision situations to be served, and describe each one in terms of its locus, criteria, decision rules, timing, and decision alternatives.
2. Define the system to be evaluated.
3. Define the evaluation specifications.

Collection of Information

1. Specify each item of information that is to be collected.
2. Specify the populations, sources, and sampling procedures for information collection.
3. Specify the instruments and methods for information collection.
4. Specify the arrangements and schedule for information collection.

Organization of Information

1. Specify a format for organizing the information.
2. Specify a means for coding, organizing, storing, and retrieving the information.

Analysis of Information

1. Specify the procedures for analyzing the information.
2. Specify a means for performing the analysis of information.

Reporting of Information

1. Specify the audiences for the evaluation reports.
2. Specify formats for the evaluation reports and reporting sessions.
3. Specify a means for providing the information to the audiences.
4. Specify a schedule for reporting the information to the specified audiences.

Administration of the Evaluation

1. Summarize the evaluation schedule.
2. Define staff and resource requirements and plans for meeting these requirements.
3. Specify means for meeting policy requirements for conduct of the evaluation.

*The logical structure of the evaluation design is the same for all types of evaluation, whether context, input, process, or product.

Figure 2 (continued)

4. Appraise the potential of the evaluation design for providing information which is valid, reliable, credible, timely, and pervasive.
5. Specify and schedule means for periodic updating of the evaluation design.
6. Provide a budget for the total evaluation program.

Source: Egon G. Guba and Daniel L. Stufflebeam, Evaluation: The Process of Stimulating, Aiding, and Abetting Insightful Action, Monograph Series in Reading Education, No. 1 (Bloomington, Indiana: Indiana University Press, 1970), p. 29.

Figure 3

Sampling Patterns Used in Developing
Measurement Analysis Model

Pattern	Description
A	50 percent of cases from both upper and lower extremes.
B	40 percent of cases from both upper and lower extremes.
C	38.10 percent of cases from both upper and lower extremes.
D	33.33 percent of cases from both upper and lower extremes.
E	27.03 percent of cases from both upper and lower extremes.
F	16.67 percent of cases from both upper and lower extremes.
G	percentile values of .05, .15, .25, .35, .45, .55, .65, .75, .85, and .95 with a constant multiplier of .10 for Test Mean.
H	six cases taken from every second case starting at each extreme.
J	seven cases taken from every second case starting at each extreme.
K	seven cases using extreme cases and then every second case starting at each extreme.
L	eight cases using extreme cases and then every second case starting at each extreme.
M	percentile values of .03, .10, .20, .30, .50, .50, .70, .80, .90, and .97 with a constant multiplier of .1104 for the Standard Deviation.
N	percentile values of .985, .95, .90, .84, .75, .25, .16, .10, .05, and .015 with a constant multiplier of .0739 for the Standard Deviation.
P	three cases from each extreme and then four cases from every second case starting at each extreme.
Q	ten cases from each extreme as determined by an iterative process.
R	five cases from extremes and then five cases from every second case starting at each extreme.
S	seven cases from extremes and then three cases from every second case starting at each extreme.
T	seven cases from extremes and then four cases from every second case starting at each extreme.

Figure 4

Positions to be Sampled in Estimating Test Mean

Class Size	Positions to be Sampled									
20	1	3	5	7	9	12	14	16	18	20
21	1	3	5	7	9	13	15	17	19	21
22	1	3	6	8	10	13	15	17	20	22
23	1	3	6	8	10	14	16	18	21	23
24	1	4	6	8	11	14	17	19	21	24
25	1	4	6	9	11	15	17	20	22	25
26	1	4	7	9	12	15	18	20	23	26
27	1	4	7	9	12	16	19	21	24	27
28	1	4	7	10	13	16	19	22	25	28
29	1	4	7	10	13	17	20	23	26	29
30	2	5	8	11	14	17	20	23	26	29
31	2	5	8	11	14	18	21	24	27	30
32	2	5	8	11	14	19	22	25	28	31
33	2	5	8	12	15	19	22	26	29	32
34	2	5	8	12	15	20	23	27	30	33
35	2	5	9	12	16	20	24	27	31	34
36	2	5	9	13	16	21	24	28	32	35
37	2	6	9	13	17	21	25	29	32	36
38	2	6	10	13	17	22	26	29	33	37
39	2	6	10	14	18	22	26	30	34	38
40	2	6	10	14	18	23	27	31	35	39
41	2	6	10	14	18	24	28	32	36	40
42	2	6	10	15	19	24	28	33	37	41
43	2	6	11	15	19	25	29	33	38	42
44	2	7	11	15	20	25	30	34	38	43
45	2	7	11	16	20	26	30	35	39	44
46	2	7	12	16	21	26	31	35	40	45
47	2	7	12	16	21	27	32	36	41	46
48	2	7	12	17	22	27	32	37	42	47
49	2	7	12	17	22	28	33	38	43	48
50	2	8	12	18	22	29	33	39	43	49

Figure 5

Positions to be Sampled in Estimating
Both Test Mean and Test Standard Deviation

Class Size	Positions to be Sampled									
20	1	2	4	6	10	11	15	17	19	20
21	1	2	4	6	10	12	16	18	20	21
22	1	2	4	7	11	12	16	19	21	22
23	1	2	5	7	11	13	17	19	22	23
24	1	2	5	7	12	13	18	20	23	24
25	1	2	5	8	12	14	18	21	24	25
26	1	3	5	8	13	14	19	22	24	26
27	1	3	5	8	13	15	20	23	25	27
28	1	3	6	8	14	15	21	23	26	28
29	1	3	6	9	14	16	21	24	27	29
30	1	3	6	9	15	16	22	25	28	30
31	1	3	6	9	15	17	23	26	29	31
32	1	3	6	10	16	17	23	27	30	32
33	1	3	7	10	16	18	24	27	31	33
34	1	3	7	10	17	18	25	28	32	34
35	1	4	7	10	17	19	26	29	32	35
36	1	4	7	11	18	19	26	30	33	36
37	1	4	7	11	18	20	27	31	34	37
38	1	4	8	11	19	20	28	31	35	38
39	1	4	8	12	19	21	28	32	36	39
40	1	4	8	12	20	21	29	33	37	40
41	1	4	8	12	20	22	30	34	38	41
42	1	4	8	13	21	22	30	35	39	42
43	1	4	9	13	21	23	31	35	40	43
44	1	4	9	13	22	23	32	36	41	44
45	1	4	9	14	22	24	33	37	42	45
46	1	5	9	14	23	24	34	38	42	46
47	1	5	9	14	23	25	35	39	43	47
48	1	5	10	14	24	25	35	39	44	48
49	1	5	10	15	24	26	36	40	45	49
50	2	5	10	15	25	26	37	41	46	49

Figure 6

Positions to be Sampled in Estimating Test Standard Deviation

Class Size	Positions to be Sampled									
20	1	2	3	4	5	16	17	18	19	20
21	1	2	3	4	5	17	18	19	20	21
22	1	2	3	4	6	17	19	20	21	22
23	1	2	3	4	6	18	20	21	22	23
24	1	2	3	4	6	19	21	22	23	24
25	1	2	3	4	6	20	22	23	24	25
26	1	2	3	4	7	20	23	24	25	26
27	1	2	3	4	7	21	24	25	26	27
28	1	2	3	4	7	22	25	26	27	28
29	1	2	3	5	7	23	25	27	28	29
30	1	2	3	5	8	23	26	28	29	30
31	1	2	3	5	8	24	27	29	30	31
32	1	2	3	5	8	25	28	30	31	32
33	1	2	3	5	8	26	29	31	32	33
34	1	2	3	5	8	27	30	32	33	34
35	1	2	4	6	9	27	30	32	34	35
36	1	2	4	6	9	28	31	33	35	36
37	1	2	4	6	9	29	32	34	36	37
38	1	2	4	6	10	29	33	35	37	38
39	1	2	4	6	10	30	34	36	38	39
40	1	2	4	6	10	31	35	37	39	40
41	1	2	4	7	10	32	35	38	40	41
42	1	2	4	7	10	33	36	39	41	42
43	1	2	4	7	11	33	37	40	42	43
44	1	2	4	7	11	34	38	41	43	44
45	1	2	4	7	11	35	39	42	44	45
46	1	2	5	7	12	35	40	42	45	46
47	1	2	5	8	12	36	40	43	46	47
48	1	2	5	8	12	37	41	44	47	48
49	1	2	5	8	12	38	42	45	48	49
50	1	2	5	8	12	39	43	46	49	50

Figure 7

Sampling Distribution Developed for Estimating Major Item Analysis Components

Class Size	Positions to be Sampled																			
20	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
21	1	2	3	4	5	6	7	8	9	10	12	13	14	15	16	17	18	19	20	21
22	1	2	3	4	5	6	7	8	9	10	13	14	15	16	17	18	19	20	21	22
23	1	3	4	5	6	7	8	9	10	11	13	14	15	16	17	18	19	20	21	23
24	1	3	4	5	6	7	8	9	10	12	13	15	16	17	18	19	20	21	22	24
25	1	3	4	5	6	7	8	9	10	12	14	16	17	18	19	20	21	22	23	25
26	1	3	4	5	7	8	9	10	11	13	14	16	17	18	19	20	22	23	24	26
27	1	3	4	5	7	8	9	10	11	13	15	17	18	19	20	21	23	24	25	27
28	1	3	4	5	7	8	10	11	12	14	15	17	18	19	21	22	24	25	26	28
29	1	2	4	5	7	8	10	11	12	14	16	18	19	20	22	23	25	26	27	29
30	1	3	4	5	7	8	10	11	13	15	16	18	20	21	23	24	26	27	28	30
31	1	3	4	5	7	8	10	11	13	15	17	19	21	22	24	25	27	28	29	31
32	1	3	4	5	7	8	10	12	14	16	17	19	21	23	25	26	28	29	30	32
33	1	3	4	5	7	8	10	12	14	16	18	20	22	24	26	27	29	30	31	33
34	1	3	4	5	7	8	10	12	14	16	19	21	23	25	27	28	30	31	32	34
35	1	3	4	5	7	8	10	12	14	16	20	22	24	26	28	29	31	32	33	35
36	1	3	4	5	7	8	10	12	14	17	20	23	25	27	29	30	32	33	34	36
37	1	3	4	5	7	8	10	12	14	17	21	24	26	28	30	31	33	34	35	37
38	1	3	4	5	7	8	10	12	15	18	21	24	27	29	31	32	34	35	36	38
39	1	3	4	5	7	8	10	12	15	18	22	25	28	30	32	33	35	36	37	39
40	1	3	4	5	7	9	11	13	16	19	22	25	28	30	32	34	36	37	38	40
41	1	3	4	5	7	9	11	13	16	19	23	26	29	31	33	35	37	38	39	41
42	1	3	4	5	7	9	11	14	17	20	23	26	29	32	34	36	38	39	40	42
43	1	3	4	5	7	9	11	14	17	20	24	27	30	33	35	37	39	40	41	43
44	1	3	4	5	7	9	12	15	18	21	24	27	30	33	36	38	40	41	42	44
45	1	3	4	5	7	9	12	15	18	21	25	28	31	34	37	39	41	42	43	45
46	1	3	4	5	8	10	13	16	19	22	25	28	31	34	37	39	42	43	44	46
47	1	3	4	5	8	10	13	16	19	22	26	29	32	35	38	40	43	44	45	47
48	1	3	4	5	8	11	14	17	20	23	26	29	32	35	38	41	44	45	46	48
49	1	3	4	5	8	11	14	17	20	23	27	30	33	36	39	42	45	46	47	49
50	1	3	4	5	8	11	14	17	20	24	27	31	34	37	40	43	46	47	48	50

Figure 8

Student/Teacher Ratio in Public and Nonpublic Elementary and Secondary Institutions

Institutions	Number of Teachers	Number of Students	Student/ Teacher Ratio
Public			
Elementary	858,000	27,692,000	32.3:1
Secondary	550,000	8,589,000	15.6:1
Total	1,408,000	36,281,000	25.8:1
Nonpublic			
Elementary	133,000	4,800,000	36.1:1
Secondary	59,000	1,100,000	18.6:1
Total	192,000	5,900,000	30.7:1
All Institutions			
Elementary	991,000	32,492,000	32.8:1
Secondary	609,000	9,689,000	15.9:1
Total	1,600,000	42,181,000	26.4:1

Source: U. S. Office of Education and National Education Association. "The Magnitude of the American Educational Establishment," Saturday Review (September 19, 1970), p. 67.

Table 10

Centroid Values for Area Sampled in Upper Portion of Normal Distribution and Correction Terms Derived for Use in Developing Tables for Estimates of Point Biserial Correlation Coefficients and Test Variance

A	z	$\dot{z}$	CV	$\dot{z}'$	CT
.1667	.9672	1.499128	.667054	1.878878	.532232
.2000	.8418	1.399600	.714489	1.754138	.570080
.2703	.6120	1.223865	.817083	1.533887	.651939
.3000	.5250	1.158612	.863101	1.452105	.688656
.3333	.4308	1.090876	.916694	1.367210	.731416
.3810	.3032	1.000000	1.000000	1.253314	.797885
.4000	.2533	.965861	1.035345	1.210527	.826087
.4333	.1680	.907807	1.101555	1.137767	.878914
.4667	.0835	.851842	1.173926	1.067626	.936657
.5000	.0000	.797885	1.253314	1.000000	1.000000

A = area of upper portion of normal distribution.

z = point on z-scale of lower limit of area sampled in upper portion of normal distribution.

$\dot{z}$ = theoretical centroid intersection point on z-scale for area sampled in upper portion of normal distribution.

CV = correction term for estimate of variance.

$\dot{z}'$ = predicted true centroid intersection point on z-scale for area sampled in upper portion of normal distribution.

CT = predicted value of upper mean for use in developing table of point biserial values.

Table 11

Conversion of Deviation Scores to Test Standard Deviation and Test Variance

Deviation	Standard Deviation	Variance	Deviation	Standard Deviation	Variance
1	.07	.01	36	2.66	7.08
2	.15	.02	37	2.73	7.48
3	.22	.05	38	2.81	7.89
4	.30	.09	39	2.88	8.31
5	.37	.14	40	2.96	8.74
6	.44	.20	41	3.03	9.18
7	.52	.27	42	3.10	9.63
8	.59	.35	43	3.18	10.10
9	.67	.44	44	3.25	10.57
10	.74	.55	45	3.33	11.06
11	.81	.66	46	3.40	11.56
12	.89	.79	47	3.47	12.06
13	.96	.92	48	3.55	12.58
14	1.03	1.07	49	3.62	13.11
15	1.11	1.23	50	3.70	13.65
16	1.18	1.40	51	3.77	14.20
17	1.26	1.58	52	3.84	14.77
18	1.33	1.77	53	3.92	15.34
19	1.40	1.97	54	3.99	15.92
20	1.48	2.18	55	4.06	16.52
21	1.55	2.41	56	4.14	17.13
22	1.63	2.64	57	4.21	17.74
23	1.70	2.89	58	4.29	18.37
24	1.77	3.15	59	4.36	19.01
25	1.85	3.41	60	4.43	19.66
26	1.92	3.69	61	4.51	20.32
27	2.00	3.98	62	4.58	20.99
28	2.07	4.28	63	4.66	21.68
29	2.14	4.59	64	4.73	22.37
30	2.22	4.92	65	4.80	23.07
31	2.29	5.25	66	4.88	23.79
32	2.36	5.59	67	4.95	24.52
33	2.44	5.95	68	5.03	25.25
34	2.51	6.31	69	5.10	26.00
35	2.59	6.69	70	5.17	26.76

Table 11 (continued)

Deviation	Standard Deviation	Variance	Deviation	Standard Deviation	Variance
71	5.25	27.53	106	7.83	61.36
72	5.32	28.31	107	7.91	62.53
73	5.39	29.10	108	7.98	63.70
74	5.47	29.91	109	8.06	64.88
75	5.54	30.72	110	8.13	66.08
76	5.62	31.54	111	8.20	67.29
77	5.69	32.38	112	8.28	68.51
78	5.76	33.23	113	8.35	69.73
79	5.84	34.08	114	8.42	70.97
80	5.91	34.95	115	8.50	72.22
81	5.99	35.83	116	8.57	73.49
82	6.06	36.72	117	8.65	74.76
83	6.13	37.62	118	8.72	76.04
84	6.21	38.53	119	8.79	77.34
85	6.28	39.46	120	8.87	78.64
86	6.36	40.39	121	8.94	79.96
87	6.43	41.34	122	9.02	81.28
88	6.50	42.29	123	9.09	82.62
89	6.58	43.26	124	9.16	83.97
90	6.65	44.24	125	9.24	85.33
91	6.72	45.22	126	9.31	86.70
92	6.80	46.22	127	9.39	88.08
93	6.87	47.23	128	9.46	89.48
94	6.95	48.26	129	9.53	90.88
95	7.02	49.29	130	9.61	92.29
96	7.09	50.33	131	9.68	93.72
97	7.17	51.38	132	9.75	95.16
98	7.24	52.45	133	9.83	96.60
99	7.32	53.53	134	9.90	98.06
100	7.39	54.61	135	9.98	99.53
101	7.46	55.71	136	10.05	101.01
102	7.54	56.82	137	10.12	102.50
103	7.61	57.94	138	10.20	104.00
104	7.69	59.07	139	10.27	105.52
105	7.76	60.21	140	10.35	107.04

Table 12

Mean Proportion of Variance (MPV)

Test Variance	Test Mean in Proportions										
	.30	.35	.40	.45	.50	.55	.60	.65	.70	.75	.80
6	.035	.038	.040	.041	.042	.041	.040	.038	.035	.031	.027
7	.030	.033	.034	.035	.036	.035	.034	.033	.030	.027	.023
8	.026	.028	.030	.031	.031	.031	.030	.028	.026	.023	.020
9	.023	.025	.027	.028	.028	.028	.027	.025	.023	.021	.018
10	.021	.023	.024	.025	.025	.025	.024	.023	.021	.019	.016
11	.019	.021	.022	.023	.023	.023	.022	.021	.019	.017	.015
12	.018	.019	.020	.021	.021	.021	.020	.019	.018	.016	.013
13	.016	.018	.019	.019	.019	.019	.019	.018	.016	.014	.012
14	.015	.016	.017	.018	.018	.018	.017	.016	.015	.013	.011
15	.014	.015	.016	.017	.017	.017	.016	.015	.014	.013	.011
16	.013	.014	.015	.016	.016	.016	.015	.014	.013	.012	.010
17	.012	.013	.014	.015	.015	.015	.014	.013	.012	.011	.009
18	.012	.013	.013	.014	.014	.014	.013	.013	.012	.010	.009
19	.011	.012	.013	.013	.013	.013	.013	.012	.011	.010	.008
20	.010	.011	.012	.012	.013	.012	.012	.011	.010	.009	.008
21	.010	.011	.011	.012	.012	.012	.011	.011	.010	.009	.008
22	.010	.010	.011	.011	.011	.011	.011	.010	.010	.009	.007
23	.009	.010	.010	.011	.011	.011	.010	.010	.009	.008	.007
24	.009	.010	.010	.010	.010	.010	.010	.010	.009	.008	.007
25	.008	.009	.010	.010	.010	.010	.010	.009	.008	.008	.006
26	.008	.009	.009	.010	.010	.010	.009	.009	.008	.007	.006
27	.008	.008	.009	.009	.009	.009	.009	.008	.008	.007	.006
28	.008	.008	.009	.009	.009	.009	.009	.008	.008	.007	.006
29	.007	.008	.008	.008	.009	.009	.008	.008	.007	.007	.006
30	.007	.008	.008	.008	.008	.008	.008	.008	.007	.006	.005
35	.006	.007	.007	.007	.007	.007	.007	.007	.006	.005	.004
40	.005	.006	.006	.006	.006	.006	.006	.006	.005	.005	.004
45	.005	.005	.005	.006	.006	.006	.005	.005	.005	.004	.004
50	.004	.005	.005	.005	.005	.005	.005	.005	.004	.004	.003
60	.004	.004	.004	.004	.004	.004	.004	.004	.004	.003	.003
70	.003	.003	.003	.004	.004	.004	.003	.003	.003	.003	.002
80	.003	.003	.003	.003	.003	.003	.003	.003	.003	.002	.002
90	.002	.003	.003	.003	.003	.003	.003	.003	.002	.002	.002
100	.002	.002	.002	.002	.003	.002	.002	.002	.002	.002	.002
150	.001	.002	.002	.002	.002	.002	.002	.002	.001	.001	.001
200	.001	.001	.001	.001	.001	.001	.001	.001	.001	.001	.001

Table 13

Item Discrimination Index

Number in Lower Sample Getting Item Correct	Number in Upper Sample Getting Item Correct										
	0	1	2	3	4	5	6	7	8	9	10
0	.00	.18	.27	.34	.40	.46	.52	.59	.65	.72	.80
1	-.18	.00	.11	.20	.28	.35	.42	.49	.56	.64	.72
2	-.27	-.11	.00	.09	.17	.25	.33	.40	.48	.56	.65
3	-.34	-.20	-.09	.00	.08	.16	.24	.32	.40	.49	.59
4	-.40	-.28	-.17	-.08	.00	.08	.16	.24	.33	.42	.52
5	-.46	-.35	-.25	-.16	-.08	.00	.08	.16	.25	.35	.46
6	-.52	-.42	-.33	-.24	-.16	-.08	.00	.08	.17	.28	.40
7	-.59	-.49	-.40	-.32	-.24	-.16	-.08	.00	.09	.20	.34
8	-.65	-.56	-.48	-.40	-.33	-.25	-.17	-.09	.00	.11	.27
9	-.72	-.64	-.56	-.49	-.42	-.35	-.28	-.20	-.11	.00	.18
10	-.80	-.72	-.65	-.59	-.52	-.46	-.40	-.34	-.27	-.18	.00

Table 14

Conversion of Item Discrimination Index to Item Probability Index

Discrimination Index		Probability Index	Discrimination Index		Probability Index
From	To		From	To	
-1.00	-.57	.00	.01		.52
-.56	-.49	.01	.02		.54
-.48	-.45	.02	.03		.55
-.44	-.42	.03	.04		.57
-.41	-.39	.04	.05		.59
-.38	-.37	.05	.06		.60
-.36	-.35	.06	.07		.62
-.34		.07	.08		.63
-.33	-.32	.08	.09		.65
-.31		.09	.10		.66
-.30		.10	.11		.68
-.29		.11	.12		.69
-.28	-.27	.12	.13		.70
-.26		.13	.14		.72
-.25		.14	.15		.73
-.24		.15	.16		.75
-.23		.16	.17		.76
-.22		.18	.18		.77
-.21		.19	.19		.79
-.20		.20	.20		.80
-.19		.21	.21		.81
-.18		.22	.22		.82
-.17		.24	.23		.84
-.16		.25	.24		.85
-.15		.27	.25		.86
-.14		.28	.26		.87
-.13		.30	.27	.28	.88
-.12		.31	.29		.89
-.11		.32	.30		.90
-.10		.34	.31		.91
-.09		.35	.32	.33	.92
-.08		.37	.34		.93
-.07		.38	.35	.36	.94
-.06		.40	.37	.38	.95
-.05		.41	.39	.41	.96
-.04		.43	.42	.43	.97
-.03		.45	.44	.48	.98
-.02		.47	.49	.55	.99
-.01		.48	.56	1.00	1.00
.00		.50			

Table 15

Deviation from Actual Values for Test Mean Estimates
Using Different Sampling Patterns

Sampling Pattern	Sample Distributions										
	1	2	3	4	5	6	7	8	9	10	*$\overline{AD}$
A	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
B	.25	.15	-.01	.11	.14	.25	-.08	.06	-.16	-.16	.14
C	.38	.26	.08	.13	.17	.22	-.14	.04	-.20	-.23	.18
D	.47	.40	.18	.15	.27	.17	-.22	.02	-.25	-.32	.24
E	.73	.71	.41	.28	.41	.11	-.37	.02	-.32	-.43	.38
F	1.17	1.30	1.03	.80	.17	.17	-.27	-.23	-.80	-.87	.68
G	.07	-.30	-.17	.10	.07	-.03	.13	-.03	.10	-.07	.11
H	.09	.07	-.05	-.02	-.03	.42	-.12	.02	.05	-.12	.10
J	-.04	-.03	-.04	-.03	-.03	.17	-.01	-.02	.16	-.01	.05
K	.24	.11	-.04	.19	.11	-.04	.06	.06	-.27	-.08	.12
L	.11	.03	.03	.15	.03	.14	.01	.02	-.14	.01	.07
M	.37	.50	.43	.30	.17	-.23	-.17	-.13	-.60	-.27	.32
N	1.07	.90	.53	.60	.27	.07	.17	-.23	-.70	-.77	.53
P	.34	.23	.20	.29	-.03	-.05	-.04	-.05	-.26	-.15	.16
Q	.02	.05	.03	.05	.12	.02	-.07	.02	-.15	-.02	.06
R	.32	.35	.33	.25	.02	.07	-.07	-.03	-.25	-.17	.19
S	.42	.40	.23	.25	.10	.07	-.12	.02	-.25	-.22	.21
T	.31	.31	.26	.22	.06	-.01	-.14	.00	-.15	-.14	.16

*Mean absolute deviation from actual values.

Table 16

Deviation from Actual Values for Test Standard Deviation Estimates Using Different Sampling Patterns

Sampling Pattern	Sample Distributions										*$\overline{AD}$
	1	2	3	4	5	6	7	8	9	10	
A	-.32	-.29	.10	-.14	.44	-.01	.09	.16	.11	-.21	.19
B	-.31	-.32	-.01	-.23	.29	.06	-.07	.24	.06	-.28	.19
C	-.19	-.26	.03	-.12	.41	.14	.04	.10	.08	-.20	.16
D	-.30	-.33	-.12	-.26	.14	-.03	-.13	-.01	-.07	-.22	.16
E	-.30	-.27	-.13	-.10	-.03	-.02	-.06	-.03	-.10	-.17	.12
F	-.28	-.14	-.24	-.08	-.06	-.12	-.23	-.07	-.24	-.07	.15
G	-.37	-.61	-.35	-.45	.10	-.82	-.63	-.05	-.26	-.36	.40
H	-.57	-.49	-.23	-.70	-.13	-.38	-.72	-.27	-.24	-.48	.42
J	-.58	-.49	-.17	-.67	-.05	-.46	-.63	-.19	-.26	-.46	.40
K	-.27	-.33	.05	.04	.42	.07	.34	.12	-.15	-.19	.20
L	-.21	-.27	.13	.18	.50	.04	.39	.26	.22	-.13	.23
M	.32	-.28	-.11	-.02	-.10	.01	.00	.09	.32	-.07	.13
N	-.17	-.27	-.27	.07	-.09	-.15	-.07	.03	-.08	-.03	.12
P	-.36	-.52	-.39	-.34	-.39	-.66	-.44	-.20	-.33	-.30	.39
Q	-.22	-.14	.28	.08	.71	.21	.39	.26	.17	-.15	.26
R	-.37	-.33	-.21	-.27	-.25	-.39	-.30	-.11	-.23	-.27	.27
S	-.17	-.18	.09	.13	.35	.21	.25	.14	.07	-.12	.17
T	-.42	-.40	-.35	-.30	-.23	-.36	-.32	-.14	-.24	-.29	.30

*Mean absolute deviation from actual values.

Table 17

Deviation from Actual Values for Test Variance Estimates Using Different Sampling Patterns

Sampling Pattern	Sample Distributions										*$\overline{AD}$
	1	2	3	4	5	6	7	8	9	10	
A	-2.09	-1.71	.83	-1.55	5.47	-.12	1.18	1.24	.84	-.99	1.60
B	-2.03	-1.87	-.08	-2.53	3.56	.72	-.91	1.88	.46	-1.30	1.53
C	-1.27	-1.54	.17	-1.33	5.09	1.70	.52	.77	.61	-.95	1.40
D	-1.97	-1.93	-.97	-2.85	1.70	-.36	-1.67	-.08	-.53	-1.04	1.31
E	-1.97	-1.59	-1.05	-1.11	-.36	-.24	-.78	-.23	-.65	-.81	.88
F	-1.84	-.85	-1.91	-.89	-5.31	-1.42	-2.94	-.53	-1.76	-.34	1.78
G	-2.40	-3.39	-2.75	-4.85	1.21	-9.15	-7.69	-.38	-1.90	-1.65	3.54
H	-3.58	-2.78	-1.84	-7.37	-1.54	-4.41	-8.84	-1.99	-1.76	-2.14	3.62
J	-3.63	-2.78	-1.37	-7.08	-.60	-5.30	-7.79	-1.42	-1.90	-2.06	3.39
K	-1.78	-1.93	.42	.46	5.22	.84	4.54	.92	1.15	-.90	1.82
L	-1.40	-1.59	1.09	2.06	6.25	.48	5.22	2.04	1.71	-.62	2.25
M	-2.09	-1.65	-.89	-.22	-1.19	.12	.00	.69	2.52	-.34	.97
N	-1.14	-1.59	-2.14	.80	-1.07	-1.77	-.91	.23	-.60	-.15	1.04
P	-2.34	-2.94	-3.05	-3.70	-4.53	-7.47	-5.53	-1.49	-2.39	-1.39	3.48
Q	-1.46	-.85	2.38	.91	9.02	2.76	5.22	2.04	1.31	-.72	2.67
R	-2.40	-1.93	-1.68	-2.96	-2.94	-4.52	-3.69	-.83	-1.69	-1.26	2.39
S	-1.14	-1.08	.75	1.48	4.32	2.56	3.31	1.08	.53	-.58	1.68
T	-2.70	-2.31	-2.75	-3.28	-2.71	-4.18	-4.06	-1.05	-1.76	-1.35	2.62

*Mean absolute deviation from actual values.

Table 18

Deviation from Actual Values for Test Reliability Coefficient (K-R 21) Estimates Using Different Sampling Patterns

Sampling Pattern	Sample Distributions										*$\overline{AD}$
	1	2	3	4	5	6	7	8	9	10	
A	-.16	-.18	.04	-.02	.02	-.02	-.02	.09	.03	-.06	.06
B	-.16	-.18	-.01	-.07	.02	-.02	-.02	.09	.03	-.06	.07
C	-.16	-.18	-.01	-.02	.02	.03	-.02	.04	.03	-.06	.06
D	-.16	-.18	-.06	-.07	.02	-.02	-.02	-.01	-.02	-.06	.06
E	-.16	-.18	-.01	-.02	-.03	-.02	-.02	-.01	-.02	-.06	.05
F	-.16	-.03	-.06	-.02	-.08	-.02	-.02	-.01	-.12	.09	.06
G	-.31	-.68	-.16	-.07	-.03	-.12	-.07	-.01	-.17	-.06	.17
H	-.46	-.43	-.11	-.12	-.03	-.07	-.07	-.11	-.12	-.06	.16
J	-.46	-.43	-.06	-.12	-.03	-.07	-.07	-.11	-.17	-.06	.16
K	-.16	-.18	-.01	-.02	-.02	-.02	-.02	.04	.03	-.06	.06
L	-.16	-.18	.04	.03	.02	-.02	-.02	.14	.08	-.06	.08
M	-.16	-.18	-.01	-.02	-.03	-.02	-.02	.04	.13	-.06	.07
N	-.06	-.18	-.06	-.02	-.03	-.02	-.02	..04	-.02	-.09	.05
P	-.31	-.43	-.16	-.07	-.08	-.07	-.07	-.11	-.17	-.06	.15
Q	-.16	-.03	.04	-.02	.07	.03	-.02	.14	.08	-.06	.06
R	-.31	-.18	-.06	-.07	-.03	-.07	-.02	-.01	-.12	-.06	.09
S	-.06	-.18	-.04	.03	.02	.03	-.02	.09	.03	-.06	.06
T	-.31	-.43	-.16	-.07	-.03	-.07	-.02	-.11	-.12	-.06	.14
GN	-.06	-.18	-.11	-.02	-.03	-.02	-.02	.04	-.02	-.06	.06

*Mean absolute deviation from actual values.

Table 19

Deviation from Actual Values for Standard Score Estimates Using Different Sampling Patterns

Sampling Pattern	Sample Raw Scores										*$\overline{AD}$
	1	2	3	4	5	6	7	8	9	10	
A	-1	0	0	0	-1	-1	0	0	-1	0	.40
B	-1	0	0	0	-1	-1	0	0	-1	0	.40
C	-1	0	0	0	-1	-1	0	0	-1	0	.40
D	-1	0	0	0	-1	-1	0	0	-1	0	.40
E	-1	0	0	0	-1	-1	0	0	-1	0	.40
F	-1	0	0	0	-1	-1	0	0	-1	0	.40
G	-1	0	0	0	-1	-1	0	0	-1	0	.40
H	1	2	1	1	1	1	1	1	1	1	1.10
J	1	2	1	1	1	1	1	1	1	1	1.10
K	-1	0	0	0	-1	-1	0	0	-1	0	.40
L	-1	0	0	0	-1	-1	0	0	-1	0	.40
M	-1	0	0	0	-1	-1	0	0	-1	0	.40
N	-1	0	0	0	-1	-1	0	0	-1	0	.40
P	1	2	1	1	1	1	1	1	1	1	1.10
Q	-3	-1	-1	-1	0	0	0	0	1	2	.90
R	-1	0	0	0	-1	-1	0	0	-1	0	.40
S	-1	0	0	0	-1	-1	0	0	-1	0	.40
T	-1	0	0	0	-1	-1	0	0	-1	0	.40
GN	-1	0	0	0	-1	-1	0	0	-1	0	.40

*Mean absolute deviation from actual values.

Table 20

Deviation from Actual Values for Item Discrimination Index Estimates Using Different Sampling Patterns

Sampling Pattern	Sample Multiple Choice Items										*$\overline{AD}$
	1	2	3	4	5	6	7	8	9	10	
A	.08	-.19	.16	-.12	.05	.03	-.11	-.03	-.16	-.25	.12
B	.04	.01	.00	-.16	.01	.00	-.07	-.07	.03	-.09	.05
C	.08	.09	-.04	-.08	-.03	.00	-.02	-.03	.00	.05	.04
D	.04	.06	-.08	-.08	-.07	-.03	.06	-.03	-.03	.08	.06
E	.01	.17	.00	.04	-.08	.07	.17	-.14	-.03	.11	.08
F	.13	-.03	-.11	-.08	.03	-.05	.06	-.19	.03	.14	.08
G	-.19	-.03	.05	.09	-.07	-.12	.22	-.19	-.16	-.34	.15
H	-.01	-.36	.18	-.36	.22	.24	.13	.21	-.26	.05	.20
J	-.05	-.48	.21	-.24	.10	.11	.09	.21	-.35	-.17	.20
K	.03	.08	-.02	.10	-.12	-.11	-.25	-.13	.20	-.25	.13
L	.16	.06	.07	-.03	-.03	-.03	-.39	-.21	.04	-.23	.12
M	.46	.16	-.01	-.24	.24	-.12	-.28	-.05	-.16	-.34	.21
N	-.04	.29	-.28	.24	.03	-.05	.41	-.31	.03	.01	.17
P	.21	.02	.10	-.08	-.03	-.07	-.24	-.17	-.05	-.20	.12
Q	.12	-.03	.07	.00	.04	.20	-.11	-.21	-.16	-.16	.11
R	.21	-.11	.14	-.16	-.07	-.03	-.28	-.14	-.09	-.16	.14
S	.12	-.03	.07	-.08	-.32	-.03	-.20	.04	.00	-.09	.10
T	.20	-.07	.10	-.16	-.03	.00	-.24	.00	-.13	-.13	.11

*Mean absolute deviation from actual values.

Table 21

Deviation from Actual Values for Item Probability Index Estimates Using Different Sampling Patterns

Sampling Pattern	Sample Multiple Choice Items										*$\overline{AD}$
	1	2	3	4	5	6	7	8	9	10	
A	.07	-.17	.05	-.17	.04	.01	-.22	-.03	-.04	-.08	.09
B	.03	.00	.00	-.24	.01	.00	-.13	-.07	.00	-.01	.05
C	.07	.01	-.04	-.10	-.03	.00	-.04	-.03	.00	.00	.03
D	.03	.01	-.08	-.10	-.09	.00	.09	-.03	.00	.00	.04
E	.00	.02	.00	.04	-.10	.01	.22	-.10	.00	.00	.05
F	.15	-.02	-.13	-.10	.03	.00	.09	-.12	.00	.00	.06
G	-.06	-.02	.02	-.06	-.09	-.03	.24	-.12	-.04	-.19	.09
H	-.01	-.48	.06	-.63	.10	.01	.18	.37	-.11	.00	.20
J	-.03	-.71	.06	-.40	.07	.01	.14	.37	-.23	-.03	.20
K	.02	.01	-.02	.07	-.17	-.02	-.48	-.10	.00	-.08	.10
L	.19	.01	.03	-.03	-.03	.00	-.66	-.12	.00	-.06	.11
M	.75	.02	-.01	-.40	.10	-.03	-.53	.08	-.04	-.19	.22
N	-.02	.02	-.44	.10	.03	.00	.28	-.13	.00	.00	.10
P	.28	.00	.04	-.10	-.03	-.01	-.47	-.11	-.01	-.05	.11
Q	.13	-.02	.03	.00	.04	.01	-.22	-.12	-.04	-.03	.06
R	.28	-.07	.05	-.24	-.09	.00	-.53	-.10	-.01	-.03	.14
S	.13	-.02	.03	-.10	-.56	.00	-.40	.06	.00	-.01	.13
T	.26	-.04	.04	-.24	-.03	.00	-.47	.00	-.02	-.02	.11

*Mean absolute deviation from actual values.

Table 22

Deviation from Actual Values for Item Difficulty Index Estimates Using Different Sampling Patterns

Sampling Pattern	Sample Multiple Choice Items										*$\overline{AD}$
	1	2	3	4	5	6	7	8	9	10	
A	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
B	.00	-.01	-.05	.01	-.04	-.05	.05	-.02	.11	.03	.04
C	-.01	.01	-.07	.02	-.05	-.07	-.01	-.01	.13	.02	.04
D	.02	.02	-.10	-.03	-.07	-.08	-.03	-.05	.15	.05	.06
E	.11	-.07	-.05	.03	-.18	-.12	-.08	.09	.15	.09	.10
F	.17	-.03	.00	-.03	-.27	-.23	-.03	.20	.10	.10	.12
G	.17	-.03	.10	.07	-.07	-.03	.07	.20	.00	.00	.07
H	.00	.04	-.05	-.03	-.12	-.01	.17	-.10	.32	.03	.09
J	.03	.01	-.01	-.10	-.01	.07	.10	-.03	.19	.06	.06
K	-.04	-.06	-.01	.11	-.01	-.07	-.12	.04	-.17	-.01	.06
L	-.02	-.01	.01	.09	.01	-.05	-.08	.02	-.16	-.05	.05
M	-.03	.07	-.20	.07	.13	-.03	-.03	-.10	.00	.00	.07
N	.27	-.03	-.10	-.03	-.27	-.23	-.03	.10	.10	.00	.12
P	.06	.04	.03	.03	-.04	-.10	-.05	.07	-.10	-.02	.05
Q	-.03	-.03	.00	.02	.13	-.03	-.03	.05	.00	.00	.03
R	.02	.02	.05	.02	-.07	-.08	-.03	.10	-.05	.00	.04
S	-.03	-.03	.00	.07	.08	-.13	-.08	.10	.00	.05	.06
T	-.01	.00	.02	.06	-.05	-.07	-.06	.08	-.01	.02	.04

*Mean absolute deviation from actual values.

Table 23

Deviation from Actual Values for Test Mean Estimates Using Sampling Pattern "G" with Different Class Sizes

Subject	Class Sizes										*$\overline{AD}$
	20	21	23	25	27	30	33	38	50	60	
Deviation	.10	-.04	.23	-.40	-.33	.20	.08	.27	-.08	-.07	.18

*Mean absolute deviation from actual values.

Table 24

Deviation from Actual Values for Test Standard Deviation Estimates Using Sampling Pattern "N" with Different Class Sizes

Subject	Class Sizes										*$\overline{AD}$
	20	21	23	25	27	30	33	38	50	60	
Deviation	-.31	-.19	-.25	-.27	-.20	-.23	-.08	.07	.38	.33	.23

*Mean absolute deviation from actual values.

Table 25

Deviation from Actual Values for Test Variance Estimates Using Sampling Pattern "N" with Different Class Sizes

Subject	Class Sizes										*$\overline{AD}$
	20	21	23	25	27	30	33	38	50	60	
Deviation	-2.48	-1.42	-1.87	-2.06	-1.83	-2.03	-.76	.75	3.53	3.36	2.01

*Mean absolute deviation from actual values.

Table 26

Deviation from Actual Values for Test Reliability Coefficient (K-R 21) Estimates Using Sampling Patterns "G" and "N" with Different Class Sizes

Subject	Class Sizes										*$\overline{AD}$
	20	21	23	25	27	30	33	38	50	60	
Deviation	-.12	-.13	-.15	-.18	-.07	-.08	.01	-.01	.06	.07	.09

*Mean absolute deviation from actual values.

Table 27

Deviation from Actual Values for Standard Scores Estimates Using Sampling Patterns "G" and "N" with Different Class Sizes

Raw Scores	Class Sizes									
	20	21	23	25	27	30	33	38	50	60
1	1	1	-2	2	0	1	-1	2	-2	1
2	1	0	-2	2	0	2	-1	2	-1	1
3	1	0	-1	2	0	0	-1	1	-1	1
4	1	0	0	1	1	1	-1	1	-1	1
5	0	0	0	2	1	1	-1	1	-1	1
6	0	1	-1	1	1	0	-1	1	-1	1
7	0	1	-1	1	1	0	-1	1	-1	1
8	0	1	-1	2	1	0	-1	1	0	1
9	0	0	-1	2	1	0	-1	1	0	1
10	0	0	-1	2	1	0	-1	1	0	1
11	0	1	-1	2	1	-1	-1	1	0	1
12	0	1	0	3	1	-1	-1	0	0	1
13	0	1	0	3	1	-1	0	0	0	1
14	0	1	-1	3	1	-1	0	0	0	1
15	-1	1	-1	2	2	-2	0	0	0	1
16	0	1	0	2	2	-2	0	0	1	1
17	0	1	0	2	2	-2	0	0	1	1
18	0	2	0	2	2	-2	0	0	2	1
19	0	2	0	2	2	-2	0	0	2	1
20	-1	2	0	2	3	-2	0	-1	2	2
*$\overline{AD}$	.30	.85	.65	2.00	1.20	1.05	.60	.70	.80	1.05

*Mean absolute deviation from actual values.

Table 28

Deviation from Actual Values for Item Discrimination Index Estimates Using Sampling Pattern "D" with Different Class Sizes

Sample Items	Class Sizes									
	20	21	23	25	27	30	33	38	50	60
1	-.09	-.15	-.06	-.06	.00	.00	-.02	.03	.04	.07
2	.01	-.03	.26	-.12	.10	-.02	.01	-.23	.23	.14
3	-.18	-.07	-.07	.13	.05	.06	-.06	.05	.05	-.13
4	-.10	.04	-.02	.01	-.04	-.05	-.09	.08	.07	.10
5	-.19	.07	-.11	-.43	-.09	.01	-.15	-.03	.15	.03
6	-.15	-.15	-.14	-.28	.06	-.08	.14	-.13	.04	.22
7	.36	-.18	-.18	-.04	-.06	.00	-.14	.11	.00	.17
8	-.01	-.19	-.11	.06	-.04	-.25	-.07	-.08	.09	-.15
9	.00	.12	-.06	-.06	-.13	-.05	.07	.09	-.06	.04
10	-.12	-.14	.05	.19	.13	-.10	-.10	.00	.04	-.07
11	-.15	.01	-.13	.16	.07	.09	-.05	.00	.04	-.09
12	-.01	-.12	-.02	.00	-.05	-.06	-.05	.05	.05	-.17
13	-.06	-.07	-.13	-.12	.06	-.06	.02	-.06	.10	.14
14	-.29	.14	-.07	-.01	-.11	-.01	.04	-.11	-.13	.15
15	-.08	-.13	-.21	-.17	-.05	-.07	-.03	-.02	-.15	-.07
16	-.01	.08	.01	-.08	-.16	.06	.12	.09	.15	.06
17	-.10	-.07	-.13	.03	-.17	.13	-.04	-.08	.02	.17
18	-.06	-.09	-.06	-.01	.12	-.15	.08	-.10	-.02	.00
19	-.15	-.02	.15	-.04	.09	-.12	.05	.16	.04	.13
20	.04	-.05	.03	-.03	-.14	-.05	.10	-.17	-.01	-.07
*$\overline{AD}$	.11	.10	.10	.10	.09	.07	.07	.08	.07	.11

*Mean absolute deviation from actual values.

Table 29

Deviation from Actual Values for Item Probability Index Estimates Using Sampling Pattern "D" with Different Class Sizes

Sample Items	Class Sizes									
	20	21	23	25	27	30	33	38	50	60
1	-.15	-.19	-.07	-.11	.00	.00	-.05	.07	.02	.20
2	.01	-.05	.42	-.14	.17	-.05	-.05	-.42	.03	-.01
3	-.15	-.11	-.11	.23	.00	.01	-.14	.01	-.06	-.34
4	-.17	.05	-.04	.00	-.09	-.11	-.10	-.02	-.02	-.03
5	-.25	.11	-.18	-.23	-.09	.05	-.21	-.09	.09	-.05
6	-.11	-.11	-.01	-.29	.09	-.15	.03	-.04	-.04	.04
7	.44	-.06	-.23	-.04	-.05	-.06	-.08	.11	-.08	-.01
8	-.02	-.23	-.10	-.01	.00	-.41	-.15	-.19	.00	-.22
9	.00	.19	-.01	-.02	-.06	-.12	.15	-.01	-.16	-.09
10	-.12	-.19	.06	.33	.23	-.12	-.13	-.05	-.06	-.12
11	-.20	.01	-.20	.28	.07	.01	-.10	-.06	-.08	-.25
12	-.01	-.14	-.03	.00	-.09	-.13	-.13	-.03	-.07	.06
13	-.04	-.12	-.22	-.20	.00	-.13	-.01	-.23	-.02	.00
14	-.39	.13	-.07	-.05	-.20	-.02	.00	-.17	-.30	-.01
15	-.11	-.19	-.33	-.06	-.11	-.05	-.05	-.09	-.31	-.20
16	-.02	.12	.00	-.13	-.28	-.01	.15	-.01	.02	-.04
17	-.12	-.10	-.20	.03	-.27	.19	.01	-.02	-.04	.00
18	-.07	-.07	-.09	.00	.08	-.17	-.01	-.22	-.11	-.01
19	-.14	-.01	.19	.00	.15	-.22	-.02	.07	-.04	-.02
20	.07	-.07	.00	-.06	-.07	-.03	.21	-.19	-.11	-.18
*AD	.13	.11	.13	.11	.10	.10	.09	.10	.08	.09

*Mean absolute deviation from actual values.

Table 30

Deviation from Actual Values for Item Difficulty Index Estimates Using Sampling Pattern "D" with Different Class Sizes

Sample Items	Class Sizes									
	20	21	23	25	27	30	33	38	50	60
1	.00	.00	-.01	-.08	.00	.00	.03	.03	-.03	.08
2	.00	.04	-.03	.07	.01	.00	.02	-.05	-.01	.05
3	.00	.01	-.01	.00	.00	.00	-.02	.05	.00	.07
4	.00	.03	.00	-.03	-.03	-.01	-.06	-.06	-.03	-.02
5	.00	.04	.04	-.01	.08	.05	-.03	.01	.06	.02
6	.00	-.02	.04	.05	-.02	-.05	-.04	.01	.00	.07
7	.00	-.01	-.02	.01	.04	.00	.11	.00	.03	.02
8	.00	-.01	-.06	-.01	.03	.03	-.03	-.02	.11	.00
9	.00	-.02	.00	-.01	.00	.02	-.10	.09	-.06	-.08
10	.00	-.04	-.07	.07	-.02	-.08	.00	-.11	.08	-.05
11	.00	.04	.04	-.05	.03	-.05	.00	.02	-.07	-.25
12	.00	.02	.03	.00	.03	-.05	.03	.09	-.07	-.10
13	.00	-.03	.04	.03	-.06	.00	-.06	-.02	.03	-.05
14	.00	-.03	-.03	-.01	.03	.05	.03	.09	.03	-.08
15	.00	.02	-.02	.06	-.07	.00	.00	.01	-.13	.05
16	.00	-.02	-.04	-.05	-.09	.08	.00	-.11	.02	-.08
17	.00	.04	.04	.03	-.05	-.07	.05	-.07	.02	-.12
18	.00	-.01	.02	-.01	.01	.00	-.01	-.04	-.11	-.08
19	.00	.03	-.01	-.03	.07	-.05	-.04	.05	-.04	-.13
20	.00	-.01	.04	.04	-.04	-.08	.05	-.02	.09	-.03
*$\overline{AD}$	.00	.02	.03	.03	.04	.03	.04	.05	.05	.07

*Mean absolute deviation from actual values.

APPENDIX B: Measurement Analysis Model

Figure 9

INSTRUCTIONS

Acceptability of Decision Points. If we are to institute individualized instruction in a nongraded structure, then we must be able to measure individual differences. In this research study, you are asked to make decisions on the probability of selected multiple choice items being able to measure differences in level and extent of progress in the trait being measured. A sample pretest and actual student responses have been furnished for your necessary information. A multiple choice item involves a decision point and probability is defined as the proportion of chances a specific item or decision point has of measuring differences in students when used in similar circumstances.

DIRECTIONS: Using the MEASUREMENT ANALYSIS MODEL,* record on the answer sheet provided the probability that the selected items will measure differences in extent and level of progress in the trait being measured.

EXAMPLE:	DECISION POINT	SELECTED ITEM	PROBABILITY INDEX
	0	2	.74

This means that there are 74 chances out of 100 possible chances that Item 2 will measure differences in level and extent of progress in the trait being measured when given to a similar group. Using this same procedure, record the probabilities that the other selected items will measure differences.

PLEASE WORK INDEPENDENTLY. YOUR PERSONAL RESPONSE IS NEEDED TO VALIDATE THIS RESEARCH STUDY.

*For subjects not using the Measurement Analysis Model, this statement was changed to read, "your BEST JUDGMENT."

Figure 10

ANSWER SHEET NO. _____

DECISION POINT	SELECTED ITEMS	PROBABILITY INDEX
0	2	.74
1	3	
2	5	
3	7	
4	9	
5	11	
6	12	
7	13	
8	14	
9	15	
10	19	
11	21	
12	23	
13	24	
14	27	
15	32	
16	35	
17	38	
18	39	
19	49	
20	50	

YOUR COOPERATION IS VERY MUCH APPRECIATED!

Figure 11

Student Responses to Sample
Fifty-item Pretest

Student	Responses to Selected Items																				Total Score
	2	3	5	7	9	11	12	13	14	15	19	21	23	24	27	32	35	38	39	49	50	
154	1	0	0	1	1	0	0	0	1	1	1	1	1	1	1	1	1	0	0	1	0	33
49	1	0	1	1	1	0	1	0	0	0	0	1	0	1	1	1	1	0	1	1	0	27
66	1	1	0	0	1	0	0	1	1	0	0	1	1	1	1	0	1	1	0	1	0	27
159	1	0	1	0	1	0	1	0	0	1	0	1	1	0	1	1	1	1	0	0	0	27
61	1	1	1	0	1	0	0	0	0	0	1	1	0	1	0	0	1	1	0	1	0	27
83	1	0	1	0	1	1	0	0	0	0	0	0	1	1	0	0	1	1	1	1	1	26
12	1	0	1	0	1	1	0	0	0	0	1	1	0	1	0	1	1	0	1	0	1	25
17	1	0	0	1	1	0	1	1	1	0	0	0	0	1	1	0	1	0	0	1	1	25
36	0	0	0	0	1	0	0	0	0	1	1	1	0	1	1	0	1	0	0	1	0	24
60	1	0	0	1	1	0	0	0	0	0	0	1	0	1	1	0	1	1	1	1	1	24
45	1	1	0	1	1	0	0	0	0	0	0	0	0	1	1	0	1	0	0	0	0	24
50	1	0	1	1	0	0	0	0	0	0	1	1	0	1	1	0	1	0	0	1	1	23
88	1	0	1	0	1	0	0	1	0	0	0	0	1	1	1	1	1	0	0	1	0	23
129	1	0	0	1	1	0	1	1	0	0	0	1	0	1	1	0	1	1	0	1	1	23
85	1	1	0	0	1	1	0	0	1	0	0	0	0	1	1	0	1	0	0	1	1	23
121	1	1	0	1	0	0	0	1	1	0	0	0	0	1	0	1	1	0	0	1	1	22
30	1	1	1	0	1	0	0	0	1	0	0	1	0	0	0	0	1	1	0	0	1	22
64	0	0	0	1	1	0	1	0	1	0	1	0	0	1	1	0	1	0	0	1	1	21
63	1	1	0	1	1	0	0	0	1	0	0	1	0	0	1	0	1	0	0	0	0	21
141	0	0	0	0	1	0	1	0	1	0	0	1	0	1	1	1	1	0	0	0	0	21
169	1	0	0	0	1	0	0	0	0	0	0	1	0	0	1	0	1	0	1	1	1	21
117	1	1	1	1	1	1	0	0	0	0	1	0	0	1	1	0	1	0	0	1	1	20
130	1	1	0	0	1	0	1	0	0	0	1	0	0	0	0	0	1	0	0	1	1	20
114	1	0	0	1	1	1	0	0	0	0	0	0	0	1	0	0	1	0	1	1	1	20
157	1	0	0	1	1	1	0	0	1	0	0	1	1	1	0	0	0	1	1	0	1	20
128	1	0	0	0	0	0	1	0	1	0	0	1	0	0	1	1	0	0	0	0	1	20
148	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	1	0	18

Note: "1" indicates success in meeting requirements of decision point or item and "0" indicates nonsuccess.

MEASUREMENT ANALYSIS MODEL

PART I: PROGRAMMED ITEM ANALYSIS

Ray Hassett

University of Houston

JUNE 1970

TABLE OF CONTENTS

INTRODUCTION

Measurement and evaluation are widely misinterpreted terms. Measurement is the gathering of data or information, while evaluation is the decision or interpretation made by an evaluator after an analysis of the data.

An example of measurement data is the height of an individual obtained by using a standard measurement instrument, i.e., a ruler marked off in inches. We try to make the measurement as objective as possible so as not to bias the data and record the measurement as accurately as we can read the ruler. Now, take a second example where we subjectively record the height of an individual as "very tall," "tall," "medium," "short," or "very short." In the first example, an analysis of the data reveals that the information is relatively easy to evaluate if we are familiar with the standard unit used, inches, and it is possible to compare this data with other similar measurements. In the second example, we are at a loss as to the precise meaning of the data. How tall is "very tall?" How short is "short?" How much shorter is a person who is "tall" than one who is "very tall?" A rational or meaningful decision would be extremely difficult to make with this type of data.

In the classroom we are faced with a similar situation in gathering meaningful data concerning the extent and rate

of progress by a student. If we do not gather relatively objective information and convert it to standard units of measure, the data is relatively meaningless. The data gathered must have the same relative meaning to anyone involved in the educational process, whether they be administrator, parent, teacher, student, or consultant. Even with the best possible data, the decisions remain largely subjective, depending on the best judgment of the evaluator. Rational decisions are possible from standardized data; however, arbitrary decisions are not only possible from unstandardized data, they are highly probable.

It seems imperative that we use the best measurement data available on extent and rate of progress in the classroom. From measurement data come decisions which determine classroom activities, affect career counseling, and produce feedback. Measurement instruments must be valid (measure what we want to measure) and reliable (consistently measure what we want to measure). We help insure content validity by checking the face validity (comparison of course objectives with measurement objectives) of the instrument before administering it to the students; however, we are seldom able to accurately predict how each student will perceive either the classroom objectives or the measurement instrument. There is a need to know whether or not an individual decision point is sensitive to differences in extent and rate

of progress as determined by the student. Once we are reasonably assured that the test is reliable and valid, then we can convert raw scores to a standard unit of measure and record them.

The following Measurement Analysis Model has been developed to assist in determining the internal validity and reliability of measurement instruments used in the classroom. It does not cover evaluation or interpretation of measurement data as this necessarily changes with the different requirements and perceptions of the evaluator. It does cover the conversion of measurement information to standard units of measure so that the information may be compared with other similar measurements. The model has been validated and found reliable for groups ranging in size from 20 to 50.

Definitions, examples, and tables have been placed in Part II for easy reference. All examples and tables in Part II have the same code number as the programmed operations in Part I which refer to them.

MEASUREMENT ANALYSIS MODEL

The model is limited to three basic principles involved in gathering and recording measurement data in the classroom.

1.00 Test reliability.

2.00 Standard units of measure.

3.00 Acceptability of decision points.

Working with these principles will involve knowledge of seven major item analysis components. These are: Test Mean, Test Standard Deviation, Test Variance, Test Reliability Coefficient, Standard Scores, Item Discrimination Index, and Item Difficulty Index. An additional component, the Item Probability Index, will be developed within the model to aid in determining the efficiency with which we can predict the future success of an item when used under similar circumstances.

0.00 General item analysis procedures.

This model was developed with two basic purposes in mind: 1. Develop a model suitable for measurement analysis by teachers with a minimum of time and computational effort and 2. Develop a model which would yield valid and reliable item analysis components when used in the classroom. The final form of the model allows the teacher to work with samples of 10 papers each, whether the original class has 20 or 50 students. Fewer students than 20 would not result in stable item analysis components and few classes are expected to exceed 50. If classes do exceed 50, the model can be used by separating the class into two or more parts and then analyze each part separately. Care should be taken to assign papers to each part in a random manner.

In order to limit the number of papers in our samples to 10, it is necessary to select specific papers, so as not to destroy the original characteristics of the data. The procedure used is systematic sampling based on rank order statistics.

0.01 Score all test papers and record each total raw score on the top right hand corner of the test paper. (Note: It is often advantageous to make a checklist on which to record successful accomplishment of a decision

point, especially if we are going to conduct an item analysis.)

0.02 Arrange and number the test papers in order, assigning the numeral one to the paper with the highest raw score and the highest numeral to the paper with the lowest raw score. Tied scores should be ranked as they appear without regard to other criteria.

0.10 Test Mean.

An important component in measurement is the central tendency of a group of scores. There are several ways to interpret this point, but the one we will use is the arithmetic mean, or average score.

0.11 Add the raw scores for the ten ranks or positions given in Table 0.11 for your class size.

0.12 Multiply result found in 0.11 by .10. This will give us the Test Mean for the complete test. (Note: This is the same as moving the decimal one place to the left.)

0.13 Record the Test Mean on your worksheet. (See Table 0.00)

0.14 Divide the Test Mean by the number of items in the test. This converts the Test Mean to a proportion.

0.15 Record the Test Mean Proportion on your worksheet. (See Table 0.00)

0.20 Test Variance.

We also need to know the extent to which each test score varies from the central tendency or mean of the

test. This allows us, not only to determine the reliability of our test, but also to convert our raw scores to a standard unit of measure.

0.21 Add the raw scores for the five upper sample positions given in Table 0.21 for your class size.

0.22 Add the raw scores for the five lower sample positions given in Table 0.22 for your class size.

0.23 Subtract the results of 0.22 from 0.21. This will give us the deviation or variation between people who score high and people who score low on the test.

0.24 Turn to Table 0.24 and locate the row labelled with the value nearest the result found in 0.23.

0.25 Record the value found in the adjoining column labelled Test Standard Deviation.

0.26 Record the value found in the column labelled Test Variance.

1.00 Test reliability.

It is often desirable to determine how consistent our complete test is as a measure of progress. The complete test may represent decision points gathered during a group session, from individual tests, or from many small tests given over a period of time. There are several ways to obtain a Test Reliability Coefficient; however, for our purposes we need consider only one which is easy to compute with the help of a table and which yields a conservative index.

1.10 Test Reliability Coefficient.

1.11 Turn to Table 1.11 and locate the column headed by the value nearest to the Test Mean Proportion found in 0.14.

1.12 Continue down this column until you reach the row headed by the value nearest the Test Variance found in 0.25. The proportion in this cell is the Mean Proportion of Variance.

1.13 Multiply the proportion found in 1.12 by the number of items in the complete test.

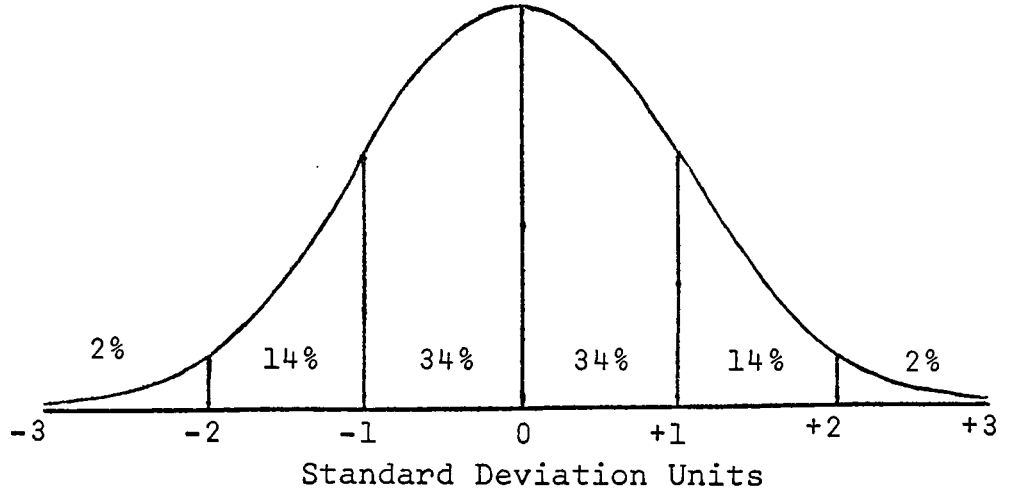
1.14 Subtract the result found in 1.13 from 1.00. The final result is the Test Reliability Coefficient. (Note: The closer the Test Reliability Coefficient is to 1.00, the more reliable or consistent the test. A Test Reliability Coefficient less than .60 is unsatisfactory

and should be improved by replacing or adding decision points. Increasing the number of decision points will usually increase the stability or consistency of the test, providing the new decision points are comparable in quality to those already in the test. A large quantity of decision points will not replace quality in the decision points used. Another reason for low test reliability could be in the selection of decision points covering more than one major topic. Decision points in a test should cover one major topic due to the varying interests, aptitudes, and rates of progress by students.

2.00 Standard units of measure.

We seldom make a decision from a single piece of data and we normally compare different pieces of data. We would never think of combining feet, meters, or miles without first converting them to a common unit, and measurement data is no exception - each instrument represents a different measure. The material is different, students are different, and the environment is different. The test may be hard or easy, long or short. In other words, if we combine raw scores we are adding different measurement units. Raw scores should first be converted to a standard unit of measure. One base for a standard unit of measurement is the normal distribution.

THE NORMAL DISTRIBUTION



Suppose that you had several different pieces of measurement information on your students concerning a specific subject or trait and you wished to group them homogeneously for a learning activity. If you had only recorded raw scores, your progress record would look like this:

	Test 1	Test 2	Test 3	Total	Average
Student A	82	88	92	262	87.3
Student B	96	94	72	262	87.3

Question: Are the levels and extent of progress on this trait the same for these two students?

Answer: You can readily think of many things which you would like to know about this data. Difficulty, length, variability of scores, etc.

Let us consider, for convenience, that each test represents three weekly tests in a single area and that none of the material covered was the same. Also, let us convert the raw scores to standard scores using actual test means and standard deviations.

	Test 1	Test 2	Test 3
Test Mean	82	88	72
Test Standard Deviation	7	3	12

	Standard Scores				
	Test 1	Test 2	Test 3	Total	Average
Student A	80	80	97	257	85.7
Student B	100	100	80	280	93.3

We now see that the level of progress for these two students is in fact very different. Once we convert to standard scores, we know the position of the mean and we know the standard deviation or variability of scores that we can expect in a similar population.

2.10 Standard Scores.

A standard score does not change the relative position of a student on a test nor does it weight individual scores. Its major purpose is to convert raw scores to a single standard unit of measure. It is recommended that standard scores be as simple and straight forward as possible and that they stay constant. A Standard Score Mean of 80 and a Standard Score Standard Deviation unit of 10 are very easy to work with and are similar to the expected range of test scores. Remember, the purpose of standardizing our scores is to develop a standard unit of measurement which can be evaluated by anyone involved in the process of education. When using the above standard unit of measure, we can expect 96 percent of our scores to fall between 60 and 100.

2.11 Take a ruler divided into millimeters and draw a line 8 centimeters long. Mark the centimeter points

on the scale. (Note: You can use the ruler to find smaller division points as they are needed.)

2.12 Label the middle centimeter point on your upper scale with your Test Mean rounded to the nearest whole number.

2.13 Add your Test Standard Deviation to your Test Mean and round the result to the nearest whole number.

2.14 Record the result found in 2.13 as the first value on the right side of your scale. Each centimeter mark is worth one Test Standard Deviation unit and each millimeter is worth one tenth of one Test Standard Deviation unit.

2.15 Subtract your Test Standard Deviation unit from your Test Mean and round the results to the nearest whole number.

2.16 Record the result found in 2.15 as the first value on the left side of your upper scale.

2.17 Complete your upper scale. The right side of the scale calls for adding one Test Standard Deviation unit for each successive centimeter mark and the left side calls for subtracting one Test Standard Deviation unit for each successive centimeter mark.

2.18 Label the middle point on your lower scale with your Standard Score Mean and then complete your lower scale using the same procedure described in 2.17. (Note:

Raw scores can now be converted to standard scores by finding their position on the upper scale and then reading off the value of the lower scale.)

2.19 Locate and record standard scores for each student. (See Table 2.19)

3.00 Acceptability of decision points.

There is a need to conserve time and energy in the classroom. Any decision point which is not sensitive to differences in extent and rate of progress is of doubtful value. We seldom have time to conduct a complete item analysis; however, it is always possible to take a small sample and closely approximate this procedure. There are three major item analysis components which we will need before we can answer the question, "What is the probability that a specific decision point can detect a significant difference in student progress?"

3.10 Item Discrimination Index.

This index is concerned with how well each decision point or item differentiates between students who score high and students who score low on the total test.

3.11 Select the ten papers with the highest total raw scores as your upper sample.

3.12 Select the ten papers with the lowest total raw scores as your lower sample.

3.13 Add the number of correct responses by students in the upper sample to Decision Point 1.

3.14 Add the number of correct responses by students in the lower sample to Decision Point 1.

3.15 Record the results from 2.13 and 2.14 in two columns for each decision point and continue the process until all decision points have been analyzed. (See Table 0.00) (Note: It is seldom necessary to analyze every item each time it is used. This depends on the stability of previous item analysis components.)

3.16 Turn to Table 3.16 and locate the column headed by the value nearest the number of students in the upper sample having Decision Point 1 correct.

3.17 Go down this column until you reach the cell in the row headed by the value nearest the number of students in the lower sample having Decision Point 1 correct. The proportion found in this cell is the Item Discrimination Index.

3.18 Record the Item Discrimination Index in the column on your worksheet labelled DIS INDEX. (Note: Be sure and retain the (-) sign when recording this index.)

3.19 Continue the process until all items have been analyzed.

3.20 Item Probability Index.

The Item Discrimination Index will fluctuate with different groups and with different numbers of people taking the test. An Item Discrimination Index of +.40 obtained on a decision point given to ten people is not as significant as an Item Discrimination Index of +.25 obtained on a

decision point given to sixty people. The Item Probability Index takes this aspect into consideration and allows us to rank each decision point in terms of our ability to predict the efficiency with which the decision point will discriminate between differences in student progress when given to similar groups.

3.21 Turn to Table 3.21 and locate the row containing the Discrimination Index found for Decision Point 1.

3.22 Record the Item Probability Index found in the adjoining column on your worksheet in the column labelled PROB INDEX. (See Table 0.00)

3.23 Continue the process until all items have been analyzed. (Note: The Item Probability Index means that we can expect this decision point to discriminate between differences in student progress so many times out of 100 when given to similar groups.)

3.30 Item Difficulty Index.

One thing that the Item Probability Index will not tell us is the difficulty level of each decision point. This index is important in developing individual mastery tests and in gearing a test to fit the needs of a specific individual or group. In other than mastery tests, it is doubtful if items below a difficulty level of 30% (.30) or above a difficulty level of 80% (.80) are of value in

diagnosing level and rate of progress. The primary argument against using these difficulty levels is the small number of people who are able to satisfy our requirements when the difficulty level is .30 or below and the small number of people who fail to meet our requirements when the difficulty level is .80 or above.

3.31 Add the number of correct responses to Decision Point 1 by both the upper and lower samples.

3.32 Multiply the result from 3.31 by .05. The result is the Item Difficulty Index for Decision Point 1.

3.33 Record the Item Difficulty Index in the column on your worksheet labelled DIFF INDEX. (Note: We have now computed the item analysis components which will assist us in determining the acceptability of each decision point. Item Probability Indexes of .80 and above are considered minimal in developing decision points for continued use in the classroom. The development of good decision points is not an easy, overnight job; however, if we pool our efforts and start weeding out and improving inferior decision points, our store of valid and reliable decision points will steadily increase. The decisions we make from the information we gather on each child will help determine the future of that child. If we make decisions from faulty information, we are not performing our function as a teacher.)

MEASUREMENT ANALYSIS MODEL

PART II: DEFINITIONS, EXAMPLES, AND TABLES

Ray Hassett

University of Houston

JUNE 1970

TABLE OF CONTENTS

DEFINITIONS OF TERMS

bias - an element which gives a distorted picture. An element which interferes with rational interpretation of data. EX: All children with bright eyes are intelligent.

central tendency - a point in a distribution where the majority of the cases tend to fall. EX: The arithmetic mean is a measure of central tendency.

data - any information which is available on a subject or trait. EX: Student personnel records contain data on a student's home environment.

decision point - the point at which we desire to assess successful or nonsuccessful achievement of a standard. EX: An item on a mathematics test may have one or more decision points.

evaluation - the judgment made after an analysis of information available on the subject or trait. EX: The guidance counselor looked over the mental and achievement records for Mary and decided that she belonged in an accelerated English class.

face validity - validity as determined by the test maker or the test taker after looking at the test in terms of their perception of the course objectives. EX: Johnny felt that the test was fair and appropriate when he took it.

feedback - data which is returned to the user for action or information concerning an ongoing program. EX: The score on a measurement instrument is feedback to the teacher, student, counselor, etc., on a student's progress.

instrument - any device which has been developed to help measure achievement on a subject or trait. EX: The Fels Behavior Rating Scale measures a child's home environment.

item analysis - a study of test results to determine difficulty, discrimination, reliability, etc. EX: Item 2 on our 8th Grade English Test has a difficulty Index of .30 and a Discrimination Index of .40.

Item Difficulty Index - the percentage or proportion of students taking the test who met the requirements of a decision point or item. EX: If 20 out of 25 students successfully answered an item, then the Item Difficulty Index would be 20/25 or .80.

Item Discrimination Index - the ability of an item or decision point to differentiate between those who score high on the total test and those who score low, expressed as a proportion. EX: If the majority of those answering an item also have high total scores on the complete test and those unable to answer the item have low total scores, the item would have a high Item Discrimination Index.

Item Probability Index - the probability than an item will discriminate between different levels and rates of progress. EX: An Item Probability Index of .99 would mean that we could expect an item to discriminate between different levels and rates of progress 99 times out of 100.

measurement - the numerical results of any test where the student is observed or measured in terms of specific requirements or rules. EX: Ethel was measured on her ability to satisfactorily meet the requirements for ten decision points on a physical dexterity test.

normal distribution - a distribution which fits the theoretical premise that all measurable traits follow a pattern. The Normal Distribution is bell-shaped and has its mean at its greatest height with equal area under both sides of the curve. EX: The height of most people will cluster around a central value with fewer and fewer people falling at the extremes.

objective - a decision based on a standard which will be relatively the same in all similar circumstances. EX: An objective test is one where a scoring key is made available as a standard.

probability - the amount of confidence which can be placed on a statement or decision point. EX: We say that there is a 90 percent probability that it will rain. This

means that there are 90 out of 100 possible chances for it to rain.

proportion - the percentage of the total component being measured which meets a standard, expressed in decimal form. EX: If 20 out of 25 students can do 50 situps, then the proportion able to do 50 situps would be .80 or 80 percent.

random - selection, by chance, of objects or students for a category or group. EX: Every third paper could be selected to form a new group. A specific paper would then fall by chance into the new group.

raw score - the total score on a test before it has been changed in any way. EX: Lucille met the requirements on 24 decision points with each decision point having equal weight. Her raw score would be 24.

reliability - consistency or stability of a test. A test would be considered reliable if it consistently produced similar results under similar circumstances. EX: Alfred scored 120 on an IQ test. Two years later he was retested with the same IQ test and again scored 120. The probability is that the IQ test used was reliable and stable.

significant - a relationship which is much used by statisticians to indicate degree of difference in trait being observed. It is a point at which error or chance can be discounted when viewing differences in performance. EX: In the normal distribution, two standard error units

(standard deviations) above or below the mean is considered to be a significant difference from the mean.

standard - a criterion which must be met, or several levels of performance criteria against which a student's performance can be measured. EX: Any student with a raw score of 80 on our Spanish test can be admitted to Second Year Spanish classes.

standard score - a score which has a preset mean and standard deviation and which uses the properties of the normal distribution to convert raw scores to equal interval data. EX: If the test mean was 20 and the standard deviation was 5, we could convert the raw scores to a standard score by selecting a new mean of 80 and a new standard deviation of 10 and then convert the raw scores through substitution.

subjective - decisions or interpretations which are based on opinions, feelings, attitudes, etc. EX: I feel it is going to rain today.

test - a group of decision points on a single trait which is administered for the purpose of assessing performance. EX: We wish to test George on the concept of "time" and develop ten decision points and administer them to him.

Test Mean - the sum of the raw scores divided by the total number taking the test. This would be the average score on the test. EX: If the sum of the raw scores on a

test equaled 120 and 10 students took the test, the mean would be 12.

Test Reliability Coefficient - a proportion obtained by finding the consistency or stability of a test. EX: If the Test Reliability Coefficient of an IQ test was .94, we would consider the test to be reliable and stable. A proportion of 1.00 would be considered perfect reliability.

Test Standard Deviation - the amount of deviation from the mean in terms of standard error units on the normal distribution. EX: If the Test Mean is 20 and the Test Standard Deviation is 5, we can expect 68% of our class to have raw scores between 15 and 25.

Test Variance - the total amount of variability between all raw scores on a test and the mean. It is also the squared value of the Test Standard Deviation. EX: If Test Standard Deviation equals 10, the Test Variance would be 100.

validity - the extent to which a test measures what it is supposed to measure. EX: Content validity refers to how well a test covers both the content and objectives of a course.

EXAMPLES

0.00 General item analysis procedures. (See Table 0.00)

0.01 See Table 0.01

0.02 See Table 0.02

0.10 Test Mean

0.11 34 + 30 + 38 + 27 + 26 + 23 + 22 + 20 + 17 + 13 = 240 (See Tables 0.02 and 0.11)

0.12 240 x .10 = 24

0.13 See Table 0.00

0.14 24 ÷ 50 = .48

0.15 See Table 0.00

0.20 Test Variance.

0.21 38 + 34 + 32 + 30 + 28 = 162 (See Tables 0.02 and 0.21)

0.22 20 + 17 + 15 + 13 + 10 = 75 (See Tables 0.02 and 0.22)

0.23 162 - 75 = 87

0.24 See Table 0.24

0.25 6.43 (See Tables 0.00 and 0.24)

0.26 41.34 (See Tables 0.00 and 0.24)

1.00 Test reliability.

1.10 Test Reliability Coefficient

1.11 See Table 1.11

1.12 .006 (See Table 1.11)

1.13 .006 x 50 = .30

1.14 1.00 - .30 = .70 (See Table 0.00)

2.00 Standard units of measure

2.10 Standard Scores.

2.11 See Table 0.00

2.12 24 (See Table 0.00)

2.13 24 = 6.43 = 30.43 or 30

2.14 See Table 0.00

2.15 24 - 6.43 = 17.57 or 18

2.16 See Table 0.00

2.17 See Table 0.00

2.18 See Table 0.00

2.19 See Table 2.19

3.00 Acceptability of decision points.

3.10 Item Discrimination Index.

3.11 Upper Sample (See Table 0.02)

Rank	Student	Raw Score	Rank	Student	Raw Score
1	1	38	6	5	29
2	9	34	7	23	29
3	28	32	8	21	28
4	17	31	9	29	28
5	10	30	10	4	27

3.12 Lower Sample (See Table 0.02)

Rank	Student	Raw Score	Rank	Student	Raw Score
30	27	10	25	14	17
29	2	13	24	8	17
28	19	15	23	30	20
27	3	16	22	15	21
26	16	17	21	13	21

3.13 3 (See Table 0.00)

3.14 1 (See Table 0.00)

3.15 See Table 0.00

3.16 See Table 3.16

3.17 .20 (See Table 0.00)

3.18 See Table 0.00

3.19 See Table 0.00

3.20 Item Probability Index.

3.21 See Table 3.21

3.22 .80 (See Tables 0.00 and 3.21)

3.23 See Tables 0.00 and 3.21

3.30 Item Difficulty Index.

3.31 3 + 1 = 4

3.32 4 x .05 = .20 (See Table 0.00)

3.33 See Table 0.00

TEST WORKSHEET

TEST Word Problems I

NUMBER OF STUDENTS TAKING TEST 30

NUMBER OF DECISION POINTS IN TEST 50

TEST MEAN (0.12) 24 STANDARD DEVIATION (0.25) 6.43

MEAN PROPORTION (0.14) .48 TEST VARIANCE (0.26) 41.34

RELIABILITY COEFFICIENT (1.14) .70

STANDARD SCORE SCALE (2.10)

Raw Scores	0	6	12	18	24	30	36	42	48
Standard Scores	40	50	60	70	80	90	100	110	120

ITEM ANALYSIS

ITEM	CORRECT RESPONSES UPPER (3.13)	LOWER (3.14)	DISC (3.17)	PROB (3.22)	DIFF (3.32)
1	3	1	.20	.80	.20
9	8	5	.25	.86	.65
17	3	0	.34	.93	.15
25	9	1	.64	1.00	.50
29	6	3	.24	.85	.45
33	6	2	.33	.92	.40
34	5	0	.46	.98	.25
35	4	5	−.08	.37	.45
39	9	5	.35	.94	.70
48	9	3	.49	.99	.60

ITEM ANALYSIS (Continued)

ITEM	CORRECT RESPONSES UPPER (3.13)	CORRECT RESPONSES LOWER (3.14)	DISC (3.17)	PROB (3.22)	DIFF (3.32)

TABLE 0.01 SAMPLE TEST PAPER

NAME _____, John TOTAL RAW SCORES 38

CONCEPT Word Problems I STANDARD SCORE 103

POSITION OF PAPER 1

DECISION POINTS

1. 1	16. 1	31. 0	46. 1
2. 0	17. 0	32. 1	47. 1
3. 0	18. 0	33. 1	48. 1
4. 0	19. 1	34. 1	49. 1
5. 1	20. 1	35. 1	50. 1
6. 1	21. 1	36. 1	51.
7. 1	22. 1	37. 1	52.
8. 1	23. 0	38. 1	53.
9. 1	24. 1	39. 1	54.
10. 1	25. 1	40. 1	55.
11. 0	26. 1	41. 1	56.
12. 0	27. 0	42. 1	57.
13. 1	28. 1	43. 0	58.
14. 1	29. 1	44. 1	59.
15. 1	30. 0	45. 1	60.

COMMENTS: Knows decimals. Needs drill on multiplication facts, and more aid in concept development of working with whole numbers in the operation of division.

TABLE 0.02 RANKING OF THIRTY SAMPLE TEST PAPERS

POSITION	STUDENT	RAW SCORE	POSITION	STUDENT	RAW SCORE
1	1	38	16	7	23
2	9	34	17	12	23
3	28	32	18	6	22
4	17	31	19	11	22
5	10	30	20	18	22
6	5	29	21	13	21
7	23	29	22	15	21
8	21	28	23	30	20
9	29	28	24	8	17
10	4	27	25	14	17
11	20	27	26	16	17
12	22	27	27	3	16
13	24	27	28	19	15
14	26	26	29	2	13
15	25	24	30	27	10

TABLE 0.11 SELECTION OF SAMPLE FOR COMPUTATION OF MEAN

CLASS SIZE	POSITIONS TO BE SAMPLED									
20	1	3	5	7	9	12	14	16	18	20
21	1	3	5	7	9	13	15	17	19	21
22	1	3	6	8	10	13	15	17	20	22
23	1	3	6	8	10	14	16	18	21	23
24	1	4	6	8	11	14	17	19	21	24
25	1	4	6	9	11	15	17	20	22	25
26	1	4	7	9	12	15	18	20	23	26
27	1	4	7	9	12	16	19	21	24	27
28	1	4	7	10	13	16	19	22	25	28
29	1	4	7	10	13	17	20	23	26	29
30	2	5	8	11	14	17	20	23	26	29
31	2	5	8	11	14	18	21	24	27	30
32	2	5	8	11	14	19	22	25	28	31
33	2	5	8	12	15	19	22	26	29	32
34	2	5	8	12	15	20	23	27	30	33
35	2	5	9	12	16	20	24	27	31	34
36	2	5	9	13	16	21	24	28	32	35
37	2	6	9	13	17	21	25	29	32	36
38	2	6	10	13	17	22	26	29	33	37
39	2	6	10	14	18	22	26	30	34	38
40	2	6	10	14	18	23	27	31	35	39
41	2	6	10	14	18	24	28	32	36	40
42	2	6	10	15	19	24	28	33	37	41
43	2	6	11	15	19	25	29	33	38	42
44	2	7	11	15	20	25	30	34	38	43
45	2	7	11	16	20	26	30	35	39	44
46	2	7	12	16	21	26	31	35	40	45
47	2	7	12	16	21	27	32	36	41	46
48	2	7	12	17	22	27	32	37	42	47
49	2	7	12	17	22	28	33	38	43	48
50	2	8	12	18	22	29	33	39	43	49

TABLE 0.21 SELECTION OF UPPER SAMPLE FOR TEST VARIANCE

CLASS SIZE	RANKS OR POSITIONS TO BE SAMPLED					CLASS SIZE	RANKS OR POSITIONS TO BE SAMPLED				
20	1	2	3	4	5	35	1	2	4	6	9
21	1	2	3	4	5	36	1	2	4	6	9
22	1	2	3	4	6	37	1	2	4	6	9
23	1	2	3	4	6	38	1	2	4	6	10
24	1	2	3	4	6	39	1	2	4	6	10
25	1	2	3	4	6	40	1	2	4	6	10
26	1	2	3	4	7	41	1	2	4	7	10
27	1	2	3	4	7	42	1	2	4	7	10
28	1	2	3	4	7	43	1	2	4	7	11
29	1	2	3	5	7	44	1	2	4	7	11
30	1	2	3	5	8	45	1	2	4	7	11
31	1	2	3	5	8	46	1	2	5	7	12
32	1	2	3	5	8	47	1	2	5	8	12
33	1	2	3	5	8	48	1	2	5	8	12
34	1	2	3	5	8	49	1	2	5	8	12
						50	1	2	5	8	12

TABLE 0.22 SELECTION OF LOWER SAMPLE FOR TEST VARIANCE

CLASS SIZE	RANKS OR POSITIONS TO BE SAMPLED					CLASS SIZE	RANKS OR POSITIONS TO BE SAMPLED				
20	16	17	18	19	20	35	27	30	32	34	35
21	17	18	19	20	21	36	28	31	33	35	36
22	17	19	20	21	22	37	29	32	34	36	37
23	18	20	21	22	23	38	29	33	35	37	38
24	19	21	22	23	24	39	30	34	36	38	39
25	20	22	23	24	25	40	31	35	37	39	40
26	20	23	24	25	26	41	32	35	38	40	41
27	21	24	25	26	27	42	33	36	39	41	42
28	22	25	26	27	28	43	33	37	40	42	43
29	23	25	27	28	29	44	34	38	41	43	44
30	23	26	28	29	30	45	35	39	42	44	45
31	24	27	29	30	31	46	35	40	42	45	46
32	25	28	30	31	32	47	36	40	43	46	47
33	26	29	31	32	33	48	37	41	44	47	48
34	27	30	32	33	34	49	38	42	45	48	49
						50	39	43	46	49	50

TABLE 0.24 CONVERSION OF DEVIATION SCORES TO STANDARD DEVIATION AND VARIANCE

DEVIATION	STANDARD DEVIATION	VARIANCE	DEVIATION	STANDARD DEVIATION	VARIANCE
1	.07	.01	36	2.66	7.08
2	.15	.02	37	2.73	7.48
3	.22	.05	38	2.81	7.89
4	.30	.09	39	2.88	8.31
5	.37	.14	40	2.96	8.74
6	.44	.20	41	3.03	9.18
7	.52	.27	42	3.10	9.63
8	.59	.35	43	3.18	10.10
9	.67	.44	44	3.25	10.57
10	.74	.55	45	3.33	11.06
11	.81	.66	46	3.40	11.56
12	.89	.79	47	3.47	12.06
13	.96	.92	48	3.55	12.58
14	1.03	1.07	49	3.62	13.11
15	1.11	1.23	50	3.70	13.65
16	1.18	1.40	51	3.77	14.20
17	1.26	1.58	52	3.84	14.77
18	1.33	1.77	53	3.92	15.34
19	1.40	1.97	54	3.99	15.92
20	1.48	2.18	55	4.06	16.52
21	1.55	2.41	56	4.14	17.13
22	1.63	2.64	57	4.21	17.74
23	1.70	2.89	58	4.29	18.37
24	1.77	3.15	59	4.36	19.01
25	1.85	3.41	60	4.43	19.66
26	1.92	3.69	61	4.51	20.32
27	2.00	3.98	62	4.58	20.99
28	2.07	4.28	63	4.66	21.68
29	2.14	4.59	64	4.73	22.37
30	2.22	4.92	65	4.80	23.07
31	2.29	5.25	66	4.88	23.79
32	2.36	5.59	67	4.95	24.52
33	2.44	5.95	68	5.03	25.25
34	2.51	6.31	69	5.10	26.00
35	2.59	6.69	70	5.17	26.76

DEVIATION	STANDARD DEVIATION	VARIANCE	DEVIATION	STANDARD DEVIATION	VARIANCE
71	5.25	27.53	106	7.83	61.36
72	5.32	28.31	107	7.91	62.53
73	5.39	29.10	108	7.98	63.70
74	5.47	29.91	109	8.06	64.88
75	5.54	30.72	110	8.13	66.08
76	5.62	31.54	111	8.20	67.29
77	5.69	32.38	112	8.28	68.51
78	5.76	33.23	113	8.35	69.73
79	5.84	34.08	114	8.42	70.97
80	5.91	34.95	115	8.50	72.22
81	5.99	35.83	116	8.57	73.49
82	6.06	36.72	117	8.65	74.76
83	6.13	37.62	118	8.72	76.04
84	6.21	38.53	119	8.79	77.34
85	6.28	39.46	120	8.87	78.64
86	6.36	40.39	121	8.94	79.96
87	6.43	41.34	122	9.02	81.28
88	6.50	42.29	123	9.09	82.62
89	6.58	43.26	124	9.16	83.97
90	6.65	44.24	125	9.24	85.33
91	6.72	45.22	126	9.31	86.70
92	6.80	46.22	127	9.39	88.08
93	6.87	47.23	128	9.46	89.48
94	6.95	48.26	129	9.53	90.88
95	7.02	49.29	130	9.61	92.29
96	7.09	50.33	131	9.68	93.72
97	7.17	51.38	132	9.75	95.16
98	7.24	52.45	133	9.83	96.60
99	7.32	53.53	134	9.90	98.06
100	7.39	54.61	135	9.98	99.53
101	7.46	55.71	136	10.05	101.01
102	7.54	56.82	137	10.12	102.50
103	7.61	57.94	138	10.20	104.00
104	7.69	59.07	139	10.27	105.52
105	7.76	60.21	140	10.35	107.04

TABLE 1.11 MEAN PROPORTION OF VARIANCE (MPV)

Test Mean in Proportions

Test Variance	.30	.35	.40	.45	.50	.55	.60	.65	.70	.75	.80
6	.035	.038	.040	.041	.042	.041	.040	.038	.035	.031	.027
7	.030	.033	.034	.035	.036	.035	.034	.033	.030	.027	.023
8	.026	.028	.030	.031	.031	.031	.030	.028	.026	.023	.020
9	.023	.025	.027	.028	.028	.028	.027	.025	.023	.021	.018
10	.021	.023	.024	.025	.025	.025	.024	.023	.021	.019	.016
11	.019	.021	.022	.023	.023	.023	.022	.021	.019	.017	.015
12	.018	.019	.020	.021	.021	.021	.020	.019	.018	.016	.013
13	.016	.018	.019	.019	.019	.019	.019	.018	.016	.014	.012
14	.015	.016	.017	.018	.018	.018	.017	.016	.015	.013	.011
15	.014	.015	.016	.017	.017	.017	.016	.015	.014	.013	.011
16	.013	.014	.015	.016	.016	.016	.015	.014	.013	.012	.010
17	.012	.013	.014	.015	.015	.015	.014	.013	.012	.011	.009
18	.012	.013	.013	.014	.014	.014	.013	.013	.012	.010	.009
19	.011	.012	.013	.013	.013	.013	.013	.012	.011	.010	.008
20	.010	.011	.012	.012	.013	.012	.012	.011	.010	.009	.008
21	.010	.011	.011	.012	.012	.012	.011	.011	.010	.009	.008
22	.010	.010	.011	.011	.011	.011	.011	.010	.010	.009	.007
23	.009	.010	.010	.011	.011	.011	.010	.010	.009	.008	.007
24	.009	.010	.010	.010	.010	.010	.010	.010	.009	.008	.007
25	.008	.009	.010	.010	.010	.010	.010	.009	.008	.008	.006
26	.008	.009	.009	.010	.010	.010	.009	.009	.008	.007	.006
27	.008	.008	.009	.009	.009	.009	.009	.008	.008	.007	.006
28	.008	.008	.009	.009	.009	.009	.009	.008	.008	.007	.006
29	.007	.008	.008	.008	.009	.009	.008	.008	.007	.007	.006
30	.007	.008	.008	.008	.008	.008	.008	.008	.007	.006	.005
35	.006	.007	.007	.007	.007	.007	.007	.007	.006	.005	.004
40	.005	.006	.006	.006	.006	.006	.006	.006	.005	.005	.004
45	.005	.005	.005	.006	.006	.006	.005	.005	.005	.004	.004
50	.004	.005	.005	.005	.005	.005	.005	.005	.004	.004	.003
60	.004	.004	.004	.004	.004	.004	.004	.004	.004	.003	.003
70	.003	.003	.003	.004	.004	.004	.003	.003	.003	.003	.002
80	.003	.003	.003	.003	.003	.003	.003	.003	.003	.002	.002
90	.002	.003	.003	.003	.003	.003	.003	.003	.002	.002	.002
100	.002	.002	.002	.002	.003	.002	.002	.002	.002	.002	.002
150	.001	.002	.002	.002	.002	.002	.002	.002	.001	.001	.001
200	.001	.001	.001	.001	.001	.001	.001	.001	.001	.001	.001

TABLE 2.19 SAMPLE CONVERSION OF RAW SCORES TO STANDARD SCORES

STUDENT	RAW SCORE	POSITION	STANDARD SCORE
1	38	1	103
9	34	2	97
28	32	3	93
17	31	4	92
10	30	5	90
5	29	6	88
23	29	7	88
21	28	8	87
29	28	9	87
4	27	10	85
20	27	11	85
22	27	12	85
24	27	13	85
26	26	14	83
25	24	15	80
7	23	16	78
12	23	17	78
6	22	18	77
11	22	19	77
18	22	20	77
13	21	21	75
15	21	22	75
30	20	23	73
8	17	24	68
14	17	25	68
16	17	26	68
3	16	27	67
19	15	28	65
2	13	29	62
27	10	30	57

TABLE 3.16 ITEM DISCRIMINATION INDEX

Number in Lower Sample Getting Item Correct (3.14)	Number in Upper Sample Getting Item Correct (3.13)										
	0	1	2	3	4	5	6	7	8	9	10
0	.00	.18	.27	.34	.40	.46	.52	.59	.65	.72	.80
1	-.18	.00	.11	.20	.28	.35	.42	.49	.56	.64	.72
2	-.27	-.11	.00	.09	.17	.25	.33	.40	.48	.56	.65
3	-.34	-.20	-.09	.00	.08	.16	.24	.32	.40	.49	.59
4	-.40	-.28	-.17	-.08	.00	.08	.16	.24	.33	.42	.52
5	-.46	-.35	-.25	-.16	-.08	.00	.08	.16	.25	.35	.46
6	-.52	-.42	-.33	-.24	-.16	-.08	.00	.08	.17	.28	.40
7	-.59	-.49	-.40	-.32	-.24	-.16	-.08	.00	.09	.20	.34
8	-.65	-.56	-.48	-.40	-.33	-.25	-.17	-.09	.00	.11	.27
9	-.72	-.64	-.56	-.49	-.42	-.35	-.28	-.20	-.11	.00	.18
10	-.80	-.72	-.65	-.59	-.52	-.46	-.40	-.34	-.27	-.18	.00

TABLE 3.21 CONVERSION OF DISCRIMINATION INDEX TO PROBABILITY INDEX

DISCRIMINATION INDEX		PROBABILITY INDEX	DISCRIMINATION INDEX		PROBABILITY INDEX
From	To		From	To	
-1.00	-.57	.00	.01		.52
- .56	-.49	.01	.02		.54
- .48	-.45	.02	.03		.55
- .44	-.42	.03	.04		.57
- .41	-.39	.04	.05		.59
- .38	-.37	.05	.06		.60
- .36	-.35	.06	.07		.62
- .34		.07	.08		.63
- .33	-.32	.08	.09		.65
- .31		.09	.10		.66
- .30		.10	.11		.68
- .29		.11	.12		.69
- .28	-.27	.12	.13		.70
- .26		.13	.14		.72
- .25		.14	.15		.73
- .24		.15	.16		.75
- .23		.16	.17		.76
- .22		.18	.18		.77
- .21		.19	.19		.79
- .20		.20	.20		.80
- .19		.21	.21		.81
- .18		.22	.22		.82
- .17		.24	.23		.84
- .16		.25	.24		.85
- .15		.27	.25		.86
- .14		.28	.26		.87
- .13		.30	.27	.28	.88
- .12		.31	.29		.89
- .11		.32	.30		.90
- .10		.34	.31		.91
- .09		.35	.32	.33	.92
- .08		.37	.34		.93
- .07		.38	.35	.36	.94
- .06		.40	.37	.38	.95
- .05		.41	.39	.41	.96
- .04		.43	.42	.43	.97
- .03		.45	.44	.48	.98
- .02		.47	.49	.55	.99
- .01		.48	.56	1.00	1.00
.00		.50			

APPENDIX C: Validation of Measurement Analysis Model

Table 31

Absolute Deviations of Group I Estimates from Actual Values of Item Probability Indexes

Subject	All Items	Odd Items	Even Items
1	2.11	1.04	1.07
5	2.72	1.41	1.31
7	2.47	1.24	1.23
11	2.29	.95	1.34
13	2.80	1.18	1.62
15	2.43	.94	1.49
21	3.57	1.44	2.13
25	3.07	.94	2.13
31	2.17	.94	1.23
33	2.87	.94	1.93
35	2.17	.94	1.23
39	2.17	.94	1.23
41	3.12	1.34	1.78
45	2.17	.94	1.23
49	2.12	.94	1.18
51	6.31	3.33	2.98

Table 32

Absolute Deviations of Group II Estimates from Actual Values of Item Probability Indexes

Subject	All Items	Odd Items	Even Items
2	6.27	3.00	3.27
6	5.91	2.57	3.34
10	8.29	4.50	3.79
14	6.91	3.92	2.99
16	8.05	4.42	3.63
22	6.53	3.61	2.92
24	6.06	2.13	3.93
28	6.80	3.70	3.10
30	6.69	3.66	3.03
34	5.39	2.48	2.91
36	5.96	3.20	2.76
40	6.21	3.28	2.93
44	5.41	2.15	3.26
46	8.57	3.88	4.69
50	6.79	3.63	3.16
54	5.38	2.37	3.01

Table 33

Absolute Deviations of Group III Estimates from Actual Values of Item Probability Indexes

Subject	All Items	Odd Items	Even Items
101	2.13	.98	1.15
103	2.40	.94	1.46
105	2.36	.98	1.38
107	2.40	1.17	1.23
109	2.43	.94	1.49
111	2.17	.93	1.24
113	3.47	1.30	2.17
115	2.17	.94	1.23
117	3.43	1.72	1.71
123	2.17	.94	1.23
125	2.17	.94	1.23
129	2.93	.94	1.99
131	2.17	.94	1.23
133	2.17	.94	1.23
135	2.43	.94	1.49
139	2.24	.94	1.30

Table 34

Absolute Deviations of Group IV Estimates from Actual Values of Item Probability Indexes

Subject	All Items	Odd Items	Even Items
102	4.08	2.09	1.99
104	4.79	2.63	2.16
106	8.30	4.56	3.74
108	6.94	3.62	3.32
110	6.29	3.25	3.04
112	6.80	3.70	3.10
118	8.15	3.87	4.28
120	4.62	1.81	2.81
122	7.26	3.68	3.58
124	8.44	4.69	3.75
126	6.61	2.68	3.93
130	6.36	3.73	2.63
134	6.21	3.42	2.79
136	6.38	2.26	4.12
138	6.82	3.76	3.06
140	7.82	4.10	3.72

Table 35

Measurement Analysis Model Estimates and Actual Values of Item Probability Indexes

Item	Model	Actual	Deviation
3	.35	.36	-.01
5	.94	.92	+.02
7	.37	.56	-.19
9	.88	.95	-.07
11	.35	.16	+.19
12	.37	.40	-.03
13	.88	.72	+.16
14	.25	.48	-.23
15	.93	1.00	-.07
19	.63	.80	-.17
21	.86	.94	-.08
23	.88	.99	-.11
24	.94	.97	-.03
27	.50	.69	-.19
32	.76	.94	-.18
35	.93	.99	-.06
38	.94	.91	+.03
39	.50	.28	+.22
49	.76	.76	.00
50	.15	.02	+.13

Table 36

Measurement Analysis Model Estimates and Actual Values of Major Item Analysis Components for Teacher-made Tests

Test	Mean		Standard Deviation		Variance		Reliability		Cases
	Model	Actual	Model	Actual	Model	Actual	Model	Actual	
A	80.80	81.12	9.46	8.71	89.48	75.94	.80	.81	33
B	77.50	77.44	6.28	6.28	39.46	39.43	.60	.56	34
C	83.00	83.00	9.09	9.26	82.62	85.71	.80	.84	21
D	76.00	75.38	8.06	8.80	64.88	77.37	.70	.77	21
E	71.90	72.00	8.57	9.06	73.49	82.10	.70	.76	21
F	75.10	74.80	9.46	9.89	89.48	97.76	.80	.82	20
G	76.80	76.80	6.28	6.53	39.46	42.67	.50	.59	43
H	77.40	77.37	4.95	4.97	24.52	24.74	.20	.30	43
J	86.10	86.22	5.32	5.22	28.31	27.29	.50	.57	23
K	82.60	82.27	7.46	8.31	55.71	69.01	.80	.80	22
L	91.60	91.81	4.51	4.50	20.32	20.25	.60	.64	42
M	90.60	90.76	11.08	11.31	122.88	127.91	.90	.94	42
N	76.90	76.75	5.10	5.62	26.00	31.59	.30	.27	20
P	78.00	77.05	8.20	9.11	67.29	83.05	.80	.80	20
Q	75.00	74.95	7.32	8.35	53.53	69.65	.60	.74	20
R	75.10	75.30	7.24	8.08	52.45	65.21	.60	.72	20
S	77.30	77.03	7.02	7.72	49.29	59.56	.60	.71	33
T	75.50	75.45	5.91	5.99	34.95	35.90	.50	.49	33
U	75.80	75.67	5.69	6.04	32.38	36.48	.40	.51	21
V	81.70	81.76	9.90	10.29	98.06	105.92	.80	.87	21
F	75.10	74.80	9.46	9.89	89.48	97.76	.80	.82	20
G	76.80	76.80	6.28	6.53	39.46	42.67	.50	.59	43
H	77.40	77.37	4.95	4.97	24.52	24.74	.20	.30	43

Table 37

Measurement Analysis Model Estimates and Actual Values of Standard Scores for Teacher-made Tests

*Raw Score	Tests									
	A		B		C		D		E	
	Model	Actual	Model	Actual	Model	Actual	Model	Actual	Model	Actual
1	99	99	100	100	93	93	108	106	98	98
2	90	90	100	100	92	92	90	90	98	98
3	90	90	94	95	91	91	89	89	90	90
4	88	88	91	92	91	91	88	88	89	89
5	84	84	91	92	91	91	86	86	89	89
6	84	84	87	87	88	88	85	85	89	89
7	84	84	87	87	87	86	81	82	87	87
8	84	84	83	84	86	85	80	81	86	86
9	79	78	83	84	86	85	78	78	84	84
10	79	78	83	84	82	82	75	76	83	83
11	79	78	71	73	82	82	75	76	81	81
12	76	75	71	73	80	80	74	75	78	78
13	76	75	71	73	79	79	72	74	78	78
14	76	75	70	71	78	78	72	74	72	72
15	73	73	70	71	77	77	71	73	72	72
16	70	70	69	70	76	76	70	72	72	72
17	68	67	69	70	72	72	69	70	68	68
18	68	67	69	70	64	65	68	69	67	67
19	68	67	64	67	60	61	66	68	67	67
20	60	59	64	67	60	61	66	68	67	67

*Twenty raw scores were selected from each test by using a table of random numbers.

Table 37 (continued)

*Raw Score	Tests									
	F		G		H		J		K	
	Model	Actual	Model	Actual	Model	Actual	Model	Actual	Model	Actual
1	102	100	102	101	102	101	98	97	97	95
2	98	96	102	101	102	101	98	97	97	95
3	93	92	97	96	102	101	98	97	97	95
4	91	90	93	93	102	101	98	97	97	95
5	88	87	92	91	96	95	88	87	97	95
6	84	84	88	88	90	89	86	85	96	94
7	84	84	85	85	88	87	86	85	83	83
8	82	82	85	85	82	81	86	85	83	83
9	81	81	85	85	82	81	78	78	76	77
10	81	81	85	85	82	81	78	78	76	77
11	80	80	78	79	80	79	78	78	76	77
12	76	76	77	78	80	79	78	78	73	75
13	76	76	73	74	78	77	78	78	69	71
14	76	76	73	74	76	75	78	78	69	71
15	74	75	73	74	74	73	78	78	69	71
16	74	75	70	71	74	73	68	68	69	71
17	64	66	70	71	72	71	68	68	69	71
18	63	65	70	71	72	71	66	66	69	71
19	63	65	68	70	72	71	66	66	67	70
20	63	65	68	70	70	69	66	66	67	70

*Twenty raw scores were selected from each test by using a table of random numbers.

Table 37 (continued)

*Raw Score	Tests									
	L		M		N		P		Q	
	Model	Actual	Model	Actual	Model	Actual	Model	Actual	Model	Actual
1	88	89	88	88	106	104	101	100	106	103
2	88	89	88	88	94	93	100	99	104	100
3	88	89	88	88	90	89	100	99	100	97
4	86	87	88	88	90	89	94	93	97	94
5	86	87	88	88	88	88	94	93	80	80
6	84	85	86	86	88	88	79	80	80	80
7	84	85	86	86	88	88	78	79	79	79
8	84	85	86	86	86	86	78	79	79	79
9	84	85	86	86	82	82	76	78	79	79
10	82	83	84	84	78	79	72	74	79	79
11	82	83	84	84	78	79	72	74	77	78
12	82	83	83	83	74	75	71	73	77	78
13	80	80	83	83	74	75	71	73	71	73
14	80	80	81	81	72	73	70	72	71	73
15	78	78	80	80	72	73	70	72	71	73
16	78	78	78	78	70	72	70	72	71	73
17	76	76	78	78	68	70	70	72	70	72
18	76	76	78	78	66	68	70	72	70	72
19	64	63	70	70	64	66	70	72	70	72
20	62	60	59	60	62	64	70	72	66	68

*Twenty raw scores were selected from each test by using a table of random numbers.

Table 37 (continued)

*Raw Score	Tests									
	R		S		T		U		V	
	Model	Actual	Model	Actual	Model	Actual	Model	Actual	Model	Actual
1	110	106	97	96	103	104	97	97	96	96
2	100	97	97	96	98	99	92	92	96	96
3	89	87	97	96	95	96	90	90	93	93
4	87	86	96	94	87	88	90	90	92	92
5	87	86	94	93	78	79	88	89	88	88
6	84	83	91	90	78	79	88	89	86	86
7	83	82	87	86	78	79	85	86	85	85
8	83	82	83	83	78	79	85	86	82	82
9	83	82	77	77	78	79	85	86	82	82
10	83	82	77	77	78	79	83	84	81	81
11	81	81	77	77	78	79	80	81	80	80
12	80	80	76	76	77	78	80	81	80	80
13	80	80	76	76	75	76	80	81	77	77
14	73	73	74	75	75	76	73	74	77	77
15	73	73	70	71	70	71	70	71	75	75
16	73	73	70	71	70	71	70	71	74	74
17	71	72	69	70	70	71	70	71	73	73
18	66	67	69	70	70	71	68	69	72	72
19	64	66	69	70	70	71	60	61	70	71
20	64	66	69	70	70	71	60	61	70	71

*Twenty raw scores were selected from each test by using a table of random numbers.

Table 38

Measurement Analysis Model Estimates and Actual Values of Major Item Analysis Components for Teacher-made Tests

Item	Major Item Analysis Components					
	Discrimination Index		Probability Index		Difficulty Index	
	Model	Actual	Model	Actual	Model	Actual
1	.18	.16	.77	.75	.95	.95
2	.28	.61	.88	1.00	.75	.75
3	.27	.66	.88	1.00	.90	.90
4	.33	.66	.92	1.00	.60	.60
5	.27	.27	.88	.88	.90	.90
6	.17	.06	.76	.60	.30	.30
7	-.27	-.28	.12	.12	.90	.90
8	.65	.66	1.00	1.00	.60	.60
9	.16	.20	.75	.80	.60	.60
10	.25	.33	.86	.92	.65	.65
11	.25	.41	.86	.96	.65	.65
12	.27	.66	.88	1.00	.90	.90
13	.25	.28	.86	.88	.35	.35
14	.20	.58	.80	1.00	.80	.80
15	.33	.46	.92	.98	.40	.40
16	.35	.60	.94	1.00	.70	.70
17	.16	.43	.75	.97	.50	.50
18	.35	.60	.94	1.00	.70	.70
19	.00	.00	.50	.50	1.00	1.00
20	.18	.16	.77	.75	.95	.95
21	.52	.70	.99	1.00	.70	.70
22	.49	.56	.99	1.00	.60	.60
23	.00	-.08	.50	.37	.90	.90
24	.27	.19	.88	.79	.90	.90
25	.46	.58	.98	1.00	.75	.75

BIBLIOGRAPHY

BIBLIOGRAPHY

Aleamoni, Lawrence M., and Richard E. Spencer. "A Comparison of Biserial Discrimination, Point Biserial Discrimination, and Difficulty Indices in Item Analysis Data," Educational and Psychological Measurement, XXIX (Summer, 1969), 353-358.

American Association of School Administrators. National Educational Assessment: Pro and Con. Washington, D.C.: National Education Association, 1966.

Ausubel, David P. Educational Psychology--A Cognitive View. New York: Holt, Rinehart and Winston, Inc., 1968.

Beatty, Walcott (ed.). Improving Educational Assessment & An Inventory of Measures of Affective Behavior. Washington, D.C.: Association for Supervision and Curriculum Development, 1969.

Bruning, James L., and B. L. Kintz. Computational Book of Statistics. Glenview, Illinois: Scott, Foresman and Company, 1968.

Colver, Robert M. "Estimating Item Indices by Nomographs," Psychometrika, XXIV (June, 1959), 179-185.

Cox, Richard C. "An Empirical Investigation of the Effect of Item Selection Techniques on Achievement Test Construction." Unpublished doctoral dissertation, Michigan State University of Agriculture and Applied Science, 1964.

Davis, Frederick B. Item-Analysis Data: Their Computation, Interpretation, and Use in Test Construction. Cambridge, Massachusetts: Howard University, 1949.

De Lay, Donald H., and David Nyberg. "If Your School Stinks, CRAM It," Phi Delta Kappan, LI (February, 1970), 310-312.

Diederich, Paul B. Short-cut Statistics for Teacher-made Tests. Princeton, New Jersey: Educational Testing Service, 1960.

Dixon, Wilfred J., and Frank J. Massey, Jr. Introduction to Statistical Analysis. New York: McGraw-Hill Book Company, Inc., 1951.

Downie, N. M. Fundamentals of Measurement: Techniques and Practices. New York: Oxford University Press, 1967.

Ebel, Robert L. "Content Standard Test Scores," Educational and Psychological Measurement, XXII (Spring, 1962), 15-25.

_____ (ed.). Encyclopedia of Educational Research. 4th ed. New York: Macmillan Co., 1969.

Edwards, Allen Jr., and Dale P. Scannell. Educational Psychology--The Teaching-Learning Process. Scranton, Pennsylvania: International Textbook Company, 1968.

Fan, C. T. "Note on Construction of an Item Analysis Table for the High-Low-27-Per-Cent Group Method," Psychometrika, XIX (September, 1954), 231-237.

Ferguson, George A. "The Factorial Interpretation of Test Difficulty," Psychometrika, VI (October, 1941), 323-329.

_____. "Item Selection by the Constant Process," Psychometrika, VII (February, 1942), 19-29.

_____. Statistical Analysis in Psychology and Education. New York: McGraw-Hill Book Company, 1966.

Flanagan, John C. "The Effectiveness of Short Methods for Calculating Correlation Coefficients," Psychological Bulletin, IL (March, 1952), 342-348.

_____. "General Considerations in the Selection of Test Items and a Short Method of Estimating the Product-Moment Coefficient from Data at the Tails of the Distribution," Journal of Educational Psychology, XXX (December, 1939), 674-680.

Flynn, John T., and Herbert Garber (eds.). Assessing Behavior: Readings in Educational and Psychological Measurement. Reading, Massachusetts: Addison-Wesley Publishing Company, 1967.

Gallup, George. "Second Annual Survey of the Public's Attitude Toward the Public Schools," Phi Delta Kappan, LII (October, 1970), 97-112.

Greenfield, T. B. "Administration and Systems Analysis," The Canadian Administrator, III (April, 1964), 25-30.

Guba, Egon G., and Daniel L. Stufflebeam. Evaluation: The Process of Stimulating, Aiding, and Abetting Insightful Action. Monograph Series in Reading Education, No. 1. Bloomington, Indiana: Indiana University Press, 1970.

Guilford, J. P. "The Difficulty of a Test and Its Factor Composition," Psychometrika, VI (April, 1941), 67-77.

_____. Psychometric Methods. New York: McGraw-Hill Book Company, Inc., 1954.

Gulliksen, Harold. "The Relation of Item Difficulty and Inter-Item Correlation to Test Variance and Reliability," Psychometrika, X (June, 1945), 79-91.

_____. Theory of Mental Tests. New York: John Wiley and Sons, Inc., 1950.

Helmstadter, G. C. Principles of Psychological Measurement. New York: Appleton-Century-Crofts, 1964.

Hoyt, Cyril. "Test Reliability Estimated by Analysis of Variance," Psychometrika, VI (June, 1941), 153-160.

Kelley, Truman L. "The Selection of Upper and Lower Groups for the Validation of Test Items," Journal of Educational Psychology, XXX (January, 1939), 17-24.

Kirkpatrick, James J., and Edward E. Cureton. "Simplified Tables for Item Analysis," Educational and Psychological Measurement, XIV (1954), 709-714.

Lazarsfeld, Paul F. and Neil W. Henry (eds.). Readings in Mathematical Social Science. Chicago: Science Research Associates, Inc., 1966.

Lindgren, Henry Clay. Educational Psychology in the Classroom. New York: John Wiley and Sons, Inc., 1967.

Lord, Frederick M. "Sampling Fluctuations Resulting from the Sampling of Test Items," Psychometrika, XX (March, 1955), 1-22.

Luce, R. Duncan. Individual Choice Behavior--A Theoretical Analysis. New York: John Wiley and Sons, Inc., 1959.

Lyerly, Samuel B. "Significance Levels for the Kuder Richardson (21) Reliability Coefficient," Educational and Psychological Measurement, XIX (Spring, 1959), 73-75.

Mayo, Samuel T. "The Methodology and Technology of Educational and Psychological Testing," Review of Educational Research, XXXVIII (February, 1968), 92-101.

McKay, Richard M. "State-wide Information System," Phi Delta Kappan, LI (November, 1969), 178.

Mitzel, Harold E. "The Impending Instruction Revolution," Phi Delta Kappan, LI (April, 1970), 434-439.

Mosteller, Frederick. "On Some Useful 'Inefficient' Statistics," Annals of Mathematical Statistics, XVII (1946), 377-408.

Payne, David A. The Specification and Measurement of Learning Outcomes. Waltham, Massachusetts: Blaisdell Publishing Company, 1968.

Perry, Norman C., and William B. Michael. "The Reliability of a Point Biserial Coefficient of Correlation," Psychometrika, XIX (December, 1954), 313-325.

Shapiro, S. S. and M. B. Wilk. "An Analysis of Variance Test for Normality (Complete Samples)," Biometrika, LII (December, 1965), 591-611.

Smith, Carl B., and Roger Farr. Evaluation Training Package, Bloomington: Indiana University, 1969.

Stufflebeam, Daniel L. "The Use and Abuse of Evaluation in Title III," Theory Into Practice, VI (June, 1967), 126-133.

U. S. Office of Education and National Education Association. "The Magnitude of the American Educational Establishment," Saturday Review (September 19, 1970), p. 67.

Walker, H. M. Elementary Statistical Methods. New York: Henry Holt and Company, Inc., 1943.

Wood, Dorothy A. Test Construction-Development and Interpretation of Achievement Tests. Columbus, Ohio: Charles E. Merrill Books, Inc., 1961.

Wright, Benjamin and Nargis Panchapakesan. "A Procedure for Sample-Free Item Analysis." Paper developed at University of Chicago, January, 1968, Chicago, Illinois. (Mimeographed).

Yost, Earl K., Jr. "Joint Estimation of Mean and Standard Deviation by Percentiles." Unpublished Master's thesis, University of Oregon, 1948.

Zavala, Albert. "Development of the Forced-Choice Rating Technique," Psychological Bulletin, LXIII (August, 1965), 117-124.