DOCUMENT RESUME

ED 224 823 TM 830 005

AUTHOR	Choppin, Bruce; And Others
TITLE	A Critical Comparison of Psychometric Models for Measuring Achievement. Methodology Project.
INSTITUTION	California Univ., Los Angeles. Center for the Study of Evaluation.
SPONS AGENCY	National Inst. of Education (ED), Washington, DC.
PUB DATE	Nov 82
GRANT	NIE-G-80-0112
NOTE	279p.
PUB TYPE	Reports - Research/Technical (143)
EDRS PRICE	MF01/PC12 Plus Postage.
DESCRIPTORS	*Academic Achievement; Achievement Tests; Comparative Analysis; Data Analysis; Item Analysis; *Latent Trait Theory; *Mathematical Models; Psychometrics; *Testing; Testing Problems; Test Theory
IDENTIFIERS	Generalizability Theory; Rasch Model; Three Parameter Model

ABSTRACT

A detailed description of five latent structure models of achievement measurement is presented. The first project paper, by David L. McArthur, analyzes the history of mental testing to show how conventional item analysis procedures were developed, and how dissatisfaction with them has led to fragmentation. The range of distinct conceptual and methodological approaches to achievement testing that now exist are discussed. The second paper, by Kenneth A. Sirotnik, analyzes measurement in achievement as a central and continuing problem in mental testing, highlighting the differences between the modern alternatives. Five papers by David L. McArthur, Bruce Choppin, Ronald K. Hambleton, Rand R. Wilcox and Noreen Webb individually treat Student-Problem (S-P) chart analysis, the Rasch model in item analysis, a three-parameter logistic model, latent class models, and generalizability theory. An analysis of reading comprehension data by four of the contributors and Raymond Moy is presented. J. Ward Keesling presents a summary paper on the empirical work carried out so far in testing different models on common sets of data. (Author/PN)

ED224823

Deliverable - November 1982

METHODOLOGY PROJECT

A CRITICAL COMPARISON OF PSYCHOMETRIC MODELS FOR MEASURING ACHIEVEMENT

Bruce Choppin
Study Director

Grant Number
NIE-G-80-0112, P3

TM 830 005

CENTER FOR THE STUDY OF EVALUATION
Graduate School of Education
University of California, Los Angeles

The project presented or reported herein was performed pursuant to a grant from the National Institute of Education, Department of Education. However, the opinions expressed herein do not necessarily reflect the position or policy of the National Institute of Education, and no official endorsement by the National Institute of Education should be inferred.

A Critical Comparison of Psychometric Models for Measuring Achievement

TABLE OF CONTENTS

INTRODUCTION AND OVERVIEW

Bruce Choppin
Center for the Study of Evaluation, UCLA

The research project reported here developed out of a growing concern at the fragmentation that is occurring within the psychometric field. Dissatisfaction with the limitations inherent in traditional forms of mental test analysis (as typified by the norm-referenced multiple-choice test of achievement), has led in recent years to a variety of new psychometric theories and procedures. The traditional approach to testing was developed in order to provide ranking of students and/or to select relatively small proportions of students for special treatment. In these tasks it was fairly effective, but it is increasingly seen as inadequate for the broader spectrum of questions that educational measurement is now called upon to address. Novel applications have stimulated new psychometric models and methods, each shaped to deal with the specific problems of the particular situation. The last two decades have seen the development of new types of tests, new scoring methods, new procedures for item analysis, and entirely new conceptions of the mental measurement process.

A marked characteristic of the professional literature on these novel approaches to measurement is its parochialism. Many of the most prolific psychometricians display little interest in models other than their own, and there have been few, and mostly inadequate, attempts to integrate theories and results. The proponents of different models have different objectives; implicitly or explicitly they make differing assumptions; and they frequently use the same words and phrases to mean different things (e.g., reliability, accuracy, guessing, error and true-score). Separate methodologies based on different models have diverged to a point where it is no longer possible to identify a mainstream approach to educational measurement, and where informed and balanced advice on the full range of alternative approaches is almost impossible to obtain.

The present project was designed to take advantage of the wide range of interest and experience of different approaches to measurement jointly held by the professional researchers who constitute the "methodology group" at CSE. The project had two related goals. The first was to document in some detail the philosophy, assumptions, mathematical procedures, advantages, limitations, etc. of each of five different approaches to the measurement of achievement that currently command considerable psychometric interest. We have tended to describe these five approaches as alternative models of achievement measurement, and in the strict scientific sense this is true, though a comprehensive mathematical formulation is easier for some than for others. This detailed documentation would enable us to clarify our understanding of the similarities and differences among the models so that we might explore with real data the consequences of adopting one analytic strategy rather than another.

The second purpose,arising from the first, was to develop a much needed "user's quide" that would set out, fairly and comprehensively, the rationale underlying each of the separate approaches and provide sound advice to the potential user regarding the selection of an approach and how these models may be operationalized.

The models we consider all belong to the class of latent structure models in that their analysis is directed to the inferential classification of test items and/or persons, based on theoretical assumptions concerning the structure of test data and conceptual theories of measurement. Within this framework, the different models may be seen as attempts at the solution of a variety of measurement problems. Sometimes, even when the models or procedures appear similar, the issues of central concern to one may not be of any particular interest to the other. In the measurement area, we meet variations in philosophy and value systems as well as in statistical referents.

A qood example of this can be found in the recent controversy over latent trait models. Although the Rasch one-parameter model and the three-parameter model developed by Birnbaum and Lord appear to have a lot in common (the Rasch model is mathematically a special case of Lord's model) they are conceptually quite distinct. Lord began some thirty years ago with large quantities of item response data which he wished to understand and explain. For him it was important to find a model that fitted his data and could make sense of it. Today his disciples view the Rasch model as a model that does not fit their data well. It is founded on assumptions (e.g., no guessing) which are often not met in practice. This group of measurement specialists rightly discard the (inexpensive) Rasch model in favor of a more complex analysis that better meets their need to "fit" data. On the other hand, Rasch was developing his model (during the 1950's), not on the basis of actual test data, but rather on a series of principles and axioms for measurement systems that he extracted from other realms of scientific experience. He did not create his model primarily to explain existing data sets, but rather to form the basis for constructing new measurement systems. For his followers, <u>test items must "fit" the model</u> if they are to be useful for measurement. The goal is to find items that do fit the model so as to permit the construction of test instruments with the optimal properties that Rasch described.

Unfortunately, many psychometricians in each camp have not been able to appreciate the distinction between these two approaches. There have been public debates during with Item Response Theorists have condemned the Rasch model for not "fitting" real data, while the Rasch practitioners attack Item Response Theory for dealing with models whose parameters cannot be satisfactorily estimated and which do not satisfy the requirements for "objective measurement". The criticisms are sound in themselves, but they do not relate to the issues that the other side holds to be important.

There are other, though perhaps less dramatic, examples of where different priorities and different concerns have led to some breakdown in communication. For example, Generalizability Theory is directly concerned with measures, and with analyzing the "errors" associated with them. However, it treats these on a grouped basis as "error variance" and makes certain assumptions about their distribution. By contrast, latent trait theorists use "standard error of measurement" on an individual basis, finding it to be a more useful concept than the conventional one of test reliability. Latent trait theorists also make assumptions about the distribution of these errors, and in general these assumptions are not compatible with those of G-theory. Both approaches are useful for solving specific measurement problems, but their areas of application are very different. The extent to which the two approaches may be regarded as complementary, and may indeed support one another, is not well understood.

Our work has addressed these and other questions. We have brought some illumination to previously dark and shaddowy areas where two or more of the models come together.

However, we do not feel that we have yet reached our second objective of developing a comprehensive and useful guide to practice. More empirical work in comparing the effects of the different models needs to be done, and the handbook we wish to develop will contain more demonstrations using real data than are found in this report. There has not been time in the last twelve months to carry out as much of this work as we would have liked, but we feel that we are on the right track and that our work is sufficiently important for its completion to be given some priority.

The format of the present report is described below. There are two introductory chapters. The first analyzes the history of mental testing to show how conventional item analysis procedures were developed (in response to which pressures and constraints), and how dissatisfaction with them has led to fragmentation and the range of distinct conceptual and methodological approaches to achievement testing that now exist. The second paper analyzes in depth a central and continuing problem in mental testing, and one which not merely illustrates the shortcomings of the traditional approach, but highlights the differences between the modern alternatives.

There follow five papers treating each of the selected approaches individually but according to a standard format.

These "models" are: the S-P Chart Analysis developed by Sato which may be viewed as a simplified form of Guttman scaling; two latent trait logistic models (Rasch with one item parameter and Lord with three item parameters) given separate treatment because of the philosophical and conceptual contrast cited above; a latent class model to which the estimation of true scores is central; and Generalizability Theory which, though somewhat different in scope from those mentioned earlier, offers a different mathematical model for test data, and some powerful statistical procedures for interpreting them.

Following this we present a summary of the empirical work carried out so far in testing out different models on common sets of data.

In conclusion, a chapter (available only in outline at the present) summarizes and synthesizes the earlier parts of the report and draws some definitive conclusions regarding the applicability of the various models to different measurement problems.

EDUCATIONAL TESTING AND MEASUREMENT: A BRIEF HISTORY

David McArthur
Center for the Study of Evaluation, UCLA

Educational assessment in the Western tradition has a long but very irregular history. Seven centuries ago, one English college was deemed remiss in its responsibilities because its founder had determined that its recent graduates "...expressed themselves very inaccurately in the learned languages..." (Sylvester, 1970, p.19) the method of such determination was not described. A tradition of oral examinations was built up over several centuries, only to disintegrate almost completely by the time Isaac Newton attended college about 1660; not only were there no examinations but frequently the lecturers themselves simply never showed up for classes. However, in another hundred years, both Oxford and Cambridge, recognizing the deteriorated situation, decided to improve their curriculum and instituted regular written examinations in a variety of topics. The exams of this era were almost exclusively essay questions emphasizing factual recall; one extant example shows eight questions each in history and geography, and six in grammar, primarily Latin and Greek. In the education of the younger pupils, examinations began to become more prevalent as textbooks for the grammar school came to be formulated into distinct grade levels.

> The new sequences of textbooks allowed a more precise grading to be implemented in schools in various parts of Europe...Within the school a further step was the development and application of the principal of a child's regular progression through grades at various intervals of about a year (Bower, 1975, p.419).

The Jesuits, finding that such a procedure fit perfectly into their concept of ratio (the systematically ordered body of knowledge) took up the idea with vigor, and it rapidly spread across Europe.

Meanwhile, in China, civil service examinations were already several millenia old. The earliest proficiency testing dates from 2200 B.C., and formal procedures for examination date from 1115 B.C. Despite a concentration on literary rather than managerial skills, the system was to be the model for a number of efforts at standardizing competition for civil service positions in Europe and the U.S. during the 19th century. But in China the testing system was abolished in reforms at the beginning of the 20th century, as Western technologies and educational orientations intruded into the Orient (DuBois, 1964, 1967).

In the United States, it was not until 1845, following Horace Mann's advocacy of written examinations, that testing was incorporated into educational practice. The first recorded examination was administered in Boston that year, and the concept took hold quickly (Englehart, 1950). Within thirty-five years, promotion from grade to grade was no longer made by personal recommendation but instead invariably was judged by success or failure, scored as a percentage, on a written exam. Mann's viewpoint of testing, while not using the word "objective," carried with it a decided bias towards objective measurement and standard tests (Ruch, 1929). The earliest objective educational tests are found in a book complete with questions, answers and scales, by an English schoolmaster, dated 1864 (Kelley, 1927). Objective tests in spelling and arithmetic were in place in the U.S.

by the 1870's. Then, in 1881, the superintendent of schools in Chicago, expressing a strong sentiment against testing in particular (if not against science in general) decreed that advancement of students was to be carried out only by direct recommendations of teachers and principals. Testing for purposes of grade-level advancement was prohibited. His viewpoint was widely shared; suddenly, the impetus for "objective" measurement and assessment was on the wane. "Examinations for grade promotions were gradually abolished in all the best schools," claimed the superintendent's successor. "The person best qualified to judge of a child's ability to go on is his teacher...To say that any other test is necessary is a travesty on common sense" (Bright, 1895, pp.274-275). By the end of the nineteenth cnetury, educational testing had achieved a bad name. Teachers were "teaching on the test," devoting weeks of preparation and drill to extant editions of upcoming exams, and the public was not pleased.

A completely separate thread in the fabric of educational measurement is found in a review of the history of statistics. The first lectures in statistics date around 1660; the first use of the word "statistic" is placed at 1749, in reference to the accounting of all the things that make up a kingdom (Meitzen, 1891). While extensive developments in mathematics were being made during this time (Newton, for example, was solving problems in differential calculus by 1676), the setting out of facts and figures in the social sciences for many years was limited to tabulations of various facts, actuarial tables, and census taking, the first about 1769 in Denmark.

Interestingly, some recognition of the importance of understanding individual differences in mental abilities is found in the field of astronomy by 1822 (Freeman, 1926). It was not until this century that the word "statistics" came to refer exclusively to quantitative approaches; its origins apparently are tied to the Germanic discipline called "Staatenkunde" or study of governments and politics. The profession suffered a decline as the old teachers passed away, and the task of statistics was made increasingly narrow.

> In 1806 and 1807 a passionate controversy arose against the brainless bungling of the number statisticians, the slaves of the tables, the skeleton-makers of statistics...The opponents in the sharp attack were themselves, however, not sufficiently clear how new and precise limits for their science should be determined. (Meitzen, 1891, pp.49-50).

An International Statistical Congress was formed to attempt to resolve the confusion; it met first in 1853 and showed a surprising degree of success. Even though its members chose to stay out of issues of statistical theory, in 1869 one of their resolutions declared:

> ...that in all statistical researches it is important to know the number of observations...; the qualitative value is to be measured by the divergences of the numbers among themselves as well as the average...; it is desirable to calculate...the average deviations (Meitzen, 1891, p.80).

These principles formed the basis for technical developments in educational statistics into the twentieth century: one of the first texts (Rugg, 1917) devoted most of its efforts to tabulation, averages, frequencies and variabilities. Despite several pioneering studies in educational attainment, in large measure the collection and analysis of data at this time was confined to tabulations of school attendance and costs. The statistical societies of the day were deeply embroiled in social problems, especially the relations of education to crime, and spent no time at all on assessing educational achievement beyond such indices as the ability to sign one's own name (Cullen, 1975).

By the middle of the nineteenth century, considerable progress had been made in the analysis of experimental data from agricultural research. Good experimental designs, including factorial and split-plot techniques, were in place about 1850. Galton spent time investigating how mathematical solutions might best be developed for data from studies of Charles Darwin, building a number of statistical tools in the process, and was the first to attempt measuring characteristics of individual intelligence (1883). But it was not until Pearson's chi-square test (1900), and Student's t-test (1908) that appropriate quantification of educational data could be developed, although the latter, surprisingly, took a number of years to catch on (Cochran, 1976). Fisher's analysis of variance (1924) drew heavily on these precursors but it too was relatively slow in being incorporated into the repertoire of educational statisticians. Guilford's text on fundamental statistics in 1942 awards analysis of variance fewer than nine pages, embedded in a chapter on reliability.

In 1890 appeared the first study of reliability (Edgeworth, 1890). In the same year the seminal short article by Cattell (1890) marked the first time the words "mental tests" were used together. Following Galton's lead, several investigators in Germany began to develop mental tests, and in the U.S. there was extensive interest in the relationship of mental capacities to physical characteristics. The American Psychological Association set up a standing committee in

1895 to consider cooperative efforts in mental and physical statistics;.the American Association for the Advancement of Science did likewise the following year. Binet, who had been working on problems in mental reasoning since 1886, wrote an important article in 1898 on the utility of measurement and scaling in the appraisal of human intelligence. However, two major studies of testing around this time (Sharp, 1899; Wissler, 1901) concluded that many of the available tests used for psychological research fell far short of their claims, in both content and method (Peterson, 1925). In education, Rice's (1897) study of spelling attainment, using a single list of 50 words in a test administered to 30,000 children, was a pioneering study, which circulated widely but gained few supporters (Wilds & Lottich, 1970).

About the turn of the century there was a fair degree of public discouragement about educational testing. **However, about this time,** the first survey of school facilities and educational **practice was** conducted, the College Entrance Examination Board was established, and in 1902 the first course in educational measurement was taught (by Thorndike at Columbia) (Meyer, 1965). Concurrently, interest in the concept of general intelligence was being pursued by a number of investigators, following a suggestion by Galton in 1883 and a study of 1,500 children conducted in 1891 (Burt, 1909). In the analysis of results from the latter investigation, however, came the explicit realization that statistical methods for educational measurement were in desperate need of thoughtful improvement. Burt speculated that the consistent failures of research investigations in the area of general intelligence before the turn of the century

> were largely due to their reliance for discovery of correlations upon mere inspections of the data they obtained, instead of upon quantitative determination and mathematical deduction (pp.94-95).

During the first decade of the twentieth century, the growing impetus for increased statistical rigor could be felt in several areas; measurement successes in anthropometry and biology provided much needed support for such improvement. In 1904, Toulouse and Pieron's two volume manual on laboratory experiments included sections on intelligence and the measurement of individual differences. In 1906 the American Psychological Association created a permanent committee charged with evaluating requirements for standard laboratory technique and appraising both group and individual tests with attention to practical applications. Binet's test for intelligence (1905) and Thorndike's book on mental measurement (1904) had particular significance during this time, as did Spearman's (1904) paper on general intelligence. By 1910, a vast number of tests in skills like English, spelling, handwriting, reading and arithmetic had emerged, followed closely by more technical articles on topics like numerical analysis, standardization, validity and correlations.

> ...American educators quickly realized that the scale idea could be applied not only to intelligence but to achievement as well. There followed a phenomenally creative period during which testmakers developed instruments for virtually every aspect of educational practice (Cremin, 1961, p. 186).

In 1913, the National Council of Education released a major report on standards and tests for measuring school efficiency, and expressed this sentiment:

> We are only begining to have measurement undertaken in terms of standards or units which are, or may become, commonly recognized. Such standards will undoubtedly be developed by means of applying scientifically derived scales of measurement to many systems of schools. From such measurements it will be possible to describe accurately the accomplishment of children and to derive a series of standards...(Strayer, 1913, p.4).

Graves, reviewing the condition of education in 1913, expressed the sentiment that the application of mathematics to measurements in education was one of the most significant movements of that time.

Developments in objective measurement of intelligence and educational achievement came to a head with the crisis of the Great War. Work in Germany on the screening of inductees had been in progress since 1905; Binet and Simon (1910) discussed the application of intelligence testing in the French army (Peterson, 1925). In the U.S., Terman's revision of the Binet scale was completed by 1917, and was applied soon thereafter to the testing of 1.7 million recruits. A small team of educational psychologists produced the Army Alpha and Beta tests of intelligence between May 28 and June 10, 1917; a copy of the examiner's manual was enroute to the printer within a month. Immediately after the war, as the Army was selling thousands of unused test blanks, both educational specialists and the public began to realize that objective test results had to be taken with some degree of caution. One of the originators of the Army Alpha expressed the sentiment unambiguously: "We do not know what intelligence is and it is doubtful if we will ever know what knowledge is" (Goddard, 1922, quoted in Spring, 1972, p.5). Even so, by 1920, objective testing formed the core of educational assessment methods. The Journal of Educational Measurement devoted several issues in 1921 to a symposium on scientific measurement of intelligence.

During the decade that followed, the objective assessment of intelligence "swept America, and to a lesser extent Canada, like an educational crusade...The critics were numerous but few in comparison to the advocates..."(Marks, 1976, p.10). McCall's (1922) book on educational measurement and Monroe's (1923) the following year were the first to set out the procedures for a "new type examination," the multiple-choice and true-false tests. Principles of test construction began to earn chapters of their own, and the variety of interpretations and uses of tests was becoming a major consideration for many educators (Monroe, 1945). Then came the first contributions to what is now recognized as classical test theory: Thurstone's (1925, 1926, 1927) articles on the scoring of individual performance, Ruch and DeGraff's (1926) study of corrections for guessing, Ruch's (1929) The Objective or New Type Examination, and Thurstone's The Reliability and Validity of Tests, 1931.

The concept of reliability is illustrative of the historical development of educational measurement. Because of its basis in correlational method, which was already well advanced at the turn of the century, a number of technical articles appeared quite early concerning the statistical nature of reliability indices. By the time that a major study was launched in the late 1920's by the American Historical Association's Commission on the Social Studies into the nature of testing in social sciences education, reliability measures were regarded as essential by technical specialists but generally disregarded by practitioners. Under the counsel of Truman Kelley, a large-scale investigation was conducted on the use of tests for determining overall class and school performance, recognizing individual skill levels and individual differences, and appraising attitudes and personality traits. It also studied the utility of the "new-type" tests. In the long run both the social science specialists and the educational measurement technicians were disappointed in the results of the study. The former were not pleased by the tendency of short-answer and multiple-choice tests towards fragmentary presentation of, and limitations to, simple facts in the curriculum and the deletion of shades of meaning. The latter felt that lack of objective terms, which they saw as essential for objective measurement, obviated the study's conclusions. Kelley's feelings were sufficiently strong that he wrote a 15-page appendix entitled "A Divergent Opinion as to the Function of Tests and Testing" in which he excoriated the opponents of testing with more than a dozen carefully reasoned arguments regarding the appropriate scientific use of educational tests, plus one or two direct strikes to the more emotional nature of the argument:

> The opponents (of testing) show no awareness of the tests of reliability and validity of measuring instruments, either judgments of teachers or of test scores. We believe that such awareness is essential to any educator who is not content to work in the dark (p. 489).

In the areas of reliability and validity, technical proofs were available as early as 1910 (Spearman, 1910) providing a rationale behind error measurement and Brown (1910) giving a definition of true score. But it was some time before either term was given serious treatment in the standard texts. Taking a representative contribution from each decade, we find a half-dozen index entries in Rugg's 1917 text, 18 entries between the two in Ruch's 1929 text, four chapters in his 1942 book, and eight full chapters devoted to the two topics in Gulliksen's 1950 text. However, by the 1930's there had accumulated a variety of estimation procedures and a great deal of confusion of terms (Adams, 1936; Barthelmess, 1931; Lincoln, 1932). An attempt to resolve the issues was made in Thurstone's small book on the topic in 1931, another in Kuder and Richardson's (1937) key article in test reliability, followed by Guttman's (1945) reformulation and Cronbach's (1947) discusion of the several different kinds of reliability coefficients. The American Psychological Association tried to resolve the various discrepancies by committee in 1954. Tryon (1957) provided an extensive historical review of the reliability concept and a domain-sampling reformulation. "The extraordinarily massive literature in this topic," wrote Cattell (1964), "...has never lacked statistical finesse and mathematical virtuosity (p.1)", but he, too, felt a need to suggest substantial redefinitions for both reliability and validity, which in turn were ignored four years later with publication of a definitive mathematical analysis by Lord and Novick (1968).

The first formulations of a 'sample-free' approach to mental measurement are found in Lawley's (1943) analysis of item selection. Although the problem had been explored tangentially by Horst (1936) and more recently by Ferguson (1942), his paper was among the earliest to seek mathematically rigorous justifications for the selection of maximally discriminating test items, and to examine in some detail the concept of item characteristic curves. Tucker (1946) provided further statistical support. Gulliksen (1950) summarized the early work in true score theory, and Lord explored the application of latent trait theory to test theory with his doctoral dissertation, published as

Theory of Test Scores (1952). Interestingly, he felt that the actual utility of large portions of the theory would be limited in practice by the difficulty in obtaining sufficiently large data sets, and did not publish about the problem again for another ten years. At that point he presented an important development, the beta-binomial model of the frequency distribution of true scores and raw scores (Keats & Lord, 1962), and further refined the definition of true scores in Lord & Novick (1968). Meanwhile, Birnbaum explored certain statistical properties of normal and logistic characteristic functions in 1957 and 1958, but few other papers on this topic appeared until the 1960's.

The sentiment has been expressed more than once that the science of educational testing has progressed fitfully. Despite a plethora of statistical developments, "most of the major theoretical and technical distinctions and most of the principle points of dispute were in existence by 1925" (Thomson & Sharp, 1983). This includes such diverse topics as item analysis, test bias, the nature vs. nurture arguments regarding individual intelligence, and at least the beginnings of factor structure explanations for educational assessment.

REFERENCES

Adams, H. F. Validity, reliability and objectivity. In W. R. Miles (Ed.), Psychological studies of human variability, Psychological Monographs, 1936, 57, 329-350.

Barthelmess, H. M. The validity of intelligence test elements. New York, Teachers College, 1931.

Binet, A. La mesure en psychologie individuelle. Revue Philosophique, 1898, 46, 113-123.

Binet, A., & Simon, T. Methodes nouvelles pour le diagnostic scientifique des etats inferieurs de l'intelligence. L'Annee psychologique, 1905, 11, 163-190.

Binet, A., & Simon, T. Sur la necessite d'une methode applicable au diagnostic des arrierees militaires. Annales medico-psychologique, 1910.

Birnbaum, A. An efficient design and use of tests of a mental ability for various decision making problems. Series Report No. 58-16, USAF School of Aviation Medicine, Randolph, Texas, 1957.

Birnbaum, A. On the estimation of mental ability. Series Report No. 15, USAF School of Aviation Medicine, Rndolph, Texas, 1958.

Bower, J. A history of western education. Civilization of Europe, sixth to sixteenth century, vol. 2. New York: St. Martin's Press, 1975.

Bright, O. T. Changes - wise and unwise - in grammar and high schools. In National Education Association Journal of Proceeding and Addresses, St. Paul, NEA, 1895.

Brown, W. Some experimental results in the correlation of mental abilities. British Journal of Psychology, 1910, 3, 296-322.

Brown, W., & Thompson, G. H. The essentials of mental measurement. Cambridge, Mass: Cambridge University Press, 1940.

Brownless, V. T. & Keats, J. A. A retest method of studying partial knowledge and other factors influencing item response. Psychometrika, 1958, 23, 67-73.

Burt, C. Experimental tests of general intelligence. British Journal of Psychology, 1909, 3, 94-177.

Cattell, J. M. Mental tests and measurements. Mind, 1890, 15, 373-381.

Cattell, R. B. Validity and reliability: A proposed more basic set of concepts. Journal of Educational Psychology, 1964, 55, 1-22.

Cochran, W. G. Early development of techniques in experimentation. In D. B. Owen (Ed.) On the history of statistics and probability. New York: Dekker, 1976.

Cremin, L. The transformation of the school. New York: Knopf, 1961.

Cronbach, L. J. Test "reliability": Its meaning and determination. Psychometrika, 1947, 12, 1-16.

Cronbach, L. J. Five decades of public controversy over mental testing. American Psychologist, 1975, 30, 1-14.

Cullen, M. J. The statistical movement in early Victorian Britain: The foundations of empirical social research. New York: Barnes & Noble, 1975.

DuBois, P. H. A test-dominated society: China, 1115 B.C.-1905 A.D. ETS Invitational conference on testing problems, Princeton: ETS, 1964.

DuBois, P. H. A history of psychological testing. Boston: Allyn and Bacon, 1970.

Edgeworth, F. Y. The element of chance in competitive examinations. Journal of the Royal Statistical Society, 1890, 53, 460-475, 644-673.

Englehart, M. D. Examinations. In W. S. Monroe (Ed.), Encyclopedia of educational research. New York: MacMillan, 1950.

Ferguson, G. A. Item selection by the constant process. Psychometrika, 1942, 7, 19-29.

Fisher, R. A. Statistical methods and scientific inference. New York: Hafner, 1956.

Freeman, F. N. Mental tests: Their history, principles and applications. Boston: Houghton Mifflin, 1926.

Graves, F. P. A history of education in modern times. New York: MacMillan, 1950.

Goodenough, F. L. A critical note on the use of the term 'reliability' in mental measurement. Journal of Educational Psychology, 1936, 27, 173-178.

Goodenough, F. L. Mental testing, its history, principles and applications. New York: Rinehart, 1949.

Guilford, J. P. Psychological measurement a hundred and twenty-five years later. Psychometrika, 1961, 26, 109-127.

Gulliksen, H. The content reliability of a test. Psychometrika, 1936, 1, 189-194.

Gulliksen, H. Measurement of learning and mental abilities. Psychometrika, 1961, 26, 93-107.

Gulliksen, H. Theory of mental tests. New York: Wiley, 1950.

Guttman, L. A basis for scaling qualitative data. American Sociological Review, 1944, 9, 139-150.

Hambleton, R. K., & Cook, L. L. Latent trait models and their use in the analysis of educational test data. Journal of Educational Measurement, 1977, 14, 75-96.

Horst, A. P Item selection by means of maximizing function. Psychometrika, 1936, 1, 229-244.

Keats, J. A., & Lord, F. M. A theoretical distribution for mental test scores. Psychometrika, 1962, 27, 59-72.

Kelley, T. L. Interpretation of educational measurements. Yonkers-on-Hudson, New York, 1927.

Kelley, T. L., & Krey, A. C. Tests and measurements in the social sciences. Report of the Commission on the Social Studies, American Historical Association, Part IV. New York: Charles Scribner's Sons, 1934.

Kuder, G. F., & Richardson, M. W. The theory of the estimation of test reliability. Psychometrika, 1937, 2, 151-160.

Lawley, D. N. On problems connected with item selection and test construction. Proceedings of the Royal Society of Edinburgh, 1943, 61, Section A, 273-287.

Lazarsfeld, P. F. Latent structure analysis and test theory. In H. Gulliksen and S. Messick (Eds.), Psychological scaling: Theory and applications. New York: Wiley, 1960.

Lazarsfeld, P. F. The logical and mathematical foundations of latent struture analysis. In S. A. Stouffer, et al (Eds.), Measurement and prediction. Princeton: Princeton University Press, 1950.

Lentz, T. F., Hirshstein, B., & Finch, F. H. Evaluation of methods of evaluating test items. Journal of Educational Psychology, 1932, 23, 344-350.

Lincoln, E. A. The unreliability of reliability coefficients. Journal of Educational Psychology, 1932, 23, 11-14.

Lord, F. M. A theory of test scores. Psychometric Monographs, No. 7, 1952.

Lord, F. M., & Novick, M. R. Statistical theories of mental test scores. Reading, Mass.: Addison-Wesley, 1968.

Macready, G. B., & Dayton, C. M. The use of probabilistic models in the assessment of mastery. Journal of Educational Statistics, 1977, 2, 99-120.

Marks, R. Providing for individual differences: A history of the intelligence testing movement in North America. Interchange, 1976-1977, 7, 3-16.

McCall, W. A. How to Measure in Education. New York: Macmillan, 1922.

Meitzen, A. History, theory, and technique of statistics. Philadelphia, 1891.

Meyer, A. E. Educational history of the western world. New York: McGraw Hill, 1965.

Monroe, W. S. Introduction to the theory of educational measurement. Boston: Houghton Mifflin, 1923.

Monroe, W. S. Educational measurement in 1920 and 1945. Journal of Educational Research, 1945, 38, 334-340.

Peterson, J. Early conceptions and tests of intelligence. Yonkers-on-Hudson, New York: World, 1925.

Rasch, G. Probabilistic models for some intelligence and attainment tests. Copenhagen, Denmark: Neilsen & Lydiche, 1960.

Rice, J. M. Forum, 1897. Cited in W. H. Wilds & K. V. Lottich, Foundations of modern education. New York: Holt, Rinehart & Winston, 1970.

Ruch, G. M. The objective or new-type examination, an introduction to educational measurement. Chicago: Scott, Foresman, 1929.

Ruch, G. M., & deGraff, M. H. Corrections for chance and "guess" vs. "do not guess" instructions in multiple-response tests. Journal of Educational Psychology, 1926, 17, 368-375.

Rugg, H. O. Statistical methods applied to education. Boston: Houghton Mifflin, 1917.

Sharp, S. E. Individual psychology: A study in psychological method. American Journal of Psychology, 1899, 10, 329-391.

Spearmen, C. Correlation calculated from faulty data. British Journal of Psychology, 1910, 3, 271-295.

Spearman, C. General intelligence objectively determined and measured. American Journal of Psychology, 1904, 15, 201-292.

Spring, J. H. Psychologists and the war: The meaning of intelligence and the Alpha and Beta tests. History of Educational Quarterly, 1972, 12, 3-15.

Strayer, G. D. Standards and tests for measuring the efficiency of schools or systems of schools. Bulletin, United States Bureau of Education, 1913, Whole No. 13: Report of the Committee of the National Council of Education.

Sylvester, D. W. Educational documents 800-1816. London: Methuen, 1970.

Thompson, G. O. B., & Sharp, S. History of mental testing. In T. Husen & N. Postlethwaite (Eds.), International encyclopedia of education: Research and studies, Oxford: Pergamon Press, 1983.

Thorndike, E. L. An introduction to the theory of mental and social measurements, 1904.

Thurstone, L. L. A method of scaling psychological and educational tests. Journal of Educational Psychology, 1925, 16, 433-451.

Thurstone, L. L. The reliability and validity of tests. Ann Arbor: Edwards, 1931.

Thurstone, L. L. The scoring of individual performance. Journal of Educational Psychology, 1926, 17, 446-457.

Thurstone, L. L. The unit of measurement in educational scales. Journal of Educational Psychology, 1927, 18, 505-524.

Toulouse, E., & Pieron, H. Technique de psychologie experimentale. Paris: Doin, 1904.

Tryon, R. C. Reliability and behavior domain validity: Reformulation and historical critique. Psychological Bulletin, 1957, 54, 229-249.

Tucker, L. R. Maximum validity of a test with equivalent items. Psychometrika, 1946, 11, 1-13.

Wilds, E. H., & Lottich, K. V. Foundations of modern education. New York: Holt, Rinehart & Winston, 1970.

Wissler, C. The correlation of mental and physical tests. Psychological Review, Monograph Supplement Vol. 8, No. 16, 1901.

Yerkes, R. M. (Ed.) Psychological examining in the United States Army. Memoirs of the National Academy of Sciences, 1921, 15, 1-890.

TOWARDS MORE SENSIBLE ACHIEVEMENT MEASUREMENT: A VIEW AND REVIEW

Kenneth A. Sirotnik
Center for the Study of Evaluation, UCLA

Introduction

Much of what will follow here is a repeat of an unfamiliar--or at least unpopular--theme. The essence of this theme has been either implicit or explicit in writings dating as far back as the early 1930's and continuing up to the present. (See, for example, Walker, 1931; Guttman, 1944; Loevinger, 1947, 1948, 1954; Rasch, 1960; Lumsden, 1961; Bentler, 1971; and Wright and Stone, 1979.) Probably the most entertaining and insightful review is a rarely quoted article by Lumsden (1976). These authors all propose different techniques (or variants of the same techniques) and analytic models for scaling the items on the ordinary test of achievement. But they all have two basic things in common: (1) they are critical of, and represent alternatives to, classical test theory and (2) they operate from fundamentally the same notion of what it means to measure. The essence of the common theme is, bluntly, that classical (and classical-like) test theories are not very useful when it comes to test construction and analysis.

Why has not the nearly exclusive practice of traditional[1] test theory methods abated during the last fifty years? Why does nearly every new issue of journals like Psychometrika or Educational and Psychological Measurement contain yet another theoretical exposition involving true and error score theory or some esoteric reformulation of the same old reliability coefficient? Were the above authors and others like them just on a flight of fancy proposing crazy ideas that happened to escape the eyes of critical reviewers? No! They merely challenged what to date[2] amounts to over 70 years' worth of archives of scholarly work on test theory models bearing little resemblance to how people ordinarily think about what it really means to measure. To be sure, each challenge did not offer a completely viable alternative to common practice. But it seems to be part of the human condition to hang on tenaciously to the familiar, to the security of a large investment, at least until the market crashes and/or the tide of opinion noticeably changes through the power of advertisement.

Such has been the case recently with the increased use of latent trait models, particularly the model proposed by Rasch (1960) and popularized in the U. S. by Wright (1968, 1969 [with Panchapakeson], 1977, and 1979 [with Stone]). The point of this report is not, however, to advertise any particular measurement model. Rather, I wish to continue advertising the self-evident notion that how one conceptualizes the act of measurement should have a lot to do with how one analyses the quality of the measurement act during its development, implementation and revision phases.

I will restrict this discussion to the measurement of achievement with items of the usual correct-incorrect (1-0) variety. (However, the basic notions are generalizable to ordered response scales more typical in the measurement of values, attitudes, beliefs, opinions, etc.) My point of view regarding how the measurement act is ordinarily conceptualized is not original nor very creative. It rests simply on analogy with measurement in the physical sciences where constructs are often experienced with the senses. The measurement of length, in particular, a person's height, is the usual example and will serve well here. Certainly most constructs we attempt to measure in the behavioral sciences are not directly experienced and this, of course, constitutes the main source of difficulty. But it does not follow, necessarily, that the generic notions of measurement be any different. Nor does it follow that measurement models be deterministic, i.e., be developed in ideal terms from which deviations are unaccounted for. Probabilistic models are those wherein all deviations from the model have an expected probability of occurance. Both deterministic and probabilistic models exist in both the physical and behavioral sciences.

Implicit in this view of measurement is an assumption that the test items are all measuring the same thing (construct, trait, etc.). Extant psychometric literature is replete with confusion over what exactly is meant by this assumption and the two commonly used terms -- unidimensional and homogeneous -- referencing sometimes similar and sometimes dissimilar empirical interpretations of this assumption. The confusion, not surprisingly, reduces down to different views of the measurement act. Viewed in its original factor analytic sense, unidimensionality refers to one interpretable common factor explaining the item correlation matrix. This fits well with the notion of measurement as repeated single-item tests and the concept of reliability as internal consistency. But internal consistency is only a necessary and not a sufficient condition for a single common factor in an item set; yet, many traditional test theorists (e.g., Gulliksen, 1950; Ghiselli, 1964; Magnusson, 1966; and Allen and Yen, 1979) and practitioners have used both unidimensionality and homogeneity in reference to the internal consistency of a set of items.

To confuse the issue further, Guttman's (1944) "unidimensionality" and Loevinger's (1947) "homogeneity" both, in empirical consequence, refer to the cumulative ordering or scaling of a set of items -- a fundamentally different notion of the use of items to measure a single construct. The analogue of this notion for probabilistic models (e.g., latent class and latent trait models) is the concept of local independence, taken by many latent trait theorists (e.g., Lord & Novick, 1968: Hambleton & Cook, 1977; and Lord, 1980) as the equivalent of the assumption of unidimensionality. (But see the discussion of Traub and Wolfe, 1981, p. 387.)

From my point of view, I assume that there exist sufficiently singular achievement constructs, represented by item sets, that are psychologically interpretable and that are of potential instructional use. A reasonably successful application of a measurement strategy is necessary but not sufficient evidence for a reasonably successful effort at measuring a singular construct. In other words, a singular construct is assumed at the outset; a priori verification of the assumption, is, in essence, an exercise in content validity; necessary a posteriori evidence lies, in essence, in the degree of success in developing the measurement device; sufficient evidence, however, is accumulated only through further construct validation studies.

In what follows, a common conceptual view of the act of measurement will be presented and contrasted, in general, with the act as implied by traditional test theories. This discussion will then be punctuated by a more specific overview of several traditional test theories to illustrate the issue further. Finally, alternative models will be reviewed which are more in line with how the measurement act is ordinarily conceived.[3]

Precision and Accuracy: Disentangling the Concepts Measurement and Dependability[4]

It is important, first, to define measurement more explicity. Many definitions have been proposed resulting in disputes over what does and does not constitute measurement. My interest is not to debate the issue at a philosophical level, but rather to simply clarify how the term will be used here. It will serve my purposes well by following the lead of Torgerson (1958) who reserves the use of the term measurement as follows:

> The logic of measurement deals with the conditions necessary for the construction of a scale or measuring device. Measurement as used here refers to the process by which the yardstick is developed, and not to its use once it has been established, in, say, determining the length of a desk. It is essential that we keep this distinction in mind. The use of the established yardstick in "making a measurement" is a rather simple procedure involving merely the comparison of the quantity to be measured with standard series, or perhaps only reading the pointer or counter of an instrument designed for the purpose. We are here concerned with the more basic problem of establishing a suitable scale of measurement.
>
> ....measurement pertains to properties of objects, and not to the objects themselves. Thus, a stick is not measurable in our use of the term although its length, weight, diameter, and hardness might well be.
>
> Measurement of a property then involves the assignment of numbers to systems to represent that property. In order to represent the property, an isomorphism, i.e., a one-to-one relationship must obtain between certain characteristics of the number system involved and the relations between various quantities (instances) of the property to be measured.
>
> The essence of the procedure is the assignment of numbers in such a way as to reflect this one-to-one correspondence between these characteristics of the numbers and the corresponding relations between the quantities. (pp. 14-15)

Implicit in this usage is the preference not to use the term measurement in the broader sense of Stevens' classic definition: "Measurement is the assignment of numerals to objects or events according to rules."

(Stevens, 1951, p. 22.) Nominal scales, therefore, are not the result of measurement but of classification. Measurement presupposes, therefore, that the object has a property that exists in magnitudes that can be represented on either ordinal, interval or ratio scales. And again I align myself with Torgerson who finds it uninteresting to worry about what is or is not "permissable," in practice, with measurement scales of these several types:

> ....a major share of the results of the field of mental testing and of the quantitative assessment of personality traits has depended upon measurement by fiat. This is clear, for example, when curves are fitted by the process of least squares or when product-moment correlations, means, or standard deviations are computed. All of these presuppose that distance has meaning. Hence, either explicitly or implicitly, the experimenter is measuring the attribute on an interval scale whose order and distance characteristics have obtained meaning initially through definition alone.
>
> The discovery of stable relationships among variables so measured can be as important as among variables measured in other ways. Indeed, it really makes little difference whether [a] scale of length, for example, had been obtained originally through arbitrary definition, through a relation with other established variables, or through a fundamental process. The concept is a good one. It has entered into an immense number of simple relations with other variables. And this is, after all, the major criterion of the value of a concept. (p. 24)

The "act" of measurement, then, refers generally to both the logic of measurement and the process of constructing a test, i.e., a rule or set of procedures operationalizing the construct in a manner consistent with the logic of measurement. What, then, is a test theory? I would prefer that the phrase "test theory" denote the complete act of not only constructing the measuring instrument, but also of assessing further the validity of that instrument including its dependability[4] under specified conditions of use. In other words a theory of testing, to be complete, must include a measurement model, a dependability model and a validity theory. This last ingredient really includes (and goes beyond) the measurement and dependability models and is what justifies the usage of the term "theory." I know of no past or current "test theory" that deals explicitly with all three aspects. Traditional test theories are theories of dependability (some more restricted than others) with some validity theory. The newer latent trait models are just that, models for measuring a presumed construct. The focus of this paper is clearly on measurement, but by way of contrasting the act of measurement with the dependability of obtained measures.

Now suppose we had before us a small collection of the usual multiple-choice (or true false, completion, etc.) items of the type commonly found on a test designed to measure a specific achievement outcome. On their face, all such tests "look alike." However, depending upon the conceptual model of measurement underlying the analytical process for selecting these items, this innocent looking collection could be quite different in terms of item composition and empirical characteristics. It is the contention here that classical theory is conspicuously lacking in explicit regard for the potential value of the individual item. By this I mean that there is no explicit recognition of the measurement function served by items. Classical true and error models characterize the consequence of applying a measurement rule--they do not characterize the essence of the rule itself.

Let's consider the "essence of a measurement rule" by continuing the analogy with measuring a person's height. In measuring height, a tape measure and its properties operationalize the rule. Instead of "tape measure," let's use the simpler term "ruler." Suppose we use a ruler (of sufficient length) to measure peoples' heights. Traditional test theories have a lot to say about what to do with the obtained measurement; they have little to say, however, about how the ruler is constructed in order to obtain the measure, i.e., how the ruler is calibrated and how a numerical result eventually becomes associated with each person as a quantitative indicant of the height of the person. In other words, rather than the question of precision with which any given measurement is obtained, traditional test theories take the measurements as given and pursue the question of accuracy, i.e., how consistent the measurement rule is over repeated applications.

Precision and accuracy are cornerstone concepts of any theory of approximate numbers. They reflect fundamentally different ideas in the measurement process. Yet they are used inter-changeably in the behavioral sciences as a synonym for reliability. Two examples out of many are the following quotes:

> The physical scientist generally has expressed the accuracy of his observations in terms of the variation of repeated observations of the same event. The mean of the squared deviations of these observations about the obtained mean is the "error variance." This is a measure of precision or reliability....We regard reliability as the consistency of repeated measurements of the same event by the same process.... (Cronbach, 1947, p. 1.)

> Reliability of measurement, then, pertains to the precision with which some trait is measured by means of specified operations....Such indices will be useful for comparing different tests so we can ascertain which gives us the most precise or stable scores, and will permit us to ascertain whether the reliability with which a test measures is sufficient for our purposes....Casting reliability in terms of the coefficient of correlation between parallel tests provides another way of describing the precision of measurement. (Ghiselli, 1964, pp. 215-218.)

In the physical sciences, the concepts of precision and accuracy are clearly distinguished although not always in the same way. In the absence of empirical error, a measurement m precise to the nearest $u^{\underline{th}}$ unit has an inherent absolute error equal to $\pm u/2$. In this case, accuracy becomes relative error due to imprecision, i.e., $(u/2)/m$. But when empirical error exists--that is, error due to the measurer, the measuree, and/or the measurement circumstances--accuracy (<u>not</u> precision) is usually defined as in the first sentence of Cronbach's (1947) quote above. The dictionary is of little help in sorting out any systematic distinctions. For example, Webster's New World Dictionary (College Edition) gives us this definition: "Precision, the quality of being precise; exactness; accuracy." And in the same dictionary, is this definition: "Accuracy, the quality of being accurate or exact; precision."

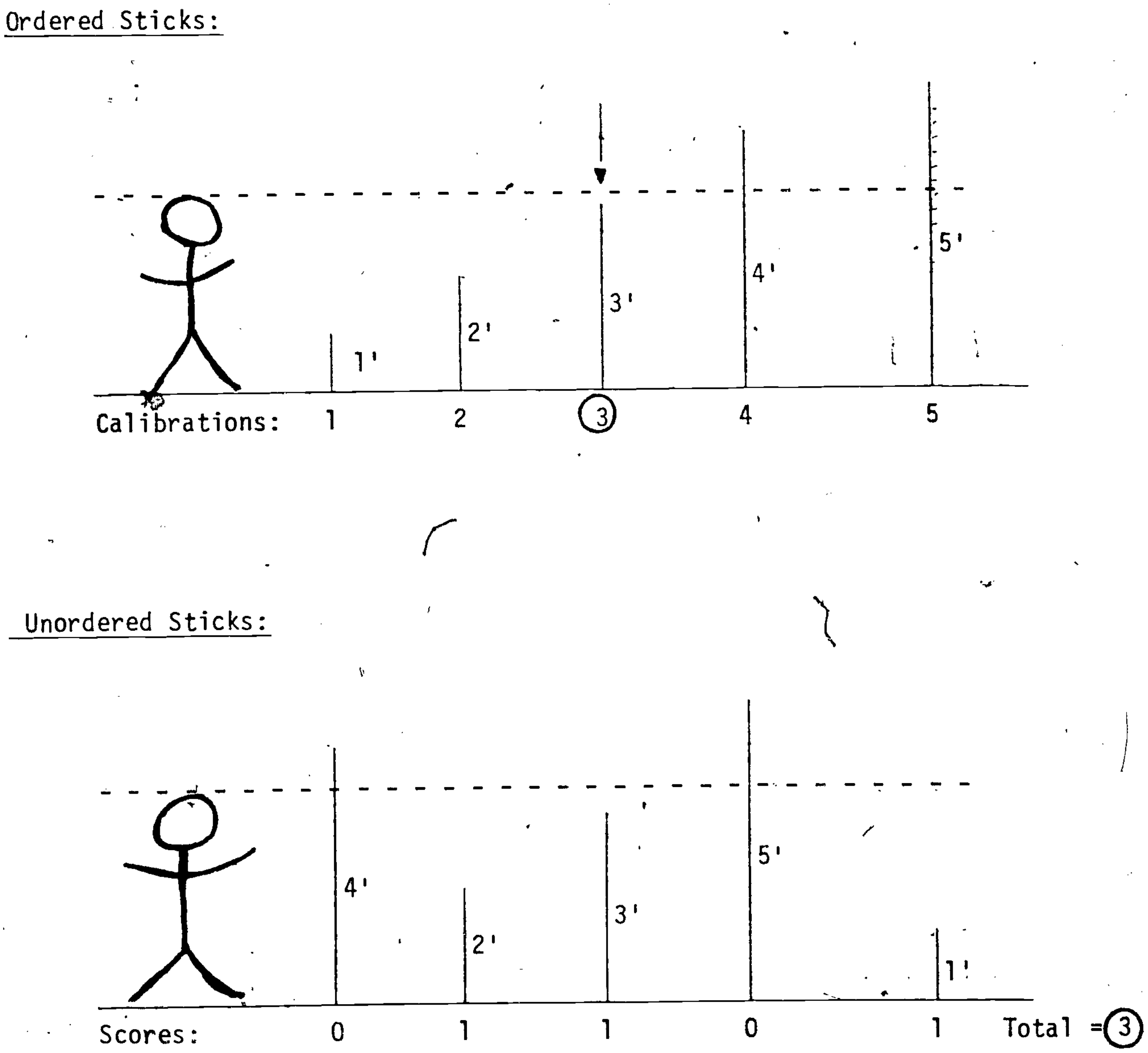
At the risk of confusing the issues further, I will elect the versions of these two concepts that serve to keep two fundamental properties of the measurement act separable. Suppose in measuring the height of a person, the ruler is marked off in feet; we can then measure anybody's height to the nearest foot. This is a statement of precision. Included in this notion of precision is the overall length of the ruler. If it is only 5 feet long, the measurement of people over 5 feet tall would necessarily be much less precise. Precision is intrinsic in the construction of the measuring instrument; it can be increased by conceptualizing and adding more hash marks to the ruler. Half feet can be added to the ruler enabling the measurement of height to be precise to the nearest half foot. It is not really necessary that the hash marks be at equal intervals, or that the addition of hash marks be midpoints of each interval.
Possibly a better conceptualization of precision is gained by defining it as the number of measurement decisions an instrument can potentially make. The ruler calibrated in half feet can potentially make twice the number of relative height decisions as can the ruler calibrated in feet.

To facilitate the analogy with test items, the ruler can be reconceptualized as a collection of straight sticks consisting of a 1-foot stick, a 2-foot stick, a 3-foot stick, and so on. The more precise ruler is reconceptualized as a set consisting of a 1-foot stick, a 1½-foot stick, a 2-foot stick, a 2½-foot stick, etc. Measurement of height, then, is the process of isolating two adjacent (ordinality being assumed) sticks within which lies the height in question and judging which of these sticks is closest, i.e., to within u/2 units where u is the unit of precision. Alternatively, the measure of a person's height is the number of sticks surpassed by the person's height (plus u/2). If the person is judged to be shorter (by u/2 or more) than the stick, he/she is scored zero; if taller, he/she is scored one. The person's height is then the total score after being tested on the set of sticks. Figure 1 lays out the process schematically. Whether sticks are ordered as calibration marks on a ruler or unordered and used summatively, the result is the same: the person's height is judged to be 3 feet to the nearest foot. That is, the person's height is somewhere in the theoretical interval of 2½ to 3½ feet. Precision is inherent in the way in which the measuring instrument is calibrated and made operational.

Accuracy is reserved here as a term for describing the degree to which the use of the measuring instrument is error-free. Accuracy is an empirical concept given an already calibrated instrument. Indexing the level of accuracy involves repeated measurements under the circumstances in which accuracy is required. In the above example, to the extent that we can consistently arrive at (or close to) the same measurement of height (to the nearest foot or half-foot depending upon which ruler we use), we have an accurate measuring procedure. The more accurate the procedure the less variability in obtained measurements over repeated measurement trials.

Figure 1

Schematic Representation of the
Act of Measurement
(Height as an Example)



Ordered Sticks:

Unordered Sticks:

The complete independence of the concepts of precision and accuracy should be clear: A highly precise instrument can be grossly inaccurate (a rubber measuring stick calibrated to the 32nd of an inch) compared to the accuracy of a less precise instrument (a steel measuring stick calibrated in yards). Moreover, accuracy is a function not only of instrument "decay," but also of the circumstances under which it is used. Technically, therefore, we assess the accuracy of the measurement procedure which includes error due to the instrument itself, the person doing the measurement, the person being measured, and the environment in which the measurement process takes place.

Given this distinction, reliability (or, more generally, dependability) as defined by classical (and classical-like) test theory models is clearly a synonym for the accuracy of a test. Empirically and theoretically, the concepts of reliability and dependability have been concepts of repeated measurements. In this sense, it matters little whether the repeated measurements are replicates (strictly parallel) or samples from a domain (randomly parallel); that is, the generic concept of accuracy remains intact regardless of the conceptual changes in meaning of "true score" implied by the several classical models. So long as we envision only the composite result of the testing process, the classical models are quite analogous to the physical model of measurement. The test score is analogous to the "ruler score," i.e., the obtained height measurement. If we are interested in assessing the accuracy of a single ruler, then we could use the original classical test theory model of strictly parallel repeated measurements. If, instead, we are more interested in the accuracy of a variety of rulers (wood, steel, cloth, etc.) from different manufacturers, then the item sampling models of randomly parallel repeated measurements would be useful. The domain of generalizability changes, but the notion of accuracy does not--empirical estimates obtain through repeated measurements, either with the same ruler (strict parallelism) or with a sample of rulers (random parallelism).

However, the physical model and traditional test theory models part company when it comes to the notion of internal consistency. Inquiry into the internal consistency of a ruler would be directed at the verification of the calibrations vis-a vis the construct in question and the selected measurement unit standard--an investigation of the precision of measurement. In test theory, the inquiry is directed, as it should be, toward the items. But in traditional theories, the inquiry proceeds by simply recasting items into the same role as the test, viz., repeated measurements--an investigation of the accuracy of measurement.

Where in the traditional test theory models is the concept of precision? Conceptually speaking, the answer is, "Nowhere." Now of course precision is manifested in the test item, in particular, the difficulty[5] of the test item. A student passing a more difficult test item evidences more ability than does a student who can pass only a less difficult item. The analogy with Figure 1 should be clear. The collection of items is the ruler, conceptualized as an ordered bundle of sticks. The item difficulties are analogous to the lengths of the sticks. Measuring the ability of a student involves locating that pair of adjacent items B and A such that the student correctly answers B (and all other items easier than B) but not A (nor all other items more difficult than A). Traditionally, the student's measure is the ordinal position of item B, or, equivalently, the total number of items answered correctly by the student.

Certainly this analogy is lacking in some non-trivial respects. In particular, the determinacy in the ordering of sticks is hardly (if ever) realized in the ordering of items. If stick C is shorter than stick B, and a student's height surpasses the length of stick B, then it will surely pass that of stick C. Such is the beauty of measuring constructs we can understand with our senses. But if item C is easier than item B, and a student correctly answers item B, then it is not always a sure bet that he/she will correctly answer item C as well.[6] Such is the legacy of the attempt to measure abstract behavioral constructs. Moreover, the procedure for assigning an invariant metric to the measurement of height is straightforward; it is much less so when using items to measure ability.

But I believe these to be minor details compared to the conceptual identity between sticks and items and their role as calibrations on the "ruler." The point to be made here is that this is not the role cast for items by classical (or classical-like) test theories. Lest I may have begun to lose some readers who are rusty on classical (and what I am referring to as classical-like) test theory, I will turn to an overview of several such theories with the expressed intent of further illustrating the argument thus far presented. (Readers already familiar with these models may skip to the Discussion in the next section with little or no loss in continuity.)

Traditional Test Theories

Some would probably argue (and justifiably so) that the sampling of alternative approaches to follow should not be lumped into a single class of test theories, especially one including classical test theory. I do this here only because, in terms of their fundamental conceptualization of the measurement process and important empirical conseq'ences, they are more similar to each other than to the models to be discusse. next.

Classical Test Theory

The basic postulate of classical test theory defines a belief regarding the composition of the raw score obtained by a student, namely, that this observed score is simply the student's true score plus what's left over, commonly designated as the error score.

Using some fairly standard notation and the usual matrix layout of the scores of n students on k items, we obtain the schematic in Figure 2. Using T and E for true and error scores, the classical test theory model posits for any student s that:

$$X_s = T_s + E_s \qquad (1)$$

A number of relationships obtain from this model when several additional assumptions are made about the true and error score components of repeated measurements on any student.[7] Specifically, these assumptions are (a) errors are totally random and cancel each other out; therefore, the mean error is zero ($\overline{E} = 0$); (b) the correlation between true and error score components is zero ($\rho_{TE} = 0$); and (c) the correlation between errors over repeated measurements is zero ($\rho_{EE'} = 0$).

Assumption (b) leads directly to the variance composition of the linear model above, viz., observed score variability is the sum of variability in true and error scores:

$$\sigma_X^2 = \sigma_T^2 + \sigma_E^2 \qquad (2)$$

Figure 2

Student-by-item raw score matrix and notation. (x_{si} = 1 or 0 if student s answers item i correctly or incorrectly.)

$$\begin{array}{cc|cccccc|c}
 & & \multicolumn{6}{c|}{\text{Items}} & \text{Raw Composite Scores} \\
 & & 1 & 2 & 3 \cdots & & i & \cdots\ k & \\
\hline
 & 1 & x_{11} & x_{12} & \cdots & \cdots & x_{1i} & \cdots\ x_{1k} & X_1 \\
 & 2 & x_{21} & x_{22} & & & \vdots & & X_2 \\
 & 3 & & & & & \vdots & & X_3 \\
\text{Students} & \vdots & & & & & \vdots & & \vdots \\
 & s & & & & & x_{si} & & X_s = \sum_{i=1}^{k} x_{si} \\
 & \vdots & & & & & \vdots & & \vdots \\
 & n & x_{n1} & \cdots & \cdots & \cdots & x_{ni} & \cdots\ x_{nk} & X_n \\
\hline
\end{array}$$

Item Difficulties $\quad p_1 \cdots p_i = \frac{1}{n} \sum_{s=1} x_{si} \cdots p_n$

Assumption (c) leads further to the fundamental theorem that the covariance between observed scores on any two repeated measurements is equal to that between the true scores on these measurements:

$$\rho_{XX'}\ \sigma_X\ \sigma_{X'} = \rho_{TT'}\ \sigma_T\ \sigma_{T'} \tag{3}$$

Finally, if a fourth assumption is added--(d) the repeated measurements are parallel measurements where parallel measurements are defined as having equal true scores (T = T) and equal error variances ($\sigma_E^2 = \sigma_E^2$)--then reliability (defined as the correlation between parallel measures, $\rho_{XX'} = \rho_{XX}$) is the equivalent of the ratio of true score to observed score variance:

$$\rho_{XX} = \frac{\sigma_T^2}{\sigma_X^2} \tag{4}$$

But this is also the coefficient of determination in predicting observed scores from true scores (or vice versa), i.e., the correlation between parallel measurements is equivalent to the square of that between observed and true score components:

$$\rho_{XX} = \rho_{XT}^2 \tag{5}$$

A little bit of algebraic manipulation of equations (2) and (4) gives us an equation for the error variance in terms of reliability and observed score variance. In standard deviation terms, this equation is

$$\sigma_E = \sigma_X \sqrt{1 - \rho_{XX}} \tag{6}$$

and is commonly referred to as the standard error of measurement. Noting again the relationship in (5), this equation also represents the standard error of estimate in predicting X from T:

$$\sigma_E = \sigma_X \sqrt{1 - \rho^2_{XT}} \qquad (7)$$

So much for theory. In practice we have only what we observe--raw scores X and the variance of these scores s^2_X which we use as an estimate of σ^2_X. In view of the above theoretical relationships, if we can also estimate ρ_{XX}, then estimates for the remaining parameters can be automatically computed. The estimate of reliability (denoted r_{XX}) is usually obtained in one or more of three fundamentally different ways with attendant differences in empirical interpretation.

Reliability as Stability. This is the test-retest formulation of reliability as the correlation between two administrations of the same test over a specified interval of time. If the time interval is too long and allows for true individual changes in the construct being measured, then the test-retest correlation has little to do with reliability. But if the time interval is well-defined in relation to the expected consistency in individual true scores over that period of time, then the test-retest correlation estimates the stability form of test reliability.

Reliability as Equivalence. This is the test-retest formulation of reliability as the correlation between two administrations of parallel tests at the same (or nearly so) point in time. This procedure most closely approximates the classical reliability definition but relies heavily upon the extent of true equivalence between the tests. (The same test could, of course, be used twice, but then practice effects might lead to inflated test-retest correlation.) This procedure most closely approximates the empirical assessment of accuracy as discussed in the previous section.

Reliability as Internal Consistency. This is the test-retest paradigm taken to its logical conclusion. For example, split-half reliability is one form of internal consistency equal to the correlation between two random halves of the test when adjusted upwards by the Spearman-Brown (Spearman, 1910 & Brown, 1910) equation to correspond to the full length test. But then we could compute a "split-fourths" coefficient by averaging all possible correlations between four random quarters of the test and adjusting this average accordingly. Eventually, we get down to the item level, treating each item as a parallel replicate "test." The intraclass correlation (average inter-item correlation) stepped-up by a factor of k (the number of items on the total test) by the Spearman-Brown formula turns out to be equivalent to the mean of all possible split-half coefficients (computed using the Rulon-Guttman formula [Rulon, 1939 & Guttman, 1945]) and was originally derived by Kuder and Richardson (1937) as their formula number 20:

$$\mathrm{KR20} = \frac{k}{k-1}\left[1 - \frac{\Sigma\, p_i(1-p_i)}{s_X^2}\right] \qquad (8)$$

Since $p_i(1-p_i)$ is the variance (s_i^2) of a binary item, this formula is often written more generally as

$$\mathrm{KR20} = \frac{k}{k-1}\left[1 - \frac{\Sigma\, s_i^2}{s_X^2}\right] \qquad (9)$$

Moreover, since the total variance s_x^2 can be decomposed into an additive sum of all item variances and twice the sum of all possible inter-item covariances, this formula can also be written as

$$\text{KR20} = \frac{\overline{r_{ij}\, s_i s_j}}{\frac{1}{k}\bar{s}_i^2 + \frac{k-1}{k}\,\overline{r_{ij}\, s_i s_j}} \qquad (10)$$

$$= \frac{\text{average interitem covariance}}{\frac{1}{k}\left(\begin{matrix}\text{average}\\ \text{item variance}\end{matrix}\right) + \frac{k-1}{k}\left(\begin{matrix}\text{average interitem}\\ \text{covariance}\end{matrix}\right)}$$

From equation (10) it is evident that this estimate of reliability (a) approaches 1 as the number of items increases (so long as additional items are positively correlated with the total test score) and (b) is a measure of the extent to which items are intercorrelated--with each other or, equivalently, with the total test score. Hence, the use of the term "internal consistency." It becomes clear, then, that this is not only an index of reliability, but also an index (necessary but not sufficient) of the extent to which the set of items comprising the test are measuring the same construct (ability). In the sense of internal consistency, therefore, reliability has a direct bearing upon the construct validity of the test. As noted above, it is for this reason that many traditional test theorists and practitioners have used the terms "homogeneous" and "unidimensional" to refer to this property of a test.

In a nutshell, these are the tenets and consequences of classical test theory. I have ignored a few other important consequences, primarily those having to do with the conceptualization of validity (effects of test length, correction for attenuation, and so forth). For purposes of comparison, however, the concepts so far developed are sufficient to illustrate what I believe to be profound differences between classical test theory and other, perhaps more realistic, measurement models.

Item Sampling Theory

One of the more difficult assumptions to accept (and empirically realize) is that requiring strictly parallel tests (or items). But with a slight shift in perspective, this assumption can be avoided. Consider again the layout in Figure 1. Suppose the k items are a random sample from a conceptually infinite population (universe, domain, pool, bank, etc.) of items over which a student's score would be meaningful. This score would theoretically be the student's true score. Likewise, the n students can be conceptualized as a random sample from an infinite population of students. And an item's true "score" (difficulty) is the theoretical average score on that item for the population of students.

In essence, what we have is the well-known random effects analysis of variance design, i.e., an n-by-k, students-by-items, random matrix sample from an infinite students-by-items matrix population. Once again, a linear, additive model is assumed; adopting the convention of using Greek letters for the population parameters, any student's (s) observed score on any item (i) is decomposed as follows:

$$X_{si} = \mu + \tau_s + \pi_i + \varepsilon_{si} \qquad (11)$$

where μ = the overall mean reflecting the general level of response relative to no response zero;

τ_s = true score for students s;

π_i = true score (difficulty) for item i;

ε_{si} = residual or error effect which could also be regarded as the student-by-item interaction effect ($\tau\pi_{si}$) for a design with one random observation per cell.

With the addition of one more critical assumption--the statistical independence of student-item responses--the components of variance mean square expectations shown in Table 1 can be derived (Cornfield & Tukey, 1956).

Table 1

Components of Variance Mean Square Expectations For the n x k Random ANOVA Model

Source	df	Mean Square	Expected Mean Square
Students	n - 1	MS_S	$\sigma^2_\varepsilon + k\sigma^2_\tau$
Items	k - 1	MS_I	$\sigma^2_\varepsilon + n\sigma^2_\pi$
Error	(n - 1)(k - 1)	MS_E	σ^2_ε

Now an internal consistency form of reliability can be derived without resorting to a definition based upon strict parallelism. Already, in accordance with the model, items can be characterized as randomly "parallel." We can proceed directly by defining reliability (ρ_{XX}) as the proportion of total score variance (σ_X^2) that is the true score variance (σ_τ^2). Since the model implies that

$$\sigma_X^2 = \sigma_\tau^2 + \frac{1}{k}\sigma_\varepsilon^2 , \qquad (12)$$

reliability can be expressed as

$$\rho_{XX} = \frac{\sigma_\tau^2}{\sigma_\tau^2 + \frac{1}{k}\sigma_\varepsilon^2} \qquad (13)$$

Using mean squares as estimates of their corresponding expected values, reliability can be estimated as

$$r_{XX} = \frac{MS_S - MS_E}{MS_S} \qquad (14)$$

which, with a bit of algebraic manipulation, can be shown to be identical to equations (8), (9) and (10) above. (This form of KR20 was first derived by Hoyt, 1941.) $\sqrt{MS_E}$, of course, is the corresponding estimated standard error of measurement equivalent to equation (7).

In terms of at least two important applied consequences (and there are more), then, both classical test and item sampling theories lead to the same result. Perhaps they are more similar than one might think.

Indeed, with the exception of the strict versus randomly parallel test distinctions, both theories are formally equivalent. It can be shown that the Cornfield and Tukey (1956) assumptions of the random components model imply assumptions (a), (b) and (c) above for the classical test theory model, and vice-versa. (See Lord and Novick, 1968, section 2.7.)

Nonetheless, the ANOVA framework implied by the item sampling model provides a convenient conceptual and analytic rubric that "liberates" (Cronbach, et al., 1963) the several classical reliability notions--that is, the sampling model emphasizes the multiplicity of possible reliability coefficients depending upon practical measurement consequences. Cronbach and his associates (Cronbach, et al., 1972) have formalized these concepts under the label "generalizability theory." In the simplest design, namely that represented in Figure 1, the "generalizability" coefficient is, of course, given by equation (14), designated previously by Cronbach (1951) as coefficient alpha (α). But other more complicated designs are also relevant and are obtained by adding more factors (facets)--and, therefore, more than one kind of true score parameter each with its corresponding reliability coefficient--to the ANOVA design. Suppose, for example, n classes are observed k times by r raters on o occasions. We can now talk about (and compute) reliability coefficients not only for the main effects due to observations, raters and occasions, but for the possible interaction effects as well. Using generalized Spearman-Brown procedures, data from one study can then be used to estimate the k, r and o necessary to reach desired reliability levels in a future study. Moreover, some facets might be considered fixed and others, random; and some populations finite, others infinite--all depending upon the practical applications intended.

However, notwithstanding the considerable conceptual and applied benefits accrued through liberating classical test theory of its strict assumption of parallel measurements, both theories conceive the fundamental dynamic of an achievement test identically: Items play roles as replicate measurement rules rather than calibrations on a single measurement rule. Hence, they are first and foremost theories of accuracy--not of precision--as these concepts have been defined above.

Binomial Error Model

An interesting twist on the item sampling model occurs if we restrict our attention to the single student s and conceptualize his/her responses to a random sample of k items as k independent binary events, each with the probability ζ_s of a correct answer where ζ_s is the hypothetically true proportion correct score for student s in the population of items from whence the sample was drawn. This is the simple "loaded coin-flipping" model, i.e., a binomial model, where the probability for success (say, "heads") is p. Over repeated trials of n coin flips each, the standard deviation of the sampling distribution (i.e., the standard error) of the observed proportions of "heads" is well known to be $\sqrt{p(1-p)/n}$.

Translated to the notation and purpose here, the standard error (of measurement) for student s is the standard deviation of his/her sampling distribution of observed proportion correct scores ($\overline{X}_s$) on repeated random samples of k as described in the paragraph above. This standard error (denoted σ_{ε_s}) is given, therefore, as

$$\sigma_{\varepsilon_s} = \sqrt{\frac{\zeta_s(1-\zeta_s)}{k}} \qquad (15)$$

This standard error of measurement is estimated for each student by correcting (15) for sampling bias and substituting observed scores for true scores:

$$s_{\varepsilon_s} = \sqrt{\frac{\overline{X}_s(1-\overline{X}_s)}{k-1}} \qquad (16)$$

It should be clear from equation (15) that for item sampled tests of fixed length k, <u>different</u> standard errors of measurement obtain for different true scores. Students obtaining a score of 50 percent will have the largest estimated standard error, i.e., $.5/\sqrt{k-1}$; s_{ε_s} decreases symetrically as scores either go up towards 100 percent or go down towards 0 percent.

This outcome, of course, is completely contrary to the assumption of independence of true and error scores in the classical test theory and item sampling models. In both of these models, the standard error of measurement (equation [7]) is a <u>constant</u> for all students regardless of their observed scores.

We can, however, derive a single standard error of measurement for the binomial model by simply computing the mean of the individual s_{ε_s}. To do this requires generalizing the binomial error model for an individual's score to that for a distribution of scores. (See Lord and Novick, 1968, Chapter 23.) And in so doing, a couple of interesting results emerge. Assuming a linear relationship between true and observed scores, the usual formulation of reliability as the ratio of true score to observed score variance leads to the following estimate for internal consistency:

$$\text{KR21} = \frac{k}{k-1}\left[1 - \frac{\overline{x}\,(k - \overline{x})}{k\,\sigma_x^2}\right] \qquad (17)$$

This, of course, is Kuder and Richardson's formula 21 developed originally as an approximation to KR20. Clearly, it is a function only of the observed score mean (or mean item difficulty since $n\overline{p} = \overline{x}$) and observed score variance. KR21 will always be less than KR20 unless there is no variation in item difficulties. When all items are of equal difficulty, they are, of course, equal to their average and formula (17) becomes identical to formula (8).

Analogous comparisons hold for the standard error of measurement. For the binary model, it follows that the estimated correlation between true and observed scores is $\sqrt{\text{KR21}}$ and the estimated standard error of measurement is:

$$s_\varepsilon' = s_x\sqrt{1 - (\text{KR21})} \qquad (18)$$

It can be easily shown that s_ε' is the mean of the individual student standard errors of measurement s_{ε_s}. This quantity will always be greater than its analogue in classical and item sampling models (equation [7] with sample estimates) unless, again, item difficulties are equal.

Discussion

Thus, excepting the test construction consequences of strict versus randomly parallel items, all three "traditional" models appear, for all practical intents and purposes, to be equivalent when item difficulties are equal (or nearly so). This makes a lot of sense when one teases out the subtle differences in the conceptions of true score inherent in each model. In the general binary error model, the true score is a parameter of the item population, but each student receives a different randomly sampled set of items. Ordinarily, a student will have different true scores on each of those item samples, but these are not the true scores of interest. Rather, it is the mean of these true scores (the item population true score) that is to be estimated for each student. A similar conception of true score holds for the item sampling model except that each student responds to the same randomly sampled set of items. The classical model is a degenerative form of the item sampling model where all π_i are equal. But in the event that items are all of equal difficulties, true scores will be identical, in each item sample, and, of course, these are identical to the true score in the population. However, if this is not the case, and students respond to different item samples, more variation can be expected to enter into any summary statistics designed to reflect measurement error.

So where in these "traditional" test theories is the concept of precision as I have defined it? Where do the theories speak to the construction and calibration of the measurement device? Again, the answer is nowhere. I am not, of course, suggesting that items go unrecognized in traditional test theories. However, I am suggesting that the item parameters, for example, in the model specified by (11), are there mostly by default. Moreover, I'm suggesting that precision, which is indeed gained in the composite test score, is serendipitious--items are invariably nonparallel and tests are usually long enough with sufficient variation in item difficulties so that total scores are at least positively and monotonically related to the underlying ability continuum. Put slightly differently, I am suggesting that the wrong theoretical framework for conceptualizating the act of measurement has been used to evaluate what turns out to be a fairly common and intuitively sensible approach to the measurement of ability.

Consider this ironic outcome in terms of classical test theory: differences in item difficulties (desirable building blocks for measurement) are evidence for violating the fundamental assumption of parallelism for the internal consistency form of reliability. Moreover, such differences automatically put a ceiling on the maximum level of KR20 (or alpha) due to the ceiling on phi coefficients when marginal proportions are not identical. For these reasons, we all learned that the "best" possible test was one with items of near equal difficulty and, preferably, all at the .5 level to maximize the potential for total score variance--all nice ingredients for norm-referenced applications. Not surprisingly, it is under the "ideal" condition of equal item difficulties that all three traditional test theory models are, for practical intents and purposes, identical.

This "ideal" student-item response pattern highlights the folly of treating items as merely short (the shortest) repeated tests. As implied above, maximum KR20 obtain when items are at the .5 difficulty level and all students either get all items right or wrong. For a k-item test, then, half the students have a score of k and half have a score of 0. Clearly little information is obtained when only two decisions can be made. (Latent trait models, which attack the issue of calibrating test items directly, can not even utilize "perfect" response vectors since they have no utility in pinpointing locations on the latent continuum.) Equally ironic implications of this "ideal" score matrix occur for validity coefficients. (See Loevinger, 1954.) It is a rather sad commentary that "something fishy" about classical test theory was smelled early on by scholars who continued to propagate the methods:

> It may be, if items of graded difficulty levels are used, that counting one point for each item correct is not a proper scoring method. The score assigned should rather be a best estimate of the difficulty level reached, analogous to that used in the Binet test.... Another limitation in the theory here developed should be pointed out. The criterion of maximizing test variance cannot be pushed to extremes. Test variance is a maximum if half of the population makes zero scores, and the other half makes perfect scores. Such a score distribution is not desirable for obvious reasons, yet current test theory provides no rationale for rejecting such a score distribution. Obviously the "best" test score distribution is one which accurately reflects the "true" ability distribution in the group, but there is perhaps little hope of obtaining such a distribution by the current procedure of assigning a score based upon sheer number of correct answers. At present the only solution to such difficulties seems to lie in some type of absolute scaling theory.... (Gulliksen, 1945, pp. 90-91.)

As a final example of the ironies inherent in classical models consider the classical test theory notion of a <u>constant</u> standard error of measurement for every possible score. Does it make sense that particular high (or low) scoring students would have the same random error distributions around their true scores as would intermediate scoring students? At a purely intuitive level this doesn't make much sense at all. The binomial error model makes it clear that errors are smaller at the ends of the score distribution and larger towards the center. This makes perfect sense if we think of sampling items as analogous to sampling balls from an urn to achieve accuracy of estimation--blue balls are items answered correctly, red ones are incorrect items, and a student's estimated true score is the proportion of blue balls obtained when selecting k balls at random from the urn.

But it makes no sense if items are conceived as fundamental building blocks of the measurement process. In this case, "error" ought to become much more associated with the precision of measurement. In fact, the error pattern should be the complete reverse of that predicted by the binomial model. Errors would be larger toward the extremes of the score distribution and smaller towards the center. At the extremes, we know nothing about the ability level of persons scoring 0 or k on a k-item test. The analogy to physical measurement is again instructive. It is equivalent to selecting that bundle of sticks of appropriate length such that they can center on the person's height. If the smallest stick is too long (a 0-scorer) or the longest stick too short (a 1-scorer), we have failed to measure the person's height to within the given units of precision.

In sum, it can be said that classical (and classical-like) test theories are good models for assessing the dependability of measurements whose internal measurement properties are already well understood or at least accepted as given. (Generalizability theory becomes particularly useful in these circumstances as noted previously.) But they are poor models for directing and assessing the development of item-based measures which, as suggested by the physical measurement analogy, rely upon item difficulties as proxies for calibrations on the "ruler." Again, many achievement tests produce useful results serendipitously for the obvious reason that practitioners of classical testing methods sense the necessity for including items of varying difficulty. But the reasons for the eventual presence or absence of items on their tests are the wrong ones, being rooted in a "theory" of dependability rather than measurement. I will now turn to an illustrative survey of some measurement models which are theoretically oriented in the latter direction.

Cumulative Test Models

For lack of a better one, I am using the term cumulative to refer to a rather heterogeneous class of measurement models which explicitly acknowledge the measurement function of items as heretofore discussed. If not already obvious, the descriptive value of this term will be apparent shortly. A potpourri of these models will be presented in just enough detail to highlight how they radically differ from classical (and classical-like) test theories in their conceptual approach to the measurement act. All these cumulative models approach the measurement act directly (using the items-as-sticks notion) relying on item difficulty variance for precision and calibration and the total score(or a function of the total score) as an indicant of the ability being measured.[8]

Before beginning this survey, I wish to note a side benefit to using the "items-as-sticks" notion in developing a measurement rule (i.e., test). In 1963, a seminal article by Glazer stimulated the so-called criterion-referenced testing movement. Soon thereafter, an important article by Popham and Husek (1969) rightly noted the inappropriateness of norm-oriented classical test theory methods for handling the development and analysis of criterion-referenced tests. The literature virtually exploded with attempts to adapt classical test theory to fit the requirements of criterion-referenced tests. The focus of these efforts was quite misdirected. The fundamental issue was not testing or even purpose of testing; rather, it was an issue of measurement. The proper role of items in a test forces (or should force) the test constructor to match item content with the cognitive processes to be assessed. Assuming a singular construct and a scalable set of k items having different difficulties, k + 1 "mastery" levels can be assessed. "Criterion-referenced testing", therefore, is simply sensible measurement.[9] Of course, following sensible measurement, one can always (a) select a particular mastery level for criterion-referenced decisions or (b) compile group statistics for comparative purposes, thereby developing norm-referenced test interpretations.

Guttman's Scalogram Analysis

David Walker (1931, 1936, 1940), perhaps the first person to recognize the value of the doubly ordered raw score matrix, began a series of investigations on the relationship between response patterns and the resultant shape of score distributions. In the course of this inquiry, Walker conceptualized the ideal response pattern and attempted to index departures from this pattern, a condition he nicknamed "hig" after the term "higgledy-piggledy" to describe the apparent haphazardness in non-ideal response patterns. But his interest centered on implications for test score scatter rather than the more profound implications for measurement itself.

Guttman (1944) reversed this focus and formalized a scaling procedure for assessing the degree to which items conformed to the ideal response pattern. Figure 3a presents an example of an ideal cumulative response pattern for 20 students responding to five items. However, that this is an ideal pattern is not immediately obvious until the score matrix is arranged in rank order on both student scores and item difficulties. One such convienent "double sorting" of the score matrix orders students from highest to lowest scores and items from easiest to most difficult. In Figure 3b we see the cumulative nature of the scoring pattern inherent in the unsorted data as presented in Figure 3a. Figure 4 presents the same score distribution, but this time there are some "errors," i.e., student-item responses which do not fit the ideal pattern. For example, student 8 should have answered item 1 correctly and item 5 incorrectly, thereby contributing two student-item response errors to the total 20x5 (i.e., nk) possible student-item responses. Finally, Figure 5 depicts yet again the same score distribution but with many errors resulting in a very poor cumulative pattern.

To index the degree of cumulativeness present in the pattern, Guttman used a deterministic approach. All deviations (e) from the ideal pattern are errors, i.e., the approach makes no allowance for probable deviations. An obvious index then is the proportion of non-errors in the entire response matrix (1-e/nk). Guttman named this index the coefficient of reproducibility (REP) insofar as it reflected the extent to which the response pattern could be perfectly reproduced from the student scores or item difficulties. Thus,

$$\text{REP} = 1-\frac{e}{nk} \qquad (19)$$

Figure 3a

Unsorted Cumulative Response Pattern
for a Hypothetical Ideal Score Matrix

	ITEMS					
STUDENTS	3	2	5	1	4	X
12	0	1	0	1	0	2
19	0	0	0	0	0	0
16	0	0	0	1	0	1
11	0	1	0	1	0	2
5	1	1	0	1	1	4
15	0	1	0	1	0	2
2	1	1	1	1	1	5
13	0	1	0	1	0	2
3	1	1	0	1	1	4
9	1	1	0	1	0	3
1	1	1	1	1	1	5
6	1	1	0	1	0	3
20	0	0	0	0	0	0
14	0	1	0	1	0	2
10	1	1	0	1	0	3
17	0	0	0	1	0	1
4	1	1	0	1	1	4
8	1	1	0	1	0	3
18	0	0	0	1	0	1
7	1	1	0	1	0	3
	10	15	2	18	5	
p_i =	.50	.75	.10	.90	.25	

Figure 3b

Sorted Cumulative Response Pattern
for a Hypothethical Ideal Score Matrix
(Rep = 1.00; CS = 1.00; α = .76)

ITEMS

STUDENTS	1	2	3	4	5	X
1	1	1	1	1	1	5
2	1	1	1	1	1	5
3	1	1	1	1	0	4
4	1	1	1	1	0	4
5	1	1	1	1	0	4
6	1	1	1	0	0	3
7	1	1	1	0	0	3
8	1	1	1	0	0	3
9	1	1	1	0	0	3
10	1	1	1	0	0	3
11	1	1	0	0	0	2
12	1	1	0	0	0	2
13	1	1	0	0	0	2
14	1	1	0	0	0	2
15	1	1	0	0	0	2
16	1	0	0	0	0	1
17	1	0	0	0	0	1
18	1	0	0	0	0	1
19	0	0	0	0	0	0
20	0	0	0	0	0	0
	18	15	10	5	2	
p_i =	.90	.75	.50	.25	.10	

Figure 4

Moderately Cumulative Response Pattern
(Rep = .86; CS = .63; α = .57)

ITEMS

STUDENTS	1	2	3	4	5	X
1	1	1	1	1	1	5
2	1	1	1	1	1	5
3	1	1	1	1	0	4
4	0	1	1	1	1	4
5	1	0	1	1	1	4
6	1	1	1	0	0	3
7	0	1	1	1	0	3
8	0	1	1	0	1	3
9	1	1	1	0	0	3
10	1	0	1	1	0	3
11	1	0	0	1	0	2
12	1	1	0	0	0	2
13	1	1	0	0	0	2
14	1	0	0	1	0	2
15	1	1	0	0	0	2
16	1	0	0	0	0	1
17	1	0	0	0	0	1
18	1	0	0	0	0	1
19	0	0	0	0	0	0
20	0	0	0	0	0	0
	15	11	10	9	5	
p_i =	.75	.55	.50	.45	.25	

Figure 5

Poor Cumulative Response Pattern
(Rep = .74; CS = .46; α = .49)

	ITEMS					
STUDENTS	1	2	3	4	5	X
1	1	1	1	1	1	5
2	1	1	1	1	1	5
3	1	1	1	1	0	4
4	1	1	0	1	1	4
5	0	1	1	1	1	4
6	1	0	1	1	0	3
7	0	0	1	1	1	3
8	1	0	1	0	1	3
9	1	1	1	0	0	3
10	0	1	1	1	0	3
11	1	0	0	1	0	2
12	1	0	0	0	1	2
13	0	1	0	0	1	2
14	0	1	0	1	0	2
15	1	1	0	0	0	2
16	0	0	0	0	1	1
17	0	0	1	0	0	1
18	1	0	0	0	0	1
19	0	0	0	0	0	0
20	0	0	0	0	0	0
	11	10	10	10	9	
p_i =	.55	.50	.50	.50	.45	

But REP can never be smaller than the average of the observed item difficulties (p_i) or easinesses (q_i=1-p_i), whichever are greatest. That is:

$$\text{Min(REP)} = \frac{\Sigma\text{Max}(p_i,q_i)}{k} \qquad (20)$$

The degree of improvement (IMP) over minimum reproducibility is, therefore,

$$\text{IMP} = \text{REP-Min(REP)} \qquad (21)$$

Moreover, the maximum possible improvement is

$$\text{Max(IMP)} = 1\text{-Min(REP)} \qquad (22)$$

Thus, a more realistic appraisal of the degree to which items scale, above that expected by the marginal results alone, can be seen in the ratio of IMP to Max(IMP). Denoted the coefficient of scalability (CS) by Menzel (1967), this index can be written as follows:

$$\text{CS} = \frac{\text{REP-Min(REP)}}{1\text{-Min(REP)}} \qquad (23)$$

It has usually been recommended that reasonable scalability requires REP$\geq$ .9 and CS$\geq$ .6. The score matrices in Figures 3a, 4 and 5 depict what are ideally, moderately and weakly cumulative response patterns. These descriptors are clearly reflected in the values of REP and CS accompanying each score matrix.

There are probably three basic reasons why Guttman scaling received little favor in the achievement testing arena. First, for reasonably homogenious objective domains, it is difficult to write achievement items which scale well. In fact, Guttman devised the scalogram procedure for attitude measurement, where it is often easier to write items with distinctly different affective magnitudes (item "difficulties") covering the same essential domain. Second, Guttman made unrealistic claims regarding the power of scalogram analysis to test unidimensionality, thereby opening up the procedure to a barrage of criticism. (See, for example, Festinger, 1947 and Loevenger, 1948.) In line with the discussion of unidimensionality earlier in this monograph, Guttman would have treaded firmer ground were he to have simply suggested that a scalable set of items is necessary but not sufficient evidence that a set of items measures the same thing to within reasonable evidence of content (and/or construct) validity. Third, and probably most critical, the model was deterministic and offered no statistical (i.e., probabilistic) tests of fit. (See Torgerson, 1958.)

But no criticism was ever directed at the most important notion behind Guttman's approach, namely, the measurement role of items as, in essence, calibrations on a "yardstick." The approximation to the ideal pattern (Figure 3b) would most likely be the acknowledged goal of most achievement test constructors. Yet, instead of expending considerable effort in mapping the cognitive consequences of instructional units and writing, testing, modifying and rewriting relevant items that <u>do</u> begin to show nice cumulative properties, test constructors have been content to build tests on the classical test theory principle of redundancy, i.e., repeated measurements to realize reliability (as internal consistency).

As an interesting aside note, even the deterministic nature of Guttman scaling was rendered a non-issue by a number of writers. Perhaps the most ingenious approach was based upon Cox's (1954) analysis of covariance model for cumulative repeated measurements (see Maxwell, 1959 and Ten Houten, 1969). Other techniques were investigated by Goodman (1959), Sagi (1959) and Schuessler (1961). The point of this note is simply that attention needs to be redirected towards the underlying principles of measurement and away from the worry of more or less sensitive statistical indicators--not that the latter are unimportant, but that the former are much more so.

Loevinger's Homogeneity Analysis

In her 1947 monograph, Jane Loevinger delivered what I believe to be among the best and most provocative critiques of classical test theory; and she followed up with an equally provocative critique of item sampling theory in 1965. To be sure, some of Loevinger's criticisms were a bit overstated, particularly her judgment that the axioms of classical test theory were circular (see Novick, 1966). But generally, her view regarding the inappropriateness of treating items as repeated measurements and her switch in focus from reliability to constructing cumulative scales represents the fundamental contribution.

Like Guttman, Loevinger's approach is based upon deviations from the ideal response pattern. Unlike REP (and its derivatives), however, her homogeneity index (H) reflects these discrepancies in terms of maximum expectations given the difficulty level of the items. Assuming items are arranged in ascending order of difficulty, then for any two items i and j the usual four-fold classification table obtains:

Item j

		1	0		
Item i	1	a	b	a+b	$p_i=(a+b)/n$
	0	c	d	c+d	$q_i=(c+d)/n$
		a+c	b+d	n=a+b+c+d	
		$p_j = (a+c)/n$	$q_j = (b+d)/n$		

a, b, c, and d are the number of students in each of the respective possible score patterns. Since we have arranged the data assuming item i is easier than j, a+b must be greater than a+c; in proportion terms, $p_i > p_j$.

Ideally, no one answering the more difficult item correctly would answer the easier item incorrectly. The ideal four-fold classification table would then look like this:

Item j

		1	0	
Item i	1	a	b	a+b
	0	0	d	d
		a	b+d	n

But in the actual testing process, "errors" do occur and c, the number of students getting the more difficult item right but the easier item wrong, is often not zero. These are the deviations from the ideal scale types in Figure 4 and 5.

Loevinger's index of "homogeneity" focuses just on the outcomes a and c, that is on the easier item's scoring pattern for those students answering the more difficult item correctly (heavily outlined column in above schematics.) In other words, the index is based upon the <u>conditional probability</u> $p_{i|j}$ of answering item i correctly <u>given that</u> item j is answered correctly. In the general case, this probability is given by the number of students a who answered both items correctly divided by the total number of students a+c who answered item j correctly:

$$p_{i|j} = \frac{a}{a+c} = \frac{p_{ij}}{p_j} \qquad (24)$$

where p_{ij} is simply the proportional equivalent of a, viz., a/n, which is the probability of answering both items i and j correctly. In the ideal case, perfectly homogenious items (like in Figure 3b), c=0 and $p_{i|j}=1$. In the perfectly heterogenious case, we would expect items to function completely independently, i.e., $p_{ij}=p_ip_j$, in which case $p_{i|j}=p_i$ by (24) above. An index of homogeneity between the two items i and j can then be formed as follows:

$$H_{ij} = \frac{\text{observed improvement in } p_{i|j} \text{ over that expected under perfect heterogeneity}}{\text{maximum possible such improvement if items were perfectly homogenious}}$$

$$= \frac{p_{i|j} - p_i}{1 - p_i} \qquad (25)$$

In form and intent, this coefficient is analogous to the coefficient of scalability (23) proposed for Guttman scaling. But H_{ij} has a number of further properties. Among the more interesting is the following:

$$H_{ij} = \frac{\phi_{ij}}{\text{Max}(\phi_{ij})} \qquad (26)$$

where ϕ_{ij} is the ordinary Pearson product-moment correlation between two items which, since the items are binary, is also the fourfold point-correlation computed as:

$$\phi_{ij} = \frac{p_{ij} - p_i p_j}{p_i q_i p_j q_j} \qquad (27)$$

But ϕ_{ij} cannot reach unity unless the marginals p_i and p_j are equal, i.e., unless the item difficulties are equal. This is exectly the circumstance under which the two items are useless for purposes of precision, i.e., they replicate the same calibration information rather than add decision points to the scale. And of course this is exactly the condition most suited for classical test theory, a theory of accuracy.

However, we can "correct" ϕ_{ij} by dividing it by the maximum possible value it can assume in the case of unequal p_i and p_j. That is

$$\text{Max}(\phi_{ij}) = \frac{p_j - p_i p_j}{p_i q_i p_j q_j} \qquad (28)$$

and thus

$$\frac{\phi_{ij}}{\text{Max}(\phi_{ij})} = \frac{p_{ij} - p_i p_j}{p_j - p_i p_j.} \qquad (29)$$

Upon dividing both numerator and denominator of (29) by p_j, the equivalency given by (26) is verified.

But the result is more than algebraic. The maximum ϕ_{ij} is obtained when all the students answering item j correctly also answer item i correctly, i.e., when $p_{ij}=p_j$. This, of course, is the ideal cumulative response pattern shown in the above schematic. Thus, $\phi_{ij}/\text{Max}(\phi_{ij})$ is really measuring the extent to which this ideal is obtained and ranges from 0 to 1 accordingly. Unfortunately, this index suffers a bit from the fact that it can also be 1 in value for items of equal difficulties when the b cell is also zero. Even in the extreme case of Figure 6, the overall index (H_t) of homogeneity (see below) is unity. Guttman indices suffer from the same problem. In effect, the scaling indices being presented here are necessary but not sufficient indicators of the cumulative nature of the test items. (See footnote 8.) We must also, therefore, have some indication of item difficulty spread over the ability range of interest.

To complete the discussion of Loevinger's approach, we note that a weighted average of H_{ij} can be formed for all item pairs i and j (such that $p_i > p_j$) yielding an overall index of test homogeneity (H_t). The most straightforward approach to constructing H_t is to reconsider equation (29) which was formed as a ratio of equations (27) and (28). Since the item variances in the denominators of (27) and (28) cancelled out, (29) is, in effect, the ratio of the observed covariance of items i and j to the maximum possible covariance given the p_i and p_j. An overall index can then be formed as a ratio of the sum of the k(k-1)/2 unique observed covariances to the sum of the corresponding k(k-1)/2 maximum covariances:

Figure 6

A Degenerate Case:
The Perfect Classical Test Response Pattern
(Rep = 1; CS = 1; α = 1)

		ITEMS					
		1	2	3	4	5	X
STUDENTS	1	1	1	1	1	1	5
	2	1	1	1	1	1	5
	3	1	1	1	1	1	5
	4	1	1	1	1	1	5
	5	1	1	1	1	1	5
	6	1	1	1	1	1	5
	7	1	1	1	1	1	5
	8	1	1	1	1	1	5
	9	1	1	1	1	1	5
	10	1	1	1	1	1	5
	11	0	0	0	C	0	0
	12	0	0	0	0	0	0
	13	0	0	0	0	0	0
	14	0	0	0	0	0	0
	15	0	0	0	0	0	0
	16	0	0	0	0	0	0
	17	0	0	0	0	0	0
	18	0	0	0	0	0	0
	19	0	0	0	0	0	0
	20	0	0	0	0	0	0
		10	10	10	10	10	
	p_i =	.50	.50	.50	.50	.50	

$$H_t = \frac{\sum_{i \neq j}\sum (p_{ij} - p_i p_j)}{\sum_{i \neq j}\sum (p_j - p_i p_j)} \qquad (30)$$

$$= \frac{\overline{\text{Cov}}_{ij}}{\overline{\text{Max}(\text{Cov}_{ij})}}$$

where Cov_{ij} denotes the covariance between items i and j. Some algebraic manipulation of (30) will verify that it can also be written as

$$H_t = \frac{\sum_{i \neq j}\sum p_j q_i H_{ij}}{\sum_{i \neq j}\sum p_j q_i} , \qquad (31)$$

i.e., H_t is a weighted (by $p_j q_i$) average of $H_{ij} = \phi_{ij}/ \text{Max}(\phi_{ij})$. This makes intuitive sense since $p_j q_i$ is the expected proportion of <u>errors</u> in the completely heterogeneous (non-cumulative) case.

It should be clear that H_t is an average inter-item statistic assessing the degree to which all possible ordered item pairs are homogeneous (in the cumulative sense) on the average. Thus, it does not increase merely as a function of increased number of items as does the internal consistency coefficient α in traditional test theory. This is as it should be since H_t is intended to index the cumulative structure of items while α is aimed at assessing the reliability of repeated item measurements.

Ironically, Horst (1953), capitalizing on the seductively simple relationship between H_t and the intraclass reliability coefficient of classical test theory, has proposed "blowing up" H_t by a factor of k using the Spearman-Brown prophecy formula to correct the ceiling effect problem of unequal item difficulties in classical test theory. To his credit, Horst is among the few test theorists who has recognized conceptual differences between reliability and homogeneity and devoted ample space to Loevinger's work in his book on measurement theory (Horst, 1966). But although I can relate to the intended use of the modification offered by Horst, the modification once again confuses fundamental measurement issues by commingling the concepts of precision and accuracy.

Consider, first, the specifics of the modification. The intraclass reliability (r_{ii}) in classical test theory is the reliability of the average single-item test. It can be shown that by adjusting r_{ii} upwards by a factor of k using the classical Spearman-Brown formula, we end up with the KR20 (or α) formula for reliability at the total test level. Noting that r_{ii} can be defined as the ratio of the average inter-item covariance to the average item variance, i.e.,

$$r_{ii} = \frac{\overline{r_{ij}s_is_j}}{\overline{s_i^2}} = \frac{\overline{Cov_{ij}}}{\overline{Var_i}} \qquad (32)$$

the relationship given in equation (10) leads directly to the Spearman-Brown "correction" as follows:

$$KR20 = \frac{k\ r_{ii}}{1 + (k-1)r_{ii}} \qquad (33)$$

Now, the maximum possible r_{ii} given the disparities in item difficulties is

$$Max(r_{ii}) = \frac{\overline{Max(Cov_{ij})}}{\overline{Var_i}} \qquad (34)$$

If we correct r_{ii} in the usual manner, it is obvious that

$$\frac{r_{ii}}{Max(r_{ii})} = \frac{\overline{cov}_{ij}}{Max(Cov_{ij})} = H_t \qquad (35)$$

The suggested modification by Horst, therefore, is to substitute the corrected r_{ii}, i.e., H_t, in equation (33), thereby making it possible for KR20 to reach unity even when item difficulties are unequal.

$$\text{Corrected} \quad \text{KR20} = \frac{k\,H_t}{1 + (k-1)H_t} \qquad (36)$$

Consider, second, the implication of this formula. A test can be perfectly homogeneous by adding an infinite number of mostly heterogeneous items so long as they are positively correlated. Now this seems reasonable for achieving increasingly accurate measurements; but it does not necessarily lead to increased precision and a more scalable set of items. Suppose, for example, the test is doubled in length by adding k parallel items, i.e., items that are equal in difficulty, one-for-one, to those in the original test and that scale identically to those in the original test. We now have twice the test information at each ability level but still the same number of ability levels represented in the test. Suppose, again, that the new items are equally scalable but have difficulty levels between those of the original items. We now have the same information at each ability level but twice the number of ability levels that can be assessed. Formulas such as (36) "blow-up" the index indiscriminately thereby conflating the issues of accuracy and precision.

Horst (1966) makes an effort to distinguish reliability and homogeneity by noting that reliable items are a necessary but not sufficient condition for high H_t. Thus, high H_t is, in part a function of reliability. Now this is true for reliability at the item level. But it is not true for reliability (as internal consistency) at the test level. Again, I am trying here to clearly separate the precision obtained through calibrating a homogeneous or unidimensional test from the accuracy of test.

Bentler's Monotonicity Analysis

I include a discussion of Bentler's (1971) approach here primarily to emphasize that multidimensionality is not an intractable issue when measurement is conceived and operationalized as a cumulative scaling process. Thus far I have avoided the issue of empirical dimensionality suggesting, instead, that a scalable or homogeneous set of items plus reasonable evidence of content validity is a necessary but not sufficient condition for unidimensionality. Although I (and others) often use the terms unidimensional and homogeneous synonymously, it should be understood that the former is not an automatic consequence of the latter.

Preferring the term monotonic (instead of cumulative), Bentler quite cleverly recognized that Yule's Y coefficient (a simple function of the more familiar Yule's Q coefficient) for association in a four-fold table (see Yule; 1912) possessed none of the drawbacks of ϕ or ϕ/ϕ_{max} when subjected to an ordinary principal components factor analysis. For any two items i and j, this index, renamed the monotonicity coefficient by Bentler since he developed it in a more general form, is given as follows:

$$m = \frac{bc - ad}{bc + ad + 2\ abcd} \qquad (37)$$

where a, b, c and d are as given in the four-fold table layout in the previous section. The nice thing about Yule's association measure is that it becomes 1 (or -1) only when one (or more) cells are empty. These include exactly those four-fold response patterns of cumulative scales; and a principal components factor analysis of the inter-item m-matrix will recover two or more cumulative scales embedded in a set of items.

As an index of homogenity, m is very similar to H_{ij}. And, like Loevinger, Bentler proposes the average of all k(k - 1)/2 inter-item monotonicity coefficients, $\overline{m}$, as an overall measure of inter-item homogeneity. But then, like Horst, Bentler becomes concerned with the length of the test not being represented in the index. Thus, he proposed the same Spearman-Brown transformation of $\overline{m}$ for a final, overall measure of the test's homogeneity (h),

$$h = \frac{k\,\overline{m}}{1 + (k - 1)\overline{m}} , \qquad (38)$$

and, in my view, falls into the same trap of mixing up fundamentally distinct measurement issues.

Sato's Student-Problem (S-P) Matrix Analysis

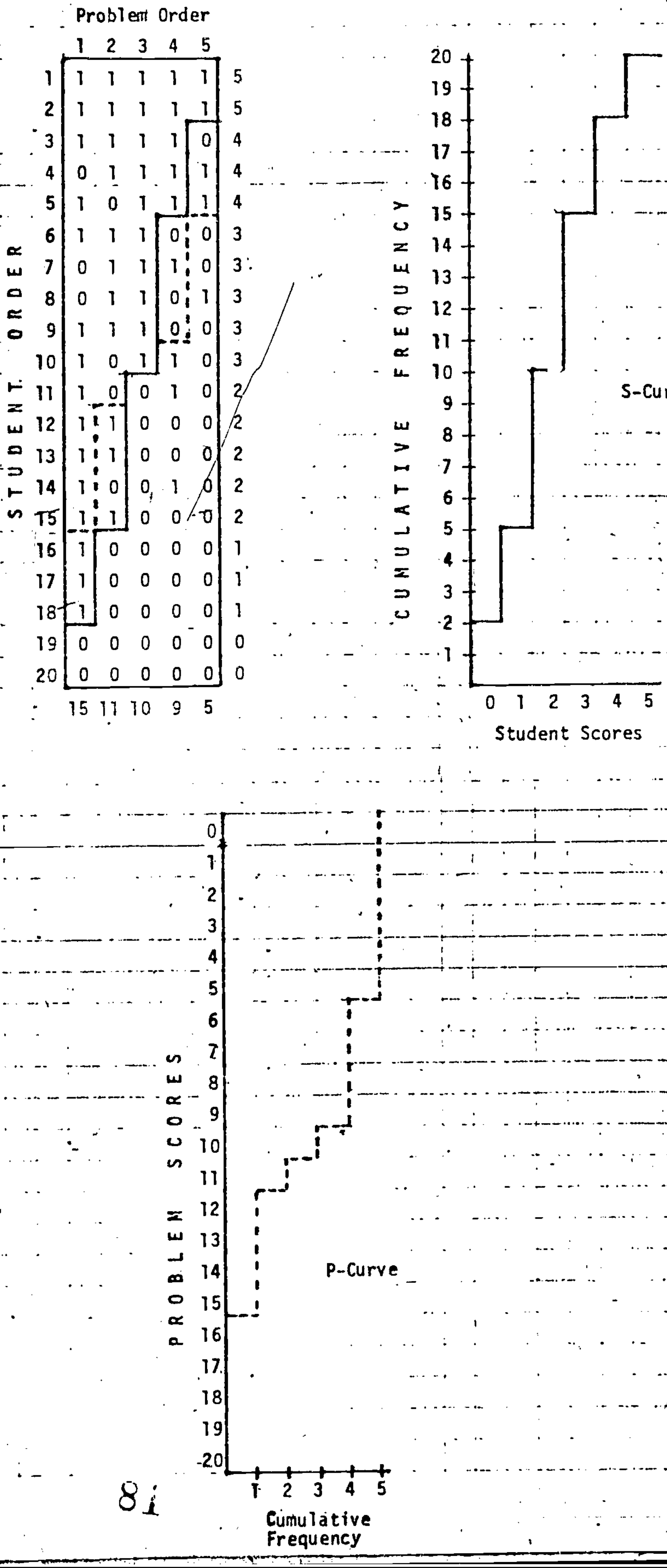
Sato (1980) developed yet another means for indexing departures from the perfect Guttman or cumulative scale. But this time the notion seems to have caught on. It is difficult to tell at this time whether it is the novelty of the procedure (and its more sophisticated mathematical basis) or whether more methodologists have begun to internalize the need to reconceptualize the proper measurement role of items. In any case, Sato's contribution reiterates the appropriate focus for understanding the measurement act, viz., the doubly ordered student-by-item (problem) matrix of raw responses (e.g., Figures 3b-5).

Interestingly, Sato's approach, unlike those discussed previously, utilizes a mathematical model of the ideal non-cumulative response pattern. An index of fit, then, is based on the extent of observed response pattern departure from the perfectly heterogeneous model. Specifically, any ordered student-by-problem (item) matrix can be partitioned into sections corresponding to the expected ideal cumulative patterns based on either the student scores, the S-curve, or problem scores (item difficulties), the P-curve.

Figure 7 depicts the process of analyzing the student-problem matrix in this manner. Figure 7 is simply Figure 4 again, but this time the cumulative student and problem score distributions are presented, separately, and superimposed, on the S-P matrix itself. As an exercise, superimpose

Figure 7

S-P Matrix and Cumulative Distributions
for Student Scores (S-Curve)
for Problem Scores (P-Curve)



	1	2	3	4	5	
1	1	1	1	1	1	5
2	1	1	1	1	1	5
3	1	1	1	1	0	4
4	0	1	1	1	1	4
5	1	0	1	1	1	4
6	1	1	1	0	0	3
7	0	1	1	1	0	3
8	0	1	1	0	1	3
9	1	1	1	0	0	3
10	1	0	1	1	0	3
11	1	0	0	1	0	2
12	1	1	0	0	0	2
13	1	1	0	0	0	2
14	1	0	0	1	0	2
15	1	1	0	0	0	2
16	1	0	0	0	0	1
17	1	0	0	0	0	1
18	1	0	0	0	0	1
19	0	0	0	0	0	0
20	0	0	0	0	0	0
	15	11	10	9	5	

the S-curves and P-curves appropriate for the matrices in Figures 3b and 5. You will discover that in the ideal case (Figure 3b) the S- and P-curves are coincident; and in the case of poor cumulative response pattern (Figure 5), the curves are quite far apart and much more so that they are for the moderately cumulative pattern exhibited here (and in Figure 4).

Thus, the area between the S- and P-curves--proportional to the number of student-item responses between the curves--reflects the degree of departure from the ideal cumulative response pattern. (In general, the number of student-item responses between the S- and P-curves is close to, but is not functionally related to, the total number of Guttman errors, viz., twice the number of 0's above, or 1's below, the S-curve.) To construct an index similar to the coefficient of scalability for Guttman scales, the maximum possible area between the S- and P-curves must be calculated for the perfectly heterogeneous student-problem response matrix of the same dimensions and mean performance. Sato models the ideal heterogeneous matrix by assuming simple binomial sampling for problems and students. Thus, the cumulative binomial distributions with parameters k and $\overline{p}$ and parameters n and $\overline{p}$ model the S- and P-curves respectively. Denoting the areas between the observed and binomial S- and P-curves as $A(n,k,\overline{p})$ and $A_B(n,k,\overline{p})$ respectively, Sato's disparity coefficient is given as follows:

$$D = \frac{A(n,k,\overline{p})}{A_B(n,k,\overline{p})} \qquad (39)$$

(A more computationally tractable estimate of D is given by Sato, 1980.)

This index reaches 1 in the case of perfect heterogeneity and 0 in the case of a perfect cumulative (homogeneous) response pattern. It therefore varies inversely (and I expect quite highly) with the other indices of homogeneity discussed in this section. Moreover, Sato (1980) defines analogous coefficients at the individual student and problem levels (called caution indices) which serve to highlight those students and items which depart considerably from ideal expectations. Loevinger (1947) developed a similar index for items whereas Guttman relied exclusively o: visual inspection of the response matrix. In the final analysis, the increasing popularity of Sato's approach is most likely due to the emphasis placed on the raw score matrix, with handy indices (for spotting aberrant cases) of great practical utility for the ordinary classroom teacher. For recent developments in the U. S., see Tatsuoka (1978), McArthur (1981), Harnisch and Linn (1981), and Miller (1981). (See also the chapter by McArthur in this monograph.)

Rasch Measurement: A Latent Trait Model

Latent trait theory, or item response theory (Lord, 1980), refers to a whole class of statistical measurement models based on the same fundamental conception of the measurement act guiding the cumulative models surveyed thus far. However, latent trait models make important allowances for those "minor" points we glossed over while drawing the analogy to the physical sciences. Specifically, these were the points relating to the variability of both the item difficulty positions as "hash marks" on the "ruler" and the underlying ability continuum itself, as one moves from one "ruler" to the next. For our purposes here, we will review only the simplest of the latent trait models, viz., the 1-parameter model, developed three decades ago by Georg Rasch. A number of good presentations and/or reviews of latent trait models generally, and the Rasch model in particular, currently exist. Some examples are: Rasch (1980 reprint of 1960 edition); Wright and Stone (1979); Hambleton and Cook (1977; see that entire issue of the <u>Journal of Educational Meament</u>); Lord (1980); and Traub and Wolfe (1981).

The Rasch model (and latent trait models generally) assumes a single invariant ability parameter and specifies a probability function over the entire 0-1 range that any item will be answered correctly by students of a given ability. Specifically, Rasch first approached the problem by imagining <u>independent</u> person and item parameters reflecting, respectively, <u>ability</u> and <u>difficulty</u> (or, its reciprocal, easiness). Second, he envisioned the same cumulative response pattern as the ideal outcome when persons with varying abilities encounter items of varying difficulties. But he modeled the process probabilistically, not only to avoid the determinism of previous approaches, but to establish an invariant measurement scale -- <u>so long as the model fits the empirical reality of the test data in question</u>.

The model he selected is a simple odds ratio, i.e., the odds (θ'_{si}) of student s with ability A_s correctly answering item i with difficulty D_i are given as

$$\theta'_{si} = \frac{A_s}{D_i} \qquad (40)$$

Instead of odds, we can use the more convenient 0-1 scale of probability. If P_{si} is the probability of student s answering item i correctly, tnen, by definition, $P_{si} = \theta_{si}/(1+\theta_{si})$. Thus equation (40) can be rewritten as

$$P_{si} = \frac{A_s}{D_i+A_s} \qquad (41)$$

It should be clear that, as hypothesized, the model predicts a lower chance of success for a student with lower ability encountering a relatively more difficult item, a higher chance of success for a student of higher ability encountering a relatively less difficult item, and a 50-50 chance of success when the ability of the student and the difficulty of the item are identical. These are invariant properties of the person and the item and are presumed to be independent of each other as well as of the other abilities of the persons being measured and the other difficulties of items doing the measuring. Again, this specific objectivity (as Rasch calls it) is operational only to the extent thac these presumptions fit the reality of the data.

Equation (40) becomes computaticnally more tractable as a simple linear function by taking the logarithm of both sides, i.e.,

$$\log (\theta_{si}) = \log (A_s) - \log (D_i) \qquad (42)$$

Likewise, equation (41) can be so converted; but it is usually expressed in exponential form using the natural base e and the substituted parameters $\alpha_s = \log (A_s)$ and $\delta_i = \log_e (D_i)$. In other words, $e^{\alpha_s} = A_s$ and $e^{\delta_i} = D_i$ and equation (41) becomes the so-called logistic function

$$P_{si} = \frac{e^{\alpha_s - \delta_i}}{1 + e^{\alpha_s - \delta_i}} \qquad (43)$$

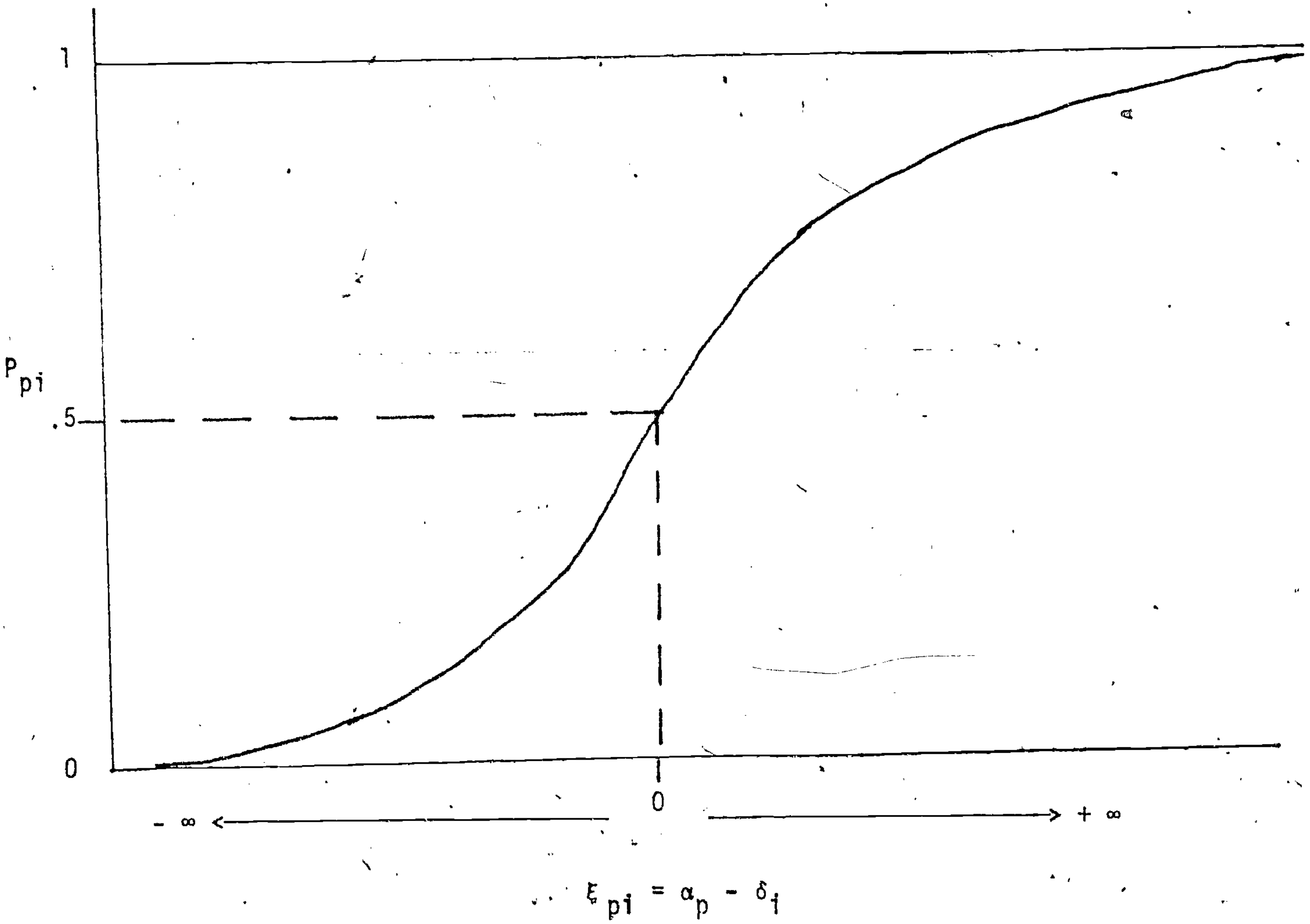
Of course, the same logic is embedded in (43) as was in (41), except now the interplay of person encountering item is reflected in the difference between the transformed ability parameter α_s and difficulty parameter δ_i. When equation (43) is graphed for all possible values of this difference, i.e., for $\xi_{si} = \alpha_s - \delta_i$ where $-\infty \leq \xi_{si} \leq +\infty$, the so-called response characteristic curve results (see Figure 8). This represents the simplest logistic model, often called the 1-parameter model, since P_{si} is really only dependent upon the single discrepancy ξ_{si}. Alternatively, for fixed difficulties δ_i or abilities α_s, the ogive in Figure 8 represents equally well the item characteristic or person characteristic curves respectively.

The rather elegant simplicity of the Rasch technique for scaling is realized through this important property of the model: the student raw scores (r_s) and observed item difficulties (p_i) are sufficient data from which to derive the best estimates of α_s and δ_i respectively. In effect, the double ordering of the student-by-item raw score matrix best estimates the ordering that would occur were we to know the actual α_s and δ_i. Thus, persons with the same raw score r from the same set of items will receive the same ability estimate α_r.

To estimate and α and δ, therefore, the n x k raw score matrix is merely collapsed row-wise such that rows now constitute the k+1 possible raw scores and cell entries are the proportions of persons in the rth raw score group correctly answering the ith item. If the index r is substituted for the

Figure 8

Item/Person Characteristic Curve



index s in equation (43), it should be clear from the above property that these cell proportions ($\hat{P}_{ri}$) are all estimates of their corresponding P_{ri}. In general, then, there are k(k+1) equations of the form

$$\hat{P}_{ri} = \frac{e^{\alpha_r - \delta_i}}{1 + e^{\alpha_r - \delta_i}}$$

with only 2k+1 unknown values of the α and δ.[10] (In practice, no information is provided by raw scores classes r = 0 or k or by observed item difficulties p = 0 or 1 and these rows and/or columns, should they occur, are eliminated for purposes of analysis.)

There are several approaches to the solution of these equations and testing the fit of the results to what the model predicts. (See references noted previously.) The important point for our argument here, however, is that this model again conforms to the measurement of a property as we ordinarily conceive of it. Moreover, when this particular model fits the data reasonably well, the parameter estimates of α and δ are reasonably independent of the particular ability and difficulty levels of specific student and item samples, thereby providing viable approaches to normally thorny testing problems such as test equating, item banking, tailored testing, and so forth.

Finally, it is interesting to note that for each person's ability estimate, there exists a so-called standard error estimate. But the only thing this estimate has in common with the standard error in traditional test theories is its name. The latent trait standard error is really based upon an information function that reflects the level of <u>precision</u> at the various ability calibrations. It bears no relationship whatsoever to any notion of item/test replicatior, i.e., accuracy (or dependability). Thus, the latent trait standard error is an index of precision and behaves accordingly, i.e., it is larger for ability estimates towards the extremes and lower for ability estimates towards the center of the item difficulty range.

Summary

To summarize the foregoing view and review, test theoreticians and practitioners must carefully distinguish their model of measurement from their model of the dependability of measurements. The former refers to the concept of precision that is applied in the construction of tests. The latter refers to the concept of accuracy that is applied to the result of testing under specified conditions of use. Items play a central role in measurement models; in models for dependability, they are of incidental importance insofar as the accuracy of estimated ability measurements is of primary importance. Clearly, truly useful test theories necessarily require both measurement and dependability models.

Classical (and classical-like) test theories are really models for the dependability of measurements. They are good for assessing the accuracy of the results of a testing process when the process is conceived as one (or several) of a great many (often infinite) measurement attempts. When each of the repeated measurements is conceived as a replicate (perfectly parallel) measure, we have classical test theory as originally developed. When the measurements are conceived as a random sample from a domain of interest (i.e., randomly parallel measures;, we have the item sampling versions of classical test theory. At the core of all of these theories, however, is the concept of repeated measurements. Whenever the results of behavioral assessments can be so conceived, classical test theories, in particular generalizability theory, enjoy a wide range of application. (See the recent review by Shavelson and Webb, 1981.)

But these test theories "dig their own grave" when they attempt to translate repeated measurements concepts to the internal structure of the test itself. Recasting items into the role of strictly parallel (or randomly parallel) measurements can't help but give rise to "test construction" procedures based on maximizing inter-item relationships. This procedure automatically eliminates items reflecting ability at the upper and lower ends of the "ruler." Thus, empirical evidence for internal consistency (in the reliability sense) or homogeneity/unidimensionality (in the construct validity sense) is based upon the wrong covariance structure.

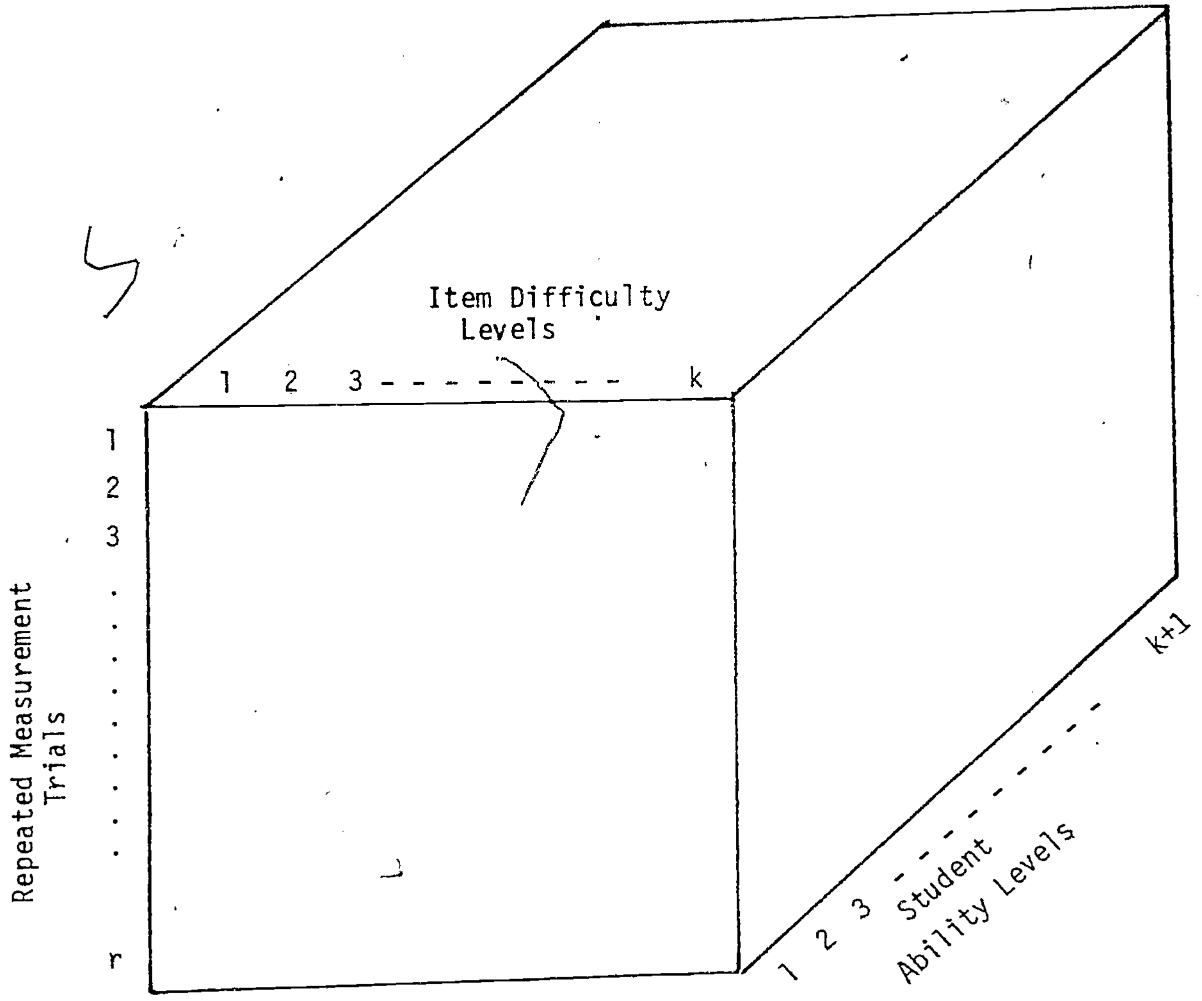
In constrast, measurement models attack the issue of test construction directly. They assume a singular construct from the start (relying primarily upon content validation) and proceed to develop items of varying difficulties analogous to hash marks on a ruler. To the extent that the set of items fits the cumulative response pattern expectation, we have evidence (necessary, but not sufficient) that our measurement goal has been achieved. Once satisfactorily constructed, it is quite appropriate that the instrument be subject to all relevant forms of dependability and validity procedures under the conditions for use in actual practice. These several ingredients comprise a complete test theory.

Moreover, it should be possible to incorporate dependability at the item level as well. The schematic in Figure 9 portrays the data box necessary to sort out -- at least in theory -- the contrasts between test precision and both item and test accuracy. Vertical slices of the data box contain the data necessary to assess the accuracy of items at each difficulty level for all ability levels. Horizontal slices contain the data necessary to test the scalability of items representing the difficulty levels for each replication. Cross slices could be used to assess the accuracy of items at the various difficulty levels holding ability constant. Collapsing the data box along the difficulty dimension produces the data matrix necessary for assessing accuracy at the test level. Of course, generalizability facets could be crossed or nested with the repeated measurement trials to assess accuracy (dependability) under different conditions. The complete empirical suggestion of Figure 9 may be quite intractable from an operational viewpoint, although, for some highly specifiable items domains (e.g., arithmetic fundamentals) on which ability varies systematically with other measurable examinee characteristics (e.g., age), it may not be too far-fetched.

In conclusion, classical test theory has probably enjoyed a long life not only because of psychological well-being through cognitive dissonance reduction, but because tests have never really been developed without variation in item difficulties. It is time now that we construct tests with varying item difficulties by design--not by happenstance--and use item analysis techniques that correspond to an appropriate theory of measurement. Moreover, it is fitting that this view forces upon us an issue of perhaps even greater importance, namely, the correspondence of item structure with

Figure 9

A Model for Contrasting Accuracy with Precision and Calibrating a Test of a Singular Achievement Construct


Item Difficulty Levels
1 2 3 - - - - - - - - - k
Repeated Measurement Trials
1
2
3
r
1 2 3 Student Ability Levels
k+1

the cognitive process to be assessed. (See, for example, the arguments recently advanced by Glaser, 1981.) It may well be that the simplistic ·ɔtions of dichotomous responses (right-wrong) to multiple choice or true-false items are unrealistic indicators of the cognitive processes underlying the abilities we try to measure. Different measurement models from those outlined here may offer more realistic solutions. (For example, see the recent latent class approaches such as Wilcox's (1981) answer-until-correct scheme.)

Footnotes

[1] I will use the term "traditional" to refer to classical and classical-like test theories, a distinction that will be clearer in the sequel.

[2] I have chosen Spearman's (1910) work, apparently inspired in 1908 by G. Udny Yule (see Yule, 1922), to mark the beginning date for classical test theory.

[3] It is important to note at the outset that I do not intend to extol any one notion of what it means to measure achievement. Rather, I wish to explicate a popular intuitive notion of measurement and the extent to which it is compatible with existing measurement theories.

[4] In general, I prefer the term "dependability" to the older term "reliability." As used in generalizability theory (Cronbach, et al., 1972), dependability denotes reliability under specified conditions of use. At times throughout this report, however, I will use the term "reliability" to facilitate the discussion of traditional test theory concepts.

[5] I am using the term "difficulty" here more in a parametric sense than as a synonym for observed p-values.

[6] The analogy could be improved upon in this regard by imagining the sticks to be subject to increases or decreases in length as a function of various and sundry effects (some random and some systematic) due to all aspects of the measurement context. This is a less sadistic equivalent of Lumsden's (1976) flogging wall test.

[7] Two classical test theory frameworks are in general use. One arises out of the definition of error as proposed originally by Spearman (1910). The other arises out of a definition of true scores as proposed originally by Brown (1910) and elaborated by Kelley (1924). The former approach is presented here since it's simpler. All derivations end up being the same so that it is a purely academic matter which approach is "better." See Gulliksen's (1950) seminal volume on classical test theory and the good historical overview by Tryon (1957).

Footnotes (continued)

8. An important caveat should be stated here: Except for the latent trait models, the illustrations I have selected do not in and of themselves provide sufficient information for calibrating items and estimating precision. Nevertheless, they are useful both historically and heuristically for underscoring the point of this discussion, viz., the contrast between dependability and measurement.

9. I am using the phrase "criterion-referenced testing" in the more profound sense rather than simply as a procedure for assessing a criterion level of performance. The criterion is, rather, the content and the attempted isomorphism between the content and the measurement rule. To quote Glaser (1963): "Criterion-referenced measures indicate the content of the behavioral repertory, and the correspondence between what an individual does and the underlying continuum of achievement." (p. 520)

10. Although useful for expository purposes here, this is not really the best procedure for estimating α and δ. (See the chapter by Choppin in this monograph.)

References

Allen, M. J., & Yen, W. M. Introduction to measurement theory. Monterey, CA: Brooks/Cole, 1979.

Bentler, P. M. Monotonicity analysis: An alternative to linear factor and test analysis. In D. R. Green, M. P. Ford & G. B. Flamer (Eds.), Measurement and Piaget. New York: McGraw Hill, 1971.

Brown, W. Some experimental results in the correlation of mental abilities. British Journal of Psychology, 1910, 3, 296-322.

Cornfield, J., & Tukey, J. W. Average values of mean squares in factorials. Annals of Mathematical Statistics, 1956, 27, 907-949.

Cox, D. R. The design of an experiment in which certain treatment arrangements are inadmissible. Biometrika, 1954, 40, 287-295.

Cronbach, L. J. Test "reliability": Its meaning and determination. Psychometrika, 1947, 12, 1-16.

Cronbach, L. J. Coefficient alpha and the internal structure of tests. Psychometrika, 1951, 16, 297-334.

Cronbach, L. J., Rajaratnam, N., & Gleser, G. C. Theory of generalizability: A liberation of reliability theory. British Journal of Statistical Psychology, 1963, 16, 137-163.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. The dependability of behavioral measurements. New York: John Wiley & Sons, 1972.

Festinger, L. The treatment of qualitative data by "scale analysis." Psychological Bulletin, 1947, 44, 149-161.

Ghiselli, E. E. Theory of psychological measurement. New York: McGraw Hill, 1964.

Glaser, R. Instructional technology and the measurement of learning outcomes. American Psychologist, 1963, 18, 519-521.

Glaser, R. The future of testing: A research agenda for cognitive psychology and psychometrics. American Psychologist, 1981, 36, 923-936.

Gulliksen, H. The relation of item difficulty and inter-item correlation to test variance and reliability. Psychometrika, 1945, 10, 79-91.

Gulliksen, H. Theory of mental tests. New York: John Wiley & Sons, 1950.

Guttman, L. A basis for scaling qualitative data. American Sociological Review, 1944, 9, 139-150.

Guttman, L. A basis for analyzing test-retest reliability. Psychometrika, 1945, 10, 255-282.

Hambleton, R. K., & Cook, L. L. Latent trait models and their use in the analysis of educational test data. Journal of Educational Measurement, 1977, 14, 75-96.

Harnisch, D. L., & Linn, R. L. Analysis of item response patterns: Questionable test data and dissimilar curriculum practices. Journal of Educational Measurement, 1981, 18, 133-146.

Horst, P. Correcting the Kuder-Richardson reliability for dispersion of item difficulties. Psychological Bulletin, 1953, 50, 371-374.

Horst, P. Psychological measurement and prediction. Belmont, CA: Wadsworth, 1966.

Hoyt, C. Test reliability estimated by analysis of variance. Psychometrika, 1941, 6, 153-160.

Kelly, T. L. Statistical methods. New York: Macmillan, 1924.

Kuder, G. F., & Richardson, M. W. The theory of the estimation of test reliability. Psychometrika, 1937, 2, 151-160.

Loevinger, J. A systematic approach to the construction and evaluation of tests of ability. Psychological Monographs, 1947, 61(4), Whole No. 285.

Loevinger, J. The technic of homogeneous tests compared with some aspects of "scale analysis" and factor analysis. Psychological Bulletin, 1948, 45, 507-529.

Loevinger, J. The attenuation paradox in test theory. Psychological Bulletin, 1954, 51, 493-504.

Lord, E. M. Applications of item response theory to practical testing problems. Hillsdale, N. J.: Lawrence Erlbaum Associates, 1980.

Lord, F. M., & Novick, M. R. Statistical theories of mental test scores. Reading, Mass.: Addison-Wesley, 1968.

Lumsden, J. The construction of unidimensional tests. Psychological Bulletin, 1961, 58, 122-131.

Lumsden, J. Test theory. In M. R. Rosenzweig & L. W. Porter (Eds.), Annual Review of Psychology (Volume 27). Palo Alto, CA: Annual Reviews, Inc., 1976.

Magnusson, D. Test theory. Reading, Mass.: Addison-Wesley, 1967.

Maxwell, A. E. A statistical approach to scalogram analysis. Educational and Psychological Measurement, 1959, 19, 337-349.

Menzel, H. A new coefficient for scalogram analysis. Public Opinion Quarterly, 1953, 17, 268-280.

Miller, M. D. Measuring between-group differences in instruction. Unpublished doctoral dissertation, University of California, Los Angeles, 1981.

Novick, M. R. The axioms and principal results of classical test theory. Journal of Mathematical Psychology, 1966, 3, 1-18.

Popham, W. J., & Husek, T. R. Implications of criterion-referenced measurement. Journal of Educational Measurement, 1969, 6, 1-9.

Rasch. G. Probabilistic models for some intelligence and attainment tests. Chicago: The University of Chicago Press, 1980. (Originally published in 1960 by the Danish Institute for Educational Research.)

Rulon, P. J. A simplified procedure for determining the reliability of a test by split-halves. Harvard Educational Review, 1939, 9, 99-103.

Sagi, P. C. A statistical test for the significance of a coefficient of reproducibility. Psychometrika, 1959, 24, 19-27.

Sato, T. The S-P chart and the caution index. NEC (Nippon Electic Company) Educational Information Bulletin. Japan: Computer and Communication Systems Research Laboratories, 1980.

Schuessler, K. F. A note on statistical significance of scalogram. Sociometry, 1961, 24, 312-318.

Shavelson, R. J., & Webb, N. M. Generalizability theory: 1973-1980. British Journal of Mathematical and Statistical Psychology, 1981, 34, 133-166.

Spearman, C. Correlation calculated with faulty data. British Journal of Psychology, 1910, 3, 271-295.

Stevens, S. S. Mathematics, measurement, and psychophysics. In S. S. Stevens (Ed.) Handbook of experimental psychology. New York: Wiley, 1951.

Tatsuoka, M. M. Recent psychometric developments in Japan: Engineers grapple with educational measurement problems. Paper presented at the Office of Naval Research Contractor's Meeting on Individualized Measurement, Columbus, Missouri, 1978.

TenHouten, W. D. Scale gradient analysis: A statistical method for constructing and evaluating Guttman scales. Sociometry, 1969, 32, 80-98.

Torgerson, W. S. Theory and methods of scaling. New York: John Wiley and Sons, 1958.

Traub, R. E., & Wolf, R. G. Latent trait theories and the assessment of educational achievement. In D. C. Berliner (Ed.), Review of Research in Education (Volume 9). American Education Research Association, 1981.

Tryon, R. C. Reliability and behavior domain validity: Reformulation and historical critique. Psychological Bulletin, 1957, 54, 229-249.

Walker, D. A. Answer-pattern and score-scatter in tests and examinations. British Journal of Psychology, 1931, 20, 73-86.

Walker, D. A. Answer-pattern and score-scatter in tests and examinations. British Journal of Psychology, 1936, 26, 301-308.

Walker, D. A. Answer-pattern and score-scatter in tests and examinations. British Journal of Psychology, 1940, 30, 248-260.

Wilcox, R. R. Solving measurement problems with an answer-until-correct scoring procedure. Applied Psychological Measurement, 1981, 5, 399-414.

Wright, B. D. Sample-free test calibration and person measurement. In Proceedings of the 1967 Invitational Conference on Testing Problems. Princeton, N. J.: Educational Testing Service, 1968.

Wright, B. D. Solving measurement problems with the Rasch model. Journal of Educational Measurement, 1977, 14, 97-116.

Wright, B. D., & Panchapakeson, N. A procedure for sample-free item analysis. Educational and Psychological Measurement, 1969, 29, 23-48.

Wright, B. D., & Stone, M. H. Best test design. Chicago: Mesa Press, 1979.

Yule, G. U. On the methods of measuring association between two attributes. Journal of the Royal Statistical Society, 1912, 75, 579-642.

Yule, G. U. An introduction to the theory of statistics. London: Charles Griffin and Co., 1922.

ANALYSIS OF PATTERNS: THE S-P TECHNIQUE

David McArthur
Center for the Study of Evaluation, UCLA

Definition of the model. A system of analyzing patterns of student responses called Student-Problem score table analysis has been developed over the last decade by a group of educational researchers in Japan (Sato, 1974, 1975, 1980, 1981a, 1981b; Sato and Kurata, 1977; Kurata and Sato, 1981; Sato, Takeya, Kurata, Morimoto and Chimura, 1981). While the mathematics associated with derivative indices in this system are relatively complex, the S-P system itself is predicated on a simple reconfiguring of test scores. Rather similar analyses of student performance on educational tests can be found in the professional literature of a half-century ago, but recent developments by Sato and colleagues represent significant improvements both in concept and execution. The method appears to hold a number of possibilities for effective and unambiguous analysis of test score patterns across subjects within a classroom, items within a test, and, by extension, to separate groups of respondents. It is a versatile contribution to the field of testing, containing minimal requirements for sample size, prior scoring, item scaling, and the like. The S-P model lends itself to extensions into polychotomous scoring analysis of multiple patterns, and analysis of patterns of item bias.

Test scores are placed in a matrix in which rows represent individual repsondents' responses to a set of items, and columns represent the responses given by a group of respondents to a set of items. The usual (and most convenient) entries in this matrix are zeros for wrong answers and ones for correct answers. Total correct

Figure 1
S-P Chart for a Six Item Test Administered to 20 Students

Items in ascending order of difficulty

rank	1	2	3	4	5	6
item #	1	5	4	2	3	6

Average passing rate p = .425
Discrepancy D* = .525

Students, in descending order of total score								Total Correct	Caution Index for Students
Rank	I.D.#								C_i*
1	02	1	1	1	1	0	0:	4	0.000
2	04	1	1	1	1	0	0:	4	0.000
3	05	1	1	1	0	1:	0	4	0.000
4	11	1	1	0	1	1:	0	4	0.034
5	12	1	0	1	0	1:	1	4	0.552 *
6	14	1	1	1	1	0:	0	4	0.000
7	20	1	1	1	0	1:	0	4	0.000
8	22	1	1	1	0	1:	0	4	0.000
9	23	1	1	1	0	0:	1	4	0.276
10	07	1	1	0	0	1:	0	3	0.033
11	17	1	1	0:	0	1	0	3	0.033
12	19	1	1:	0	1	0	0	3	0.033
13	27	1	1:	0	1	0	0	3	0.033
14	29	0	1:	1	0	1	0	3	0.433 *
15	03	1	0:	0	0	1	0	2	0.276
16	06	0	1:	0	0	1	0	2	0.448 *
17	08	1	1:	0	0	0	0	2	0.000
18	10	1	1:	0	0	0	0	2	0.000
19	15	1:	0	1	0	0	0	2	0.241
20	16	1:	0	0	1	0	0	2	0.276
21	21	1:	0	0	1	0	0	2	0.276
22	28	1:	0	1	0	0	0	2	0.241
23	01	:1	0	0	0	0	0	1	0.000
24	09	:0	1	0	0	0	0	1	0.238
25	13	:1	0	0	0	0	0	1	0.000
26	18	:0	0	0	1	0	0	1	0.619 *
27	24	:0	0	0	1	0	0	1	0.619 *
28	25	:0	1	0	0	0	0	1	0.238
29	26	:1	0	0	0	0	0	1	0.000
ITEM TOTALS:		23	18	11	10	10	02		
C_j* caution index for items		0.167	0.148	0.111	0.423 *	0.231	0.000		

101

* High caution index for unusual response pattern.

scores are calculated for each respondent, and total number of correct responses are tallied for each item. Rows are reordered by descending total number of correct responses; columns are reordered by ascending order of difficulty of items. The resulting matrix has several aspects which are particularly convenient for a detailed appraisal of respondents or items, singly or collectively. A short example, annotated and indexed with several computations to be explained below, is shown on the following page.

Two cumulative ogives are drawn over the matrix to form the framework for further analysis. Because the data is discrete, the ogives take on a stair-step appearance, but both can be thought of as approximations to curves which describe in summary form the two distinct patterns embedded in the data. The first is a curve reflecting respondents' performance as shown by their total scores; the second is a similarly overlaid ogive curve reflecting item difficulties. In one special circumstance, the two curves describe only one pattern: if the matrix of items and respondents is perfectly matched in the sense of a Guttman scale, both of the curves overlap exactly. All of the correct responses would be to the upper left while all of the incorrect responses would be to the lower right. However, as the occurance of either unanticipated errors by respondents with high scores or unanticipated successes by respondents with low scores increases, or as the pattern of responses becomes increasingly random, the respondent or student curve (S-curve) and the item or problem curve (P-curve) become increasingly discrepant. Sato has developed an index which evaluates the degree of discrepancy or lack of conformation between the S- and P-curves. This index will be zero in the special case of perfectly ordered sets, and will approach 1.0 for the case of totally random data.

For any respondent, on for any item, taken individually, the pattern of scores reflects that row or column <u>in relation to</u> the pattern established by the configuration of sorted rows and columns. For any given individual respondent or single item, the response pattern may be "perfectly ordered" in the sense used above. The row or column shares a symmetry with the associated row or column marginal; in the case of dichotomous data this symmetry is seen in a high positive point-biserial correlation. As the match between patterns declines-- that is, as the row or column under consideration shares less and less in common with the associated marginal formed from all rows or all columns--the point-biserial also declines. Unfortunately, r_{pbis} is not independent of the proportions within the data and never reaches 1.0 in practice. Cases of complete "symmetry" between row or column and the corresponding marginal which happen to differ in proportions do not yield the same correlation coefficients.

An index which is stable across differing proportions is Sato's Caution Index C, which gives a value of 0 in the condition of "perfect symmetry" between row or column and row marginal or column marginal. As unanticipated successes or failures increase and "symmetry" declines, the index increases (a modification of the Caution Index, called C*, has an upper bound of 1.0). Thus a very high index value is associated with a respondent or item for which the pattern of obtained responses is very discrepant from the overall pattern established by all members of the set.

Harnisch and Linn (1982) present the modified Caution Index as follows:

$$C_i^* = \frac{\sum_{j=1}^{n_{i.}} (1 - u_{ij}) n_{.j} - \sum_{j=n_{i.}+1}^{J} u_{ij}\, n_{.j}}{\sum_{j=1}^{n_{i.}} n_{.j} - \sum_{j=J+1-n_{i.}}^{J} n_{.j}}$$

where $i = 1,2,\ldots,I$ indexes the examinee,

$j = 1,2,\ldots,J$ indexes the item,

$u_{ij} = 1$ if the respondent i answers item j incorrectly,

0 if the respondent i answers item j incorrectly,

$n_{i.}$ = total correct for the i^th respondent, and

n_{ij} = total number of correct responses to the j^th item.

Harnisch and Linn explain that the name of the index comes from the notion that a large value is associated with respondents that have unusual response patterns. It suggests that some caution may be needed in interpreting a total correct score for these individuals. An unusual response pattern may result from guessing, carelessness, high anxiety, an unusual instructional history or other experiential set, a localized misunderstanding that influences responses to a subset of items, or copying a neighbor's answers to certain questions.

A large value may also suggest that some individuals have acquired skills in an order which is not characteristic of the whole group. The index says nothing about the most able respondents with perfect total scores, because the "symmetry" condition is met. More importantly, if a respondent gets no item correct whatsoever, both the total score and the caution index will be zero since, again, the "symmetry" condition is met; in this situation the available information about the respondent is insufficient to make any useful diagnosis. Most persons, though, will achieve total scores between the extremes and for them the caution index provides information that is not contained in the total score. A large value of the caution index raises doubts about the validity of the usual interpretation of the total score for an individual.

A related development is a modification of the Caution Index to examine patterns of responses to clusters or subtest scores and an "ideal" pattern of scores of individual subtests, the perfect Guttman pattern (Fujita and Nagaoka, 1974, in Sato, 1981).

Sato has developed an index of discrepancy to evaluate the degree to which the S and P curves do not conform either to one another or to the Guttman scale. Except in the case of perfectly ordered sets there is always some degree of discrepancy between curves. The index is explained as follows:

$$D^* = \frac{A(I,J,p)}{A_B(I,J,p)}$$

> where the numerator is the area between the S curve and the P curve in the given S-P chart for a group of I students who took J-problem test and got an average problem-passing rate p, and $A_B(I,J,p)$ is the area between the two curves as modeled by cumulative binomial distributions with parameters I,J, and p, respectively (Sato, 1980, p. 15; indices rewritten for consistency with notation of Harnisch & Linn).

The denominator is a function which expresses a truly random pattern of responses for a test with a given number of subjects, given number of items, and given average passing rate, while the numerator reflects the obtained pattern for that test. As the value of this ratio approaches 1.0, it portrays an increasingly random pattern of responses. For the perfect Guttman scale, the numerator will be 0 and thus D* will be 0. The computation of D* is functionally derived from a model of random responses, but its exact mathematical properties have not been investigated thoroughly.

Also available, but not yet studied in detail, is an index of "entropy" associated with distributions of total scores for students choosing different answers to the same question. This index explores the particular pattern of responses (right answer and all distractors included), in the context of overall correct score totals for these responses.

While most of the published work using the S-P method has concentrated on binary data (0 for wrong answer, 1 for right answer), and calculations are most tractable in that form, the indices developed from the configuration of S- and P-curves are not limited to such data. The technique can extended to multi-level scoring (see Possible Extensions to the model, below).

Measurement philosophy. A precursor to the S-P method is the concept of "higgledy-piggledy" (or "hig" for short) suggested by Thomson about 1930 and elaborated by Walker in a trio of contributions (1931, 1936, 1940), but evidently carried no further by educational researchers at that time. Walker examined right/wrong answers to a set of independent items with particular reference to score-scatter, which had been a focus of attention since the early twenties. Where scatter reflects random behaviors on the part of examinees, "hig" is said to be present. However,

> By a test being unig (the converse of hig) we mean that each score x is composed of correct answers to x easiest questions, and therefore to no other questions. Hig implies a departure from this composition. Note that it is not sufficient for our purposes to define unig by stipulating that every score x is identical in composition--there must be added the condition that it is composed of the x easiest items; in other words the score x + 1 always compromises the x items of the score x, and one more. Now if hig is absent, that is each score is unig, it is easy to show that an exact relationship exists between the n's of the answer-pattern and the N's of the score scatter (1931, p.75).

The parallel to Guttman scaling, while the latter is far more mathematically rigorous, is obvious; Sato's indices appear to address the same underlying concepts.

Guttman's (1944) statistical model for the analysis of attitudinal data was formulated to solve scaling problems in the context of morale assessment for the U.S. Army. While the initial approaches were not at all technically sophisticated and involved much sorting of paper by hand, Guttman's conceptualization was powerful; the scalogram approach, and especially its mathematical underpinnings, received extensive development during the 1950's. But by 1959, Maxwell had expressed rather strong disappointment with the narrow range of application these procedures had enjoyed, and suggested two general statistics which might serve to dissolve the arbitrary distinction between qualitative and quantitative scales, and, at the same time, reduce some of the cumbersome calculations. (One of these statistics is a regression coefficient developed from the residual between observations and perfect patterns of responses to a given set of items, which bears some conceptual resemblance to Sato's D*.) However, the primary audience for these technical contributions appears to have been educational statisticians and researchers. Only infrequently was attention given to simplifying the techniques for a broader potential audience (Green's (1956) contribution is one exception, although published in a highly sophisticated journal).

Many of the publications by Sato and colleagues in Japan seem geared directly to end-users, teachers in the classroom who, with the S-P method and handscoring or microcomputer processing, can analyze their own instructional data for purposes of understnading their students' comprehension and modifying their own instruction. The overarching concern of the Educational Measurement and Evaluation Group at the Nippon Electric Company's Computer and Communication Systems Research Laboratories has been development and dissemination of readily understandable and adaptable procedures. Evidently it has proved popular in a variety of classroom settings in Japan, and has been applied to the following areas:

- test scoring and feedback to each examinee about his/her own performance on a test
- feedback to the instructor about both individual and group performance
- analysis of types of errors made by students
- analysis of instructional process and hierarchies of instructional units
- item analysis, rating scale analysis, questionnaire analysis
- test score simulations
- development of individual performance profiles across repeated testings

Two characteristics are shared by all of these approaches: first, the central focus of the study is the degree to which items and/or respondents are heterogeneous, and second, the actual element of raw data (say, 0 or 1) is assumed to be best understood in terms of its position in a matrix with orderly properties. Interestingly, the article by Green (1956) noted above forms the only overt link between the S-P method and earlier work in English on analysis of response patterns.

Where the S-P method diverge from its predecessors can be seen in the very reduced role played by probability theory, and the absence of anything resembling tests of statistical significance (a shortcoming addressed below). Much of the work on the S-P method is either in Japanese or in English-language journals not generally available in the West. In the U.S. the small number of research presentations using the S-P method to date is small (Harnisch, 1980; Harnisch & Linn, 1981, 1982; McArthur, 1982; Tatsuoka, 1978; Tatsuoka & Tatsuoka, 1980).

Assumptions made by the model. The S-P method starts from a complete matrix of scores, doubly reordered by I rows and J columns. The model applies equally well to the trivial case of a 2 x 2 matrix, and to 2 x J and I x 2 retangular matrices; it also appears to have no functional upper limit on the number of items or respondents. However, missing data cannot be incorporated effectively. That is, each respondent and item must have complete data since all calculations are made with reference to i and j as constant values. For purposes of reordering, if two or more respondents have the same total score their ranks are tied but their positions within the sorted matrix must be unique, so ties between marginals are resolved arbitrarily (a situation which could cause some small instability in the S and P curves). In respect to both individual scores and sets of scores taken as a whole, no explicit probabilistic formulation is involved, although underlying the analysis of the matrix is a model premised on cumulative binomial or beta binomial distributions, with parameters I (number of cases), J (number of items), and p (average passing rate). No study has been made of how guessing affects the obtained pattern of responses, nor how corrections for guessing might affect the S-P chart. Because of the very small number of assumptions made by the model, its interpretation does not require a strong theoretical background, and in fact can be annotated easily by computer as an aid to the user novice. Indeed, the graphic reordering with overlay of S- and P-curves but no further statistics appears sufficient to allow teachers, with use of a brief nontechnical reference guide, to make well-reasoned instructional decisions.

One implicit assumption deserves special attention. In the derivation of a caution index for item or respondent, the entire existing configuration of I items and J respondents, whether valid or not, enters into consideration. That is, because the frame of reference does not extend beyond the data at hand, the derivative indices are inherently subject to limits on their analytic utility. However, it is important to recognize that for the great bulk of practical testing applications, such limitations in fact may be advantageous. Each index also depends on a linear interpretation of steps between marginal totals, although it is readily demonstrable that substitution of a highly discriminating item for a weakly discriminating one, <u>or</u> a very able examinee for a poor one, can alter many of the indices for <u>both</u> persons and items. Additionally, the linearity constraint treats all data elements within the matrix equally, despite unknown (and perhaps inestimable) contributions from chance correct responses. On the other hand, without further tests of significance, the resulting statistical uncertainties, which are small under most conditions, have little practical importance in the usual classroom situation.

<u>Strengths and weaknesses.</u> Obvious strengths of the S-P system are its simplicity, wide potential audience, and portability. The code required for computer processing can be exceptionally brief and with the increased availability of microcomputers, can be delivered to the classroom teacher directly. According to Harnisch and Linn (1982), the caution indices compare well with Cliff's (1977) C_{i1} and C_{i2}, Mokken's (1971) H^*_i, Tatsuoka and Tatsuoka's (1980) Norm Conformity Index (NCI), and van der Flier's (1977) U', all of which are harder to calculate as a rule. As an inherently flexible system, it appears to be suitable for a variety of test types, and for a range of analyses even within the same test. The novice user need not master the full range of calculations in order to make excellent use of more elementary portions of the results. A sophisticated user can easily iterate selectively through an existing data set, choosing particular items or persons not meeting some criterion for performance, and recasting the remaining matrix into a revised chart. Under certain conditions, addressed below, the method can be adapted to examination of test bias (McArthur, 1982).

Weaknesses include the following three general criticisms. No substantive body of psychometric or educational theory preceded the development of practical applications of the model because in fact its development was not paradigm-driven. Instead, the S-P techniques arose in response to a perceived need for classroom teachers to have a readily interpretable, minimally complex tool for test analysis. Thus, at present little can be said regarding questions of reliability, validity, true scores, scaling theory, or quality of measurement. No extant work addresses either the problem of signal/noise ratio or of model fit. The absence of a strong theoretical base dampens the development of rationally interconnected research hypotheses, although the method offers ample opportunities for direct investigation of individual performance and item characteristics. The absence of strong theory-derived hypotheses leaves a recognizable gap in the ability to draw strong inferences from the S-P method. That is, in developing a diagnostic interpretation of a student's score pattern, the teacher or researcher must make a conscious effort to balance the evidence in light of some uncertainty about what constitutes critical or significant departure from the expected.

These weaknesses do not affect the classroom teacher to any major degree. In the classroom, the technique is used for confirming knowledge about individual students gained in the course of interaction with the class, and/or to confirm that items on a particular test are reasonably well suited to the class. From the researcher's viewpoint, the weaknesses constitute rather important blocks to further development. On the other hand, because of some points of similarity between the S-P technique and less arcane aspects of a number of existing models, hypothesis building tends to proceed anyway. The absence of recognizable criteria for establishing statistical significances for degree of heterogeneity is an important technical problem. Because the various indices appear to share a great deal in common with indices having known statistical properties from other research models, an initial direction for such effort would be to examine these parallels.[1]

Present areas of application. All of the published studies in English to date utilize the S-P method exclusively in the context or right/wrong (1/0) scoring. These studies each use data collected from multiple-choice tests (generally reading or math) administered to primary or secondary level students. In this body of literature the general application is either to the task of individual student analysis, or more frequently, to item analysis. With an appropriate microcomputer--one marketed exclusively in Japan is configured exclusively for the purposes of the S-P method--classroom teachers can use the technique interactively. Science teachers in Japan are evidently the largest cluster of users, although details about acceptance and daily utilization remain sketchy.

A different application arises in the context of large-scale assessment. Harnisch (personal communication) reports that several school districts have contracted for S-P analysis of mid-year and final achievement test scores. Several thousand individuals tested on dozens of items pose no new conceptual or mathematical complexity and in this situation the results can be used to address both item-level and aggregate-level questions.

[1] Strong parallels also can be found with aspects of the analysis of planar Wiener processes and spatial patterns, from the domain of mathematical geophysics.

Possible extensions of the model. Three new directions for the S-P method are being explored. The first is the application of iterative procedures, first suggested by Green (1956) in a brief paragraph on p-tuple analysis of Guttman scales. Zimmer (1982) has collected extensive developmental data on children's perception of various tasks and attributions; this data incorporates multiple discrete levels of performance arranged, according to theory, in a logical staircase ascendency. P-tuple iterative analyis by the S-P procedure appears to offer answers to three questions: a) does broad sample of children respond in an orderly amnner to the range of tasks; b) does such order reflect known characteristics of the sample (viz. developmental level as measured on standardized procedures); and c) do deviations from the symmetrical relationship between the developmental complexity of the task and the developmental level of the child reflect consistent support for one or another competing theory of development. For this data, separate S-P analyses were made with the first developmental level scored 0 and all others 1, then the first two levels scored 0 and all others 1, and so on. Stability of person order and item order, uniformity of the staircase intervals, and relationships between item difficulty and item complexity can be studied. Preliminary evidence suggests that the S-P method provides a system of analysis for such multi-level data that exceeds the explanatory power of several extant procedures.

In p-tuple analysis, which makes use of repeated passes through data, some questions of a technical nature are unresolved at this time. For example, it is clear that successive reorderings can perturb the positional stability of any one respondent within the matrix or any one task within the matrix, to some degree. However, changes in ordering contribute to changes in the S-P indices, and whether such changes, and/or linearity assumptions and violations therein, play an important role is also under study in the context of this developmental data. Another way to think of this problem is to imagine a single matrix of persons x items with the S-P chart from each developmental level overlaid. The most difficult tasks would be accomplished only by the most developmentally advanced individuals, and below a certain competence (i.e. the highest S-curve on this compound chart) virtually no one would be expected to succeed on those tasks. The ordering of those participants who fail at all tasks of that difficulty level is arbitrary, because their total score for these most difficult tasks is zero. But their ordering would not be arbitrary on tasks of moderate or low difficulty, at which more successes might be anticipated and the corresponding S-curves would be located lower on the chart. What constitutes acceptable and interpretable slippage of this kind needs further probing. Perhaps the best analogy is to the term "seiche," drawn from the field of oceanography: it refers to regular, entirely predictable tidal motions occuring within confined bodies of water. Such seiche in a polychotomous S-P chart ought to show itself totally consistent and predictable.

The second area for development of the S-P method is in the realm of scalar data, for which a number of statistical assumptions have been developed. An example is signal detection analysis, in which the "raw element" of data is once again a 0/1 response, this time for absence or presence of perceived stimulus. A variety of complex statistical techniques have been used to investigate how such stimuli, presented across a range of intensities over a repeated number of trials, are processed by the receiver. The analog in S-P analysis might best be portrayed as a three-dimensional matrix of persons, items, and repeated trials. Items are not necessarily objectively identical from trial to trial, and responses are tempered by not one but several possible orderly progressions. Such three-dimensional and higher-dimensional data challenges the S-P method to provide cohesive summary statistics which can be evaluated probabilistically.

An extension of the S-P technique to the study of test bias has been made by McArthur (1982). Where two distinct groups have been tested on the same instrument or on two instruments one of which is an exact translation of the other, S-P analysis offers an interesting alternative to the complex techniques for detection of biased items generally in use. McArthur studied the response patterns for items on the California Test of Basic Skills, administered to both English-speaking and Spanish-speaking children, the latter taking the CTBS-Espanol. Even when proportions of children achieving correct responses to a given item differ between the two language groups, the item may not be biased. However, the D* values for the student-problem matrices calculated separately for the two groups suggest that the Spanish-language group engaged in more random responding than did their English-speaking counterparts. A significantly larger number of items for the fromer group show that those children above the P-curve (children who in a case of "symmetry" as defined earlier would be expected to do well) who gave the correct response were frequently fewer in number than the corresponding sample from the English-language group. That is, deleting cases below the P-curve, which are more likely to have engaged in random responding, leaves a finite number of respondents for whom the prediction of success is high. Obviously on easier items this reduced sample is larger than for difficult items because of the shape of the P-curve. Nonetheless, while the p values for a given item may differ significantly between one group and the other, the proportions of right answers above the P-curves can be statistically identical. To establish evidence of bias, the additional requirement is that for students in the disadvantaged group who by their pattern of performance on the test as a whole should have succeeded with a particular item, that item generated erroneous responding for one group more than for another.

REFERENCES

Cliff, N. A theory of consistency of ordering generalizable to tailored testing. Psychometrika, 1977, 42, 375-399.

Fujita, T., & Nagaoka, K. Arbitrary Ho full-marked S-P table. Institute of Electronic Communication Engineers of Japan, 1974 (In Japanese).

Green, B.F. A method of scalogram analysis using summary statistics. Psychometrika, 1956, 21, 79-88.

Guttman, L. A basis for scaling quantitative data. American Sociological Review, 1944, 9, 139-150.

Harnisch, D.L., & Linn, R.L. Analysis of item response patterns: Questionable test data and dissimilar curriculum practices. Journal of Educational Measurement, 1981, 18, 133-146.

Harnisch, D.L., & Linn, R.L. Identification of abberant response patterns. Champaign, Illinois: University of Illinois, 1982. National Institue of Education Grant No. G-80-0003, Final Report.

Kurata, T, & Sato, T. Similarity of some indices of item response patterns based on an S-P chart. Computer and Communication Systems Research Laboratories, Nippon Electric Company, Research Memorandum E181-4, 1981.

Maxwell, A.E. A statistical approach to scalogram analysis. Educational and Psychological Measurement, 1959, 19, 337-349.

McArthur, D.L. Detection of item bias using analyses of response patterns. Paper presented to the Annual meeting of the American Educational Research Association, New York, 1982.

Mokken, R.J. A theory of procedure of scale analysis. The Hague: Mouton, 1971.

Sato, T. A classroom information system for teachers, with focus on the instructional data collection and analysis. Association for Computer Machinery Proceedings, 1974, 199-206.

Sato, T. Analysis of students' pattern of response to individual subtests. Computer and Communications Systems Research Laboratories, Nippon Electric Company, Research Memorandum E181-2, 1981a.

Sato, T. Similarity of some indices of item response patterns. Computer and Communications Research Laboratories, Nippon Electric Company, *Research Memorandum* E181-1, 1981b.

Sato, T. *The construction and interpretation of S-P tables.* Tokyo: Meiji Tosho, 1975 (In Japanese).

Sato, T. The S-P chart and the caution index. Nippon Electric Company, *Educational Informatics Bulletin*, 1980.

Sato, T., & Kurata, M. Basic S-P score table characteristics. *NEC Research and Development*, 1977, *47*, 64-71.

Sato, T., Takeya, M., Kurata, M., Morimoto, Y., & Chimura, H. An instructional data analysis machine with a microprocessor -- SPEEDY. *NEC Research and Development*, 1981, *61*, 55-63.

Tatsuoka, M.M. *Recent psychometric developments in Japan: Engineers grapple with educational measurement problems.* Paper presented at the Office of Naval Research Contractors' Meeting on Individualized Measurement, Columbia, Missouri, 1978.

Tatsuoka, M.M., & Tatsuoka, K. Detection of abberant response patterns and their effects on dimensionality. Computer-based Education Research Laboratory, University of Illinois, *Research Report* 80-4, 1980.

van der Flier, H. Environmental factors and deviant response patterns. In Y.H. Poortinga (Ed.), *Basic problems in cross-cultural psychology.* Amsterdam: Swets and Seitlinger, 1977.

Walker, D.A. Answer-pattern and score-scatter in tests and examinations. *British Journal of Psychology*, 1931, *22*, 73-86; 1936, *26*, 301-308; 1940, *30*, 248-260.

Zimmer, J.M. Analysis of developmental levels of children. University of California, Santa Barbara, 1982. In preparation.

THE RASCH MODEL FOR ITEM ANALYSIS

Bruce Choppin
Center for the Study of Evaluation, UCLA

1. Definition of the Model

The so-called Rasch model now widely employed for item analysis, is only one of a complete family of models described by Rasch in his 1960 text. All may be properly called "Rasch Models" since they share a common feature which Rasch labeled "specific objectivity". This is a property of most measurement systems which requires that the comparison of any two objects that have been measured shall not depend upon which measuring instrument or instruments were used. It is a familiar feature of many everyday physical measurements (length, time, weight, etc.). In the context of mental testing, it means that the comparison of two individuals who have been tested should be independent of which items were included in the tests. Traditional test analysis based on "true scores" does not have this property since "scores" on one test cannot be directly compared to "scores" on another. (The peculiar virtues of specific objectivity and the conditions needed to achieve it are discussed later in this chapter.)

Mathematical Representation

The Rasch model is a mathematical formulation linking the probability of the outcome when a single person attempts a single item to the characteristics of the person and the item. It is thus one of the family of latent-trait models for the measurement of achievement, and is arguably the least complex member of this family. In its simplest form it can be written:

$$\text{Probability } [X_{vi} = 1] = \frac{A_v}{A_v + D_i}$$

where, X_{vi} takes the value 1 if person v responds correctly to item i, and zero otherwise,

A_v is a parameter describing the ability of person v,

and D_i is a parameter describing the difficulty of item i.

In this formulation, A and D may vary from 0 to ∞. A transformation of these parameters is usually introduced to simplify much of the mathematical analysis. This defines new parameters for person ability (α) and item difficulty (δ) to satisfy the equations:

$A_v = W^{\alpha_v}$ and $D_i = W^{\delta_i}$ for some constant W.

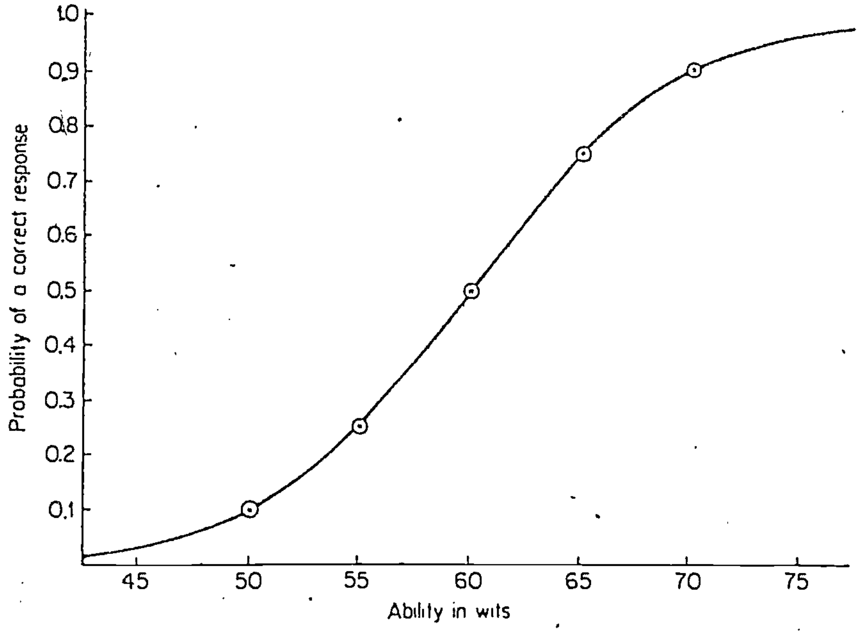


Figure 1 : Item Characteristic Curve (wits) for the Rasch Model

A further simplification, introduced by Rasch himself and used widely in the literature, is to fix the constant W to the natural logarithmic base, e. In this case the model can be written:

$$\text{(2)} \quad \text{Probability } [X_{vi} = 1] = \frac{e^t}{1 + e^t}, \text{ where } t = (\alpha_y - \delta_i).$$

In this formulation, α and δ can take all real values and measure ability and difficulty respectively on the same "logit" scale. The sign of the expression ($\alpha - \delta$) in any particular instance indicates the probable outcome of the person-item interaction. If $\alpha > \delta$ then the most probable outcome is a correct response. If $\alpha < \delta$ then the most likely outcome is an incorrect response. It should also be noted that the "odds" for getting a correct response (defined as the ratio of the probability for getting one to the probability for not getting one) take on a particularly simple form:

$$\text{Odds } [X_{vi} = 1] = \frac{\dfrac{e^t}{1 + e^t}}{1 - \dfrac{e^t}{1 + e^t}} = e^t$$

$$\text{or } t = \log_e(\text{odds})$$

For this reason, the Rasch model is sometimes referred to as the "log-odds" model.

Alternative Units

As stated above, the model based on the exponential function yields measures of people and items on a natural scale, whose unit is called a "logit". Rasch himself used the model in this form, and most of Wright's publications also make use of it. Mathematically and computationally the logit is convenient, but as an operational unit it has two drawbacks. First, a change in achievement of one logit represents a considerable amount of learning. Studies in various parts of the world indicate that in a given subject area, the typical child's achievement level would rise by rather less than half a logit in a typical school year. In practice, many of the differences in achievement level that we need to measure are much less than this, as is the precision yielded by our tests, so results are commonly expressed as decimal fractions rather than as integers.

Secondly, logits are usually ranged around a mean of zero (this is a matter of convention rather than necessity) so that half of all the values obtained for parameters are typically negative. In general, teachers dislike dealing with negative numbers, and the prospect of having to explain to an anxious parent what Jimmy's change in math achievement from -1.83 logits to -1.15 logits actually means is too much for most of them.

The solution for practical applications of the Rasch scaling technique is to use a smaller and more convenient unit. This is accomplished by setting W to some value other than e. A number of alternatives have been suggested, but the unit in the widest use after the logit is obtained by setting $W = 3^{0.2}$. This unit is known as the "wit" in the United Kingdom and United States, and as the "bryte" in Australia. Wits are typically centered around 50 with a range from about 30 to 70. One logit is equal to 4.55 wits. For many purposes it is sufficient to report wits as integers. The particular value for W is chosen so as to provide a set of easily memorized probability values, as can be seen in the Table 1.

Table 1
The Relationship of Logits and Wits to the
Probability of Correct Response

(α - δ) Measured in Logits	(α - δ) Measured in Wits	Probability of a Correct Response
-2.198	-10	0.10
-1.099	-5	0.25
0	0	0.50
+1.099	+5	0.75
+2.198	+10	0.90

It must be emphasized that the choice of a unit for reporting is an arbitrary matter. Most of the theoretical work on the model, and all the computer programs for parameter estimation in common use, work in logits--translating to wits or some other scale for reporting only if desired.

Analytic Possiblities

Parameter estimation is a difficult issue in latent-trait theories. That for Rasch model a variety of different estimation algorithms (at least six) have become available in the last fifteen years results from the mathematical simplicity of the Rasch formulation.

The basic equation models only the outcome of one particular item-person interaction, but since it does so in terms of a probability function, it is necessary to accumulate data from several such interactions in order to estimate parameters or test the fit of the model itself.

For example, the accumulation of responses of one individual to a set of items may be used to estimate the ability parameter for the individual, and the pattern of responses by several individuals to two items may be used to estimate the relative difficulty of the two items. From a (persons-by-items) response matrix it is possible to estimate both sets of parameters (abilities and difficulties), and also to check on whether the model is an acceptable generating function for the data. This calibration of items, and the test of goodness-of-fit to the model, correspond to item analysis procedures in classical test theory (but see section 5(a)).

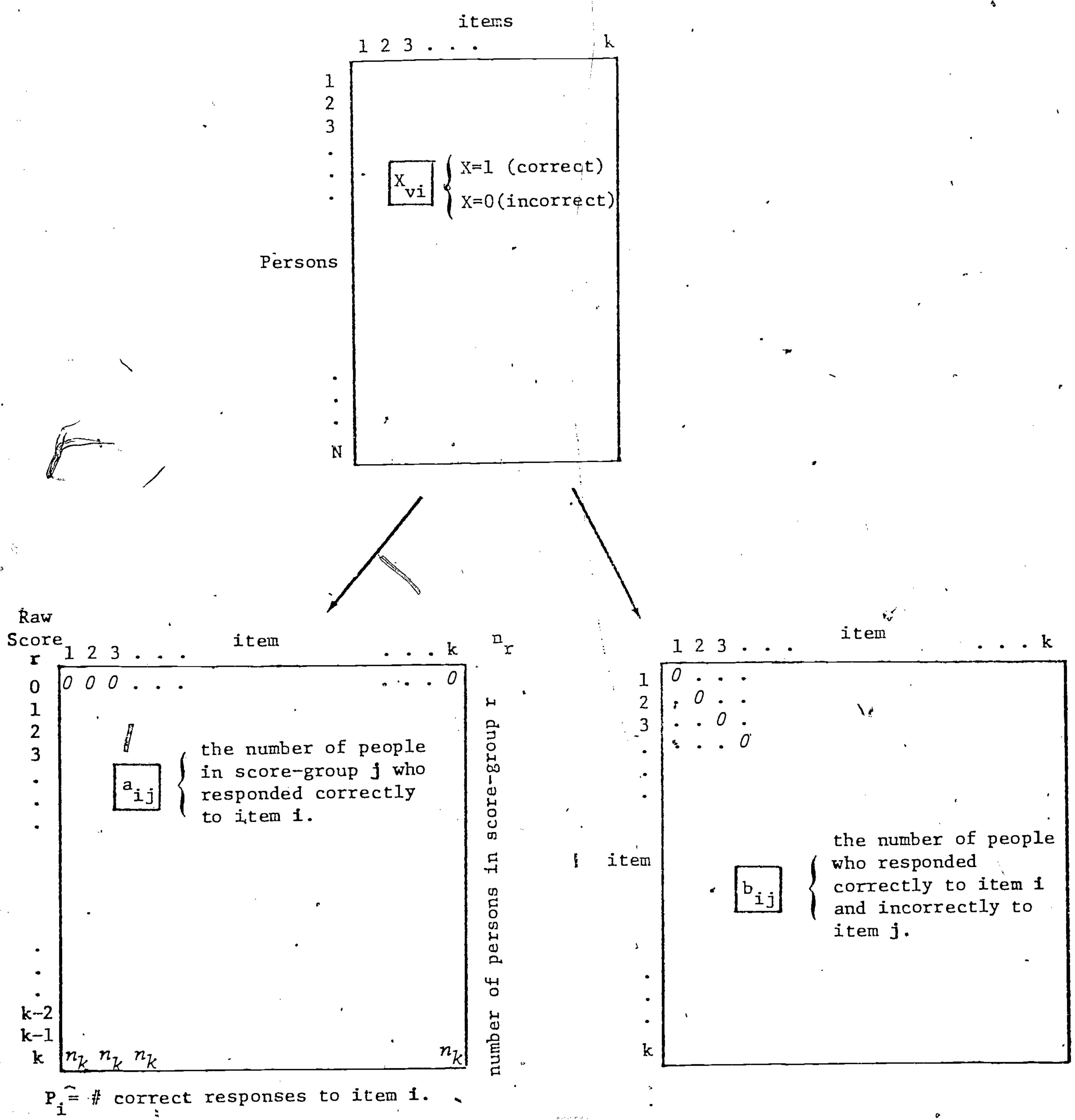
Once items have been calibrated, equations can be developed to predict the characteristics of tests composed of different samples of previously calibrated items, or the performance of previously measured people on new items. Although the simplest approach to statistical analysis requires a complete rectangular persons-by-items response matrix, other procedures are available to handle alternative data structures. For example, when a group of individuals take different but overlapping tests, the persons-by-items matrix will necessarily be incomplete, but it is still possible to calibrate the items and measure the people. An extreme example, in which a computer-managed adaptive test is individually tailored to each testee (such that the next item given depends on the responses to previous items), may lead to a situation in which every person tested may respond to a unique set of items. If the items have been calibrated in advance, it is possible to estimate the individual's ability parameter at each step of the sequence, and to discontinue testing when the ability has been measured with the desired degree of precision.

Estimation Techniques

Although this paper is not the place for a detailed presentation of the algebraic manipulation involved in the various algorithms for parameter estimation, an outline of the different approaches may be helpful.

Conventionally the starting point is taken to be a rectangular matrix of **persons** by **items** in which the elements are one if a particular person responded correctly to the appropriate item, zero if he responded incorrectly, and blank if the person was not presented with the item. Initially we shall restrict the discussion to complete matrices of ones and zeros such as occur when a group of **N** people all attempt a test of **k** items. In most applications **N** is usually much larger than **k** . Two summarizations of data contained in the **N x k** matrix leads to effective strategies for parameter estimation (see Figure 2).

One, known as the "score-group method" clusters together all those persons who had a particular raw score, and then counts within each cluster the number of correct responses to each item. This


items
1 2 3 . . . k
1
2
3
N
Persons
X vi
X=1 (correct)
X=0(incorrect)
Raw Score r
item
1 2 3 . . . . . . k
0 0 0 0
0
1
2
3
k-2
k-1
k
a ij
the number of people in score-group j who responded correctly to item i.
n k n k n k n k
n r
number of persons in score-group r
P i = # correct responses to item i.
item
1 2 3 . . . . . . k
1
2
3
0
0
0
0
item
b ij
the number of people who responded correctly to item i and incorrectly to item j.
k

2a : Score-group Summarization

2b : Pair-wise Item Summarization

Figure 2 : Data reduction strategies for Rasch parameter estimation.

produces a score-group by item matrix as in Figure 2A. The other method considers the items two at a time, and counts for each pair the number of persons who responded correctly to the first but incorrectly to the second. This is known as the "pair-wise" approach and produces an item by item matrix as in Figure 2B. (A parallel analysis comparing the people two at a time can be developed theoretically, but has found little practical application.) Both the score-group and the pair-wise approaches are described by Rasch in his 1960 book, but without the development of a maximum likelihood technique he was unable to exploit them.

The score-group method produces a (k + 1) by k matrix, but since raw scores of zero and k do not contribute to the estimation procedure, the summary yields k(k - 1) elements for use in the estimation algorithm. The pair-wise approach results in a k by k matrix in which the leading diagonal elements are always zero, so again there are k(k - 1) elements in the summary on which the estimation algorithm operates.

Analysis of score-group matrix to separate information on α and δ and thus obtain fully conditioned estimates for both the item difficulty parameters and the abilities associated with membership of score-group 1 through k - 1 is computationally demanding and expensive. The best available procedure has been programmed by Gustafsson (1977), but, though mathematically elegant and statistically sound, it is far too expensive for routine use.

However, Wright has shown that estimates developed from the margins of the score-group matrix can be developed very easily using a maximum likelihood approach. Though the simultaneous estimation of both α and δ sets of parameters introduces a bias, a simple expansion factor applied to the results can largely correct for this (Wright & Douglas, 1977; Habermann, 1977), and this method is widely used in practice. When the data are summarized in a score-group fashion, they are convenient for checking the assumption of equal discriminating power between items and the tests of fit developed by Wright and Mead (1976) concentrate on this.

By contrast, the pair-wise approach separates information about the δ's from information about the α's at the beginning. The matrix of counts summarized in Figure 2B has conditioned out all information about variations in α, so that a fully conditional estimate of the δ's (either by maximum likelihood or least squares) can be obtained. The ability estimates for each individual are developed from solving iteratively the equation:

$$r - \sum_{i=1}^{k} \frac{W^{\alpha-\delta_i}}{1 + W^{\alpha-\delta_i}} = 0$$

where r is the raw score of the person, and the summation extends only over those items that were attempted.

The test of fit applied to the pair-wise summary matrix is not very sensitive to violations of the equal discrimination power assumption (see section 3), but instead focuses on the issue of local independence between items (Choppin & Wright, in progress). In practice, therefore, the two approaches may be regarded as complementary.

Though slower than the Wright estimation algorithm based on score-group marginals, the pair-wise approach has the considerable advantage of being able to handle incomplete data matrices--corresponding to all those applications in which not every person attempts every item. It is thus of particular interest in such fields as adaptive testing and item banking (Choppin, 1978, 1982).

2. The Measurement Philosophy and Primary Focus of Interest

Although it turns out that the mathematical details have much in common with those of "item response theory", Rasch derived his models from a very different standpoint. In the first paragraph of the preface to the book which launched his ideas on measurement (Rasch, 1960) he quotes approvingly an attack by B.F. Skinner on the application of conventional statistical procedures to psychological research.

> "The order to be found in human and animal behavior should be extracted from investigations into individuals ... psychometric methods are inadequate for such purposes since they deal with groups of individuals." (Skinner, 1956, p. 221)

Group-centered statistics, which form the backbone of conventional psychometric practice (factor analysis, analysis of variance, etc.) require the clustering of individuals into discrete categories or populations, and further make assumptions about the nature of variation within these categories which Rasch viewed with grave distaste. The alternative was to develop methods which would work with individuals.

> "Individual-centered statistical techniques require models in which each individual is characterized separately and from which, given adequate data, the individual parameters can be estimated. It is further essential that comparisons between individuals become independent of which particular instruments - tests, or items or other stimuli - within the class considered have been used. Symmetrically, it ought to be possible to compare stimuli belonging to the same class - measuring the same thing - independent of which particular individuals within the class considered were instrumental for the comparison." (Rasch, 1960, p. vii)

In this excursion into what he later calls "specific objectivity", Rasch is echoing a theme developed explicitly by L.L. Thurstone three decades earlier:

> "A measuring instrument must not be seriously affected in its measuring function by the object of measurement. To the extent that its measurement function is so affected, the validity of the instrument is impaired or limited. If a yardstick measured differently because of the fact that it was a rug, a picture, or a piece of paper that was being measured, then to that extent the trustworthiness of that yardstick as a measuring device would be impaired. Within the range of objects for which the measuring instrument is intended its function must be independent of the object of measurement. " (Thurstone, 1928, p.547).

Reliance on this form of analogy to the physical sciences is quite characteristic of latent trait measurement theorists. Wright (1968, 1977) also uses the yardstick as a convenient metaphor for a test item. Others (Eysenck, 1979; Choppin, 1979, 1982) have pointed out the similarities between the measurement of mental traits and the measurement of temperature. The underlying premise is that although psychological measurement maybe rather more difficult to accomplish than is measurement in the fields of physics and chemistry, the same general principles should apply. Features which are characteristic of good measurement techniques in physics should also be found in the fields of psychology and education.

Rasch himself draws out the similarity between the development of his model, and Maxwell's analysis of Newton's laws of motion in terms of the concepts force and mass (Maxwell, 1876). The second law links force, mass and acceleration in a situation where although acceleration and its measurement have been fully discussed, the concepts mass and force are not yet defined. Rasch (1960, pp. 110-114) considers the necessity of defining the two concepts in terms of each other, and shows how appropriate manipulation of the mathematical model (the "law") and the collection of suitable data can lead to the (comparative) measurement of masses, and the (comparative) measurement of forces. He points out the close analogy to his item-response model which links ability, difficulty and probability. Ability and difficulty require related definitions since people need tasks on which to demonstrate their ability, and tasks only exhibit their difficulty when attempted by people. Since his model is "specifically objective", data can be collected so that the two sets of parameters are capable of separate estimation (as with force and mass).

This approach to measurement is the primary focus of interest for the Rasch model. Individuals are to be measured through the estimation of parameters characterizing their performance. These parameters shall be interpretable by comparison with the parameters estimated for other individuals (as in norm-referencing) and/or in conjunction with the parameter estimates for test stimuli (as in criterion-referencing).

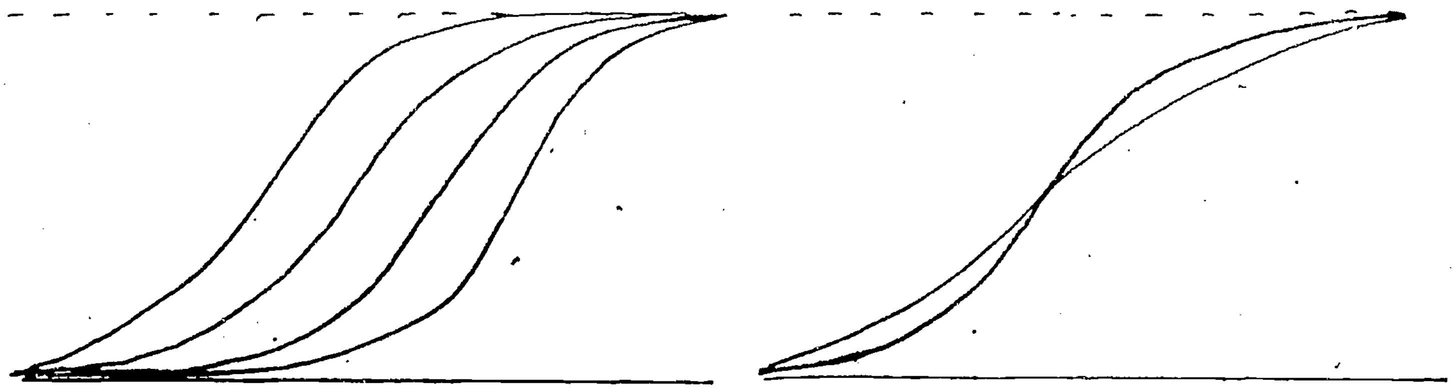
3. Assumptions made by the Rasch Model

The basic assumption is a simple yet powerful one that derives from the requirement of specific objectivity, so central to Rasch's thinking about measurement. It is that the set of people to be measured, and the set of tasks (items) used to measure them, can each be uniquely ordered in terms respectively of their ability and difficulty. (Ability and difficulty as already described.) This ordering permits a parameterization of people and tasks that fits the simple model defined in section 1 above.

The basic assumption has a number of important implications. One such assumption is that of local independence. The probability of a particular individual responding correctly to a particular item must not depend upon the responses that have been made to the previous items. If it did, then altering the sequence of items that made up a particular test, would alter the ordering of people on the underlying trait (in violation of the basic assumption). Similarly, local independence requires that the response of an individual to a particular item is not affected by the responses given by other people to the same item. If it were, then it would be possible, by selective clustering of people, to change the ordering of items in terms of their difficulty (in violation of the basic assumption).

Another implication that follows from the basic assumption of the model is sometimes stated (rather confusingly) as "equality of discrimination". It must be emphasized that this does not mean that all items are assumed to have equal point-biserial correlation indices with total test score, or with some external criterion. Rather, it means that the signal/noise ratio represented by the maximum slope of the characteristic curve of each item is assumed to be the same for all items. If the slopes were not the same, then at some point the characteristic curves for two items would cross. This would mean that the ordering of the items in terms of difficulty for persons of lower ability would not be the same as the ordering for persons of higher ability (see Figure 3). This again violates the basic assumption.

Figure 3



(a) Characteristic curves for items that fit the Rasch Model.

(b) Characteristic curves for two items with different discriminations.

Uni-dimensionality is also a consequence of the basic assumption. If the performance of people on a set of items depended on their individual standing on two or more latent traits, such that the ordering of people on these latent traits was not identical, then it would be impossible to represent the interaction of person and task with a single person parameter for ability.

A further assumption and one which is mathematically very convenient, albeit somewhat unrealistic (at least on multiple-choice items), is that there is no random guessing behavior. The model requires that for any test item, the probability of a successful response tends asymptotically to zero as the ability of the person attempting it is reduced (see Figure 1).

Similarly, there is a built in assumption, which has been much less carefully explored, that as the ability of the person being considered increases, the probability of a successful response to any given item approaches one.

4. Strengths and Weaknesses and Gaps in the Development

The strong features of the Rasch model when compared with other measurement models are:

(a) The combination of specific objectivity, a property taken for granted in the field of physical measurement, and the model's mathematical simplicity.

(b) Deriving from this, the separability property which permits the estimation of person-parameters and item-parameters separately.

(c) The existence of several algorithms for parameter estimation some of which are extremely fast and which work well with small amounts of data.

(d) The inbuilt flexibility of the system. As with other latent trait models which are defined at the item level, there is no requirement that tests be of a fixed length or contain the same items.

(e) The close parallels that exist between the Rasch model and the conventional practice of calculating raw scores based on an equal weighting of items. Rasch models are the only latent-trait models for which the raw score, as conventionally defined, is a sufficient statistic for ability (and correspondingly the raw difficulty or p-value of an item is a sufficient statistic for Rasch difficulty).

Against this it must be admitted that there are areas of considerable weakness. The most serious focuses on the assumptions made by the model. These are, in general, too strong to carry full credibility. In practice some real data appear to fit the model rather poorly. The assumptions of local independence and of no guessing (which are crucial to the model) are not strictly met in practice. Although the psychometrician may be able to reduce the guessing problem through the avoidance of objective items, and may be able to structure the test and the conditions under which it is administered to improve local independence, in real life situations these problems are rarely completely eliminated. The model also demands (as do most others) uni-dimensionality (or, as Rasch calls it, conformability), and while the items that comprise many existing tests fail to meet this criterion, the problem is less critical. If one has control over the test construction phase of a measurement program, then it is possible to build sets of items which satisfy the uni-dimensionality assumption moderately well.

One feature of the model which has been described as a weakness (Goldstein, 1979; Divgi, 1981) is that it implies a unique ordering of items, in terms of their difficulty, for all individuals. This appears not to be sufficiently sensitive to the effects of instructional and curriculum variation, ~~and~~ stands, therefore, as an important criticism (but see Bryce, 1981).

The seriousness with which such objections need to be considered depends upon the nature of the measurement task being addressed. Most educational instruction programs aim at increasing the learning of the student and thus at increasing his ability to solve relevant test items. We would usually expect the ability to solve all relevant test items to increase--but the relative difficulty of the items could (and normally would) remain unchanged. While this is the dominant goal of instruction, the model can handle the situation appropriately, and the occasional changes in relative difficulty brought about by alternative curricula (see, for example, Engel, 1976 or Choppin, 1978) can shed considerable light on the real effects of the instructional program. If, however, a section of curriculum is aimed specifically at breaking down some piece of learning and replacing it with another (i.e. making some items more difficult to solve, and other easier) such as may occur during revolutionary changes in society, then we may well feel that the simple model proposed is inadequate to describe the situation. In this case the items measuring the "old" learning and the "new" do not seem to belong on the same scale. Such circumstances, however, are not routine in the United States.

Similarly, we find in general that the ordering of item difficulties is the same with respect to all students. Where one student differs significantly in finding some item much harder or easier than predicted by the model, then we have valuable diagnostic information about that individual (Mead, 1975). In practice we rarely find evidence for such differences, and where they do occur the interpretation is usually clear and direct (for example, the student missed instruction on a particular topic). If we were attempting to measure in an area where there was no common ordering of item difficulties for most students, then the model would appear quite inappropriate. Such situations may be simulated by creating test items whose solution depends upon luck or chance, but this is far removed from purposive educational testing.

Experience over the last two decades suggests that the simplification made by the model in requiring a unique ordering of items is met adequatelyin practice. Deviations, where they do occur, are indicators of the need for further investigation (Dobby & Duckworth, 1979; Choppin, 1977). There seems little reason, therefore, to regard this as a weakness of the Rasch approach.

5. Areas of Application

The basic form of the model proposed by Rasch, and described in section 1, dealt with the simplified situation where only two possible outcomes of a person attempting a test item were considered (i.e the response is scored "right" or "wrong"). For this reason, perhaps, most of the applications so far developed have been confined to the use of "objective" test items for the measurement of achievement since these are most naturally scored in this fashion.

(a) Item Analysis

The most frequent application of the model has been for item analysis. Users have wanted to confirm that the model fits data they have already accumulated for existing tests; they seek clues as to why particular tests are not functioning as well as they should; or in the construction of new tests they seek guidance as to which items to include and which to omit.

It is probably true to say, however, that the Rasch model has not proved particularly valuable in any of these three roles. It can detect lack of homogenity among items, but is probably less sensitive to this than is factor analysis. It can identify items that do not discriminate or for which perhaps the wrong score key has been selected, but it seems no more effective at this than is the more traditional form of item analysis. The exception to this generalization probably comes when tests are being tailored for a very specific purpose. Wright and Stone explore this in "Best Test Design" (1979). Careful adherence to all the steps they outline would probably yield a test with better characteristics for the specific and intended purpose than would a test produced on the basis of only traditional forms of item analysis and the crude criteria they employ.

(b) Scaling and Equating

A serious problem of traditional testing is that the "score" produced can only be interpreted in terms of the particular test used. The development of norms for standardized tests is an attempt to overcome this problem but this too has serious limitations. Latent trait scaling has been used to tackle this question directly. With the Rasch model, the raw scores on one test are mapped onto their latent trait scale, and different tests can of course have their scores mapped onto the same scale (provided always that the dimension of ability being measured is the same). The method has been used to compare "quasi-parallel" tests (e.g., Woodcock, 1973; Willmott & Fowles, 1974); to link the tests given at different stages of a longitudinal study (Engel, 1976; Choppin, 1978); and to check on the standardization characteristics of batteries of published tests (Rentz & Bashaw, 1976, 1977).

It should perhaps be noted that although equating using the Rasch model appears more flexible than traditional procedures in that only the difficulty level of the two tests is being compared and other characteristics such as test length, the distribution of item difficulties, etc. maybe quite different, there is an implicit assumption that the "discrimination power" (in the sense discussed above) of the items in the two tests are comparable. As a rule this implies that the item types are similar. Attempts to use the Rasch model to equate multiple choice and essay type tests on the same topic have led to inconsistent and bizarre results (Willmott, 1979; Vincent, 1980).

(c) Item Banking

Item banks take the equating of test scores to its logical limit by calibrating all possible performances on all possible tests composed of items drawn from a fixed set (the bank).

> When a family of test items is constructed so that they can be calibrated along a single common dimension and when they are employed so that they retain these calibrations over a useful realm of application, then a scientific tool of great simplicity and far reaching potential becomes available. The "bank" of calibrated items can serve the composition of a wide variety of measuring tests. The tests can be short or long, easy or hard, wide in scope or sharp in focus. (Wright, 1980).

An item bank requires calibration, and although in theory there are alternative approaches, in practice the Rasch model has proved by far the most cost effective and is the most widely used (Choppin, 1979).

(d) Quality of Measurement.

An important development that is facilitated by latent trait scaling is the calculation of an index to indicate the quality of measurement for each set of test data, and if necessary for each person attempting a test or for each item. The Rasch model, for example, yields an explicit probability for each possible outcome of every interaction of a person and an item. Where, overall, the probabilities of the observed outcomes are too low we may deduce that for some reason the Rasch model does not offer an adequate description of a particular set of data. If the probabilities are generally in the acceptable range, but are low for a particular item, then we may conclude that this is an unsatisfactory item. Perhaps it does not discriminate, or is addressing some different dimension of achievement. If the probabilities are generally acceptable but are low for a specific person, then we may conclude that this person was not adequately measured by the test (perhaps he guessed at random, was insufficiently motivated, or misunderstood the use of the answer sheet). The reporting for this person of a low measurement quality index would imply that the person's score should be disregarded and that a retest is appropriate.

A recent extension of this approach involves trying to identify within the vector of item responses from a particular individual those portions which provide reliable measurement information, on which items (or groups of items) the subject appears to have guessed at random, and how the total vector of responses may be selectively edited in order to provide a more reliable estimate of the subject's level of achievement.

6. Extensions to the Basic Model

Two types of adaptation and extension will be considered here. The first centers around the notion of sequential testing in which evidence of the level of ability of the subject is accumulated in Bayesian fashion during the test session and may be used to determine which items are to be attempted at the next point of the sequence and/or when to terminate the testing session. This approach relies upon the existence of difficulty calibrations for a pool or bank of test items. Most of the reseach that has been done so far has employed computers to manage the testing session: to select items for the subject to answer, to keep track of measurement quality, to generate up-to-date estimates of the ability of the subject (together with the appropriate standard errors) and to decide when the session should be terminated. Wright and Stone (1979) point out that individual people can do most of this for themselves if provided with suitable guidelines and computational aids, and in many circumstances making the learner responsible for evaluating his own learning is a useful thing to do.

The second area of development from the basic Rasch model is in the extension from simple dichotomous scoring of items (right-wrong) to a more complex system. Two separate situations need to be considered. The first is when an item is not answered completely but enough is done to earn some partial credit. Data would then consist of scores in the range 0 to 1 for each item. The other case is that which typically occurs with rating scales or attitude measures when the respondent is asked to choose one from among a finite number of discrete categories, and each category contains information about the standing of the respondent on some latent trait. Douglas (1982) has considered the theoretical implications of generalizing the basic Rasch model to include both these cases, and it turns out that almost everything that can be done for dichotomous items can also be done for these more complex methods of scoring. For the rating scale problem both Andrich (1977) and Wright and Masters (1982) have found it convenient to concentrate on establishing the location of thresholds (the point at which the probability for reponding in one category passes the probability of responding in the next one - Figure 4). Wright and Masters have produced some interesting theorems about the importance of these thresholds being properly ordered, and about the spacing of thresholds that maximizes the information gained. There have been few practical applications of this approach to date.

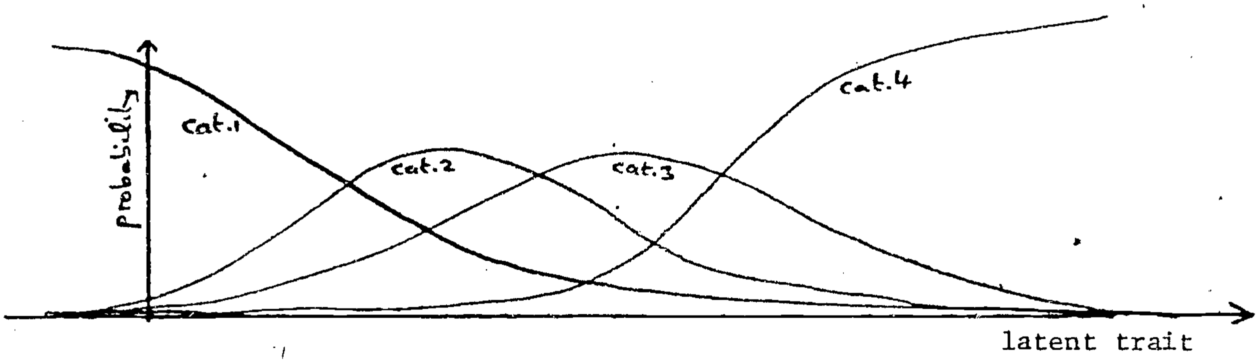


Figure 4 : The Probability of Responding in various categories.

For the analysis of "partial credit" data two computer programs (CREDIT by Masters and POLYPAIR by Choppin) have been devised and applied to real data sets. The latter program, for example, was used in the assessment of writing skills which forms part of the British National Assessment Program.

7. Points of Controversy

In some ways the Rasch model represents a revolutionary approach to educational measurement that discards many time-honored constructs in testing theory (e.g., true score, measurement error, and reliability). On the other hand, it can be viewed as providing a comprehensive and sound mathematical underpinning for the conventional practice of using raw scores, and shows that in most testing applications raw scores are all that are required. From this point of view the Rasch model may be seen as less radical than other latent trait models. Perhaps because the former view of the model was the first to catch the imagination in the United States and has dominated efforts to popularize it, it has been a subject of continuing controversy. The most strident arguments are not concerned with how best to use the Rasch model, but whether or not its use is ever appropriate.

To some extent the Rasch model has been central in the general attack on latent trait theory as applied to the measurement of student achievement. Goldstein (1979) who has led this attack on the other side of the Atlantic, stresses the fundamental difference between what he regards as well-ordered traits such as aptitude and intelligence on the one hand, and the complex pattern of behaviors that we call educational achievement on the other. In his view it makes no sense to apply any unidimensional model to the assessment of achievement.

Less extreme in their implications are the arguments within the latent trait camp about whether the Rasch (i.e., one-parameter) model is adequate for achievement testing, or whether a more complex (usually three-parameter) model is indicated.

It is important to differentiate two kinds of usage. One is in test construction where in general the users of Rasch models appear to be on firm ground in claiming that a strategy to develop and select items that conform to the Rasch model will produce better test instruments than would other more conventional strategies. The other type of usage is concerned with the analysis of existing test data (for example, the massive data sets of NAEP or the accumulated files of SAT material at ETS) where items are likely to be so varied (and in many cases so poor) that it is comparatively easy to show that the Rasch model is not appropriate. Devotees of the Rasch model react to this by dropping the non-fitting items (which may well be the majority) and working with those that are left--but this cavalier approach does not commend itself to many researchers. If one is interested in analyzing and scaling data sets which include some possibly very bad items, then something like the three-parameter model is going to be needed.

This difference of emphasis among the areas of application has its origins in contrasting views of measurement philosophy. As the next paper in this collection makes clear, the Rasch model can be regarded as a special case of the three-parameter model when the discrimination parameters are held equal, and the "guessing" parameter is fixed at zero. Mathematically, this view is undoubtedly correct--but philosophically, it is very misleading. Rasch developed his model, in ignorance of Lord's seminal work on item characteristic curves, on the basis of a set of features which were necessary for an objective measurement system. For measurements with the required properties he found that his model, or a simple mathematical transformation of it, was the mathematically unique solution. The three-parameter model that forms the basis of Lord's Item Response Thoery is not, and cannot be, "specifically objective". Those whose main interest is in understanding existing data sets, and therefore in careful modeling of observed ICCs, see little benefit or relevance in speific objectivity. Those who wish to construct instruments to measure individuals optimally tend to prefer the approach which offers the stronger and more useful system. ICCs which reflect the behavior of inefficient or ineffective items have little interest for them. As has been suggested earlier in this paper, the Rasch model supports a range of applications which goes well beyond what a latent trait model that is not specifically objective can manage.

In the view of this writer, much of the energy which has fueled professional arguments over which is the better model (and the many research studies whose main goal was to compare the effectiveness of the two models in exploring a particular set of data) stem from a failure to appreciate that the two models are basically very different, and were developed to answer different questions. Neither is ever "true". Both are merely models, and it seems clear that in some applications one is of more use than the other and vice versa.

Among users of the Rasch model there is little that is currently controversial, due in no small part to the dominance of two computer programs now in use around the world (BICAL developed by Wright and his associates in Chicago, and PAIR developed by Choppin in London). One current issue that requires clarification concerns the status of "tests of fit". It is generally conceded by Rasch users that whereas better tests of fit are available for the Rasch model than for most other psychometric models, they still leave a lot to be desired. In most cases, showing that an item does not fit the model merely requires collecting a sufficiently large body of data. The area of disagreement lies between those who prefer to treat fit/misfit as a dichotomous categorization and draw up decision rules for dealing with test data on this basis, and those who prefer to regard degree of misfit as a continuous variable which needs to be considered in the context of the whole situation. The present writer belongs in the latter camp, but is prepared to admit that many of the "rules of thumb" that have been developed lack much theoretical or empirical basis.

References

Andrich, D. A rating formulation for ordered response categories. Psychometrika, 1978, 43, 561-73.

Bryce, T.G.K. Rasch-fitting. British Educational Research Journal, 1981, 7, 137-153.

Choppin, B. The national monitoring of academic standards. Paper read to National Council on Measurement in Education, Toronto, 1977.

Choppin, B. Item banking and the monitoring of achievement. Slough, England: National Foundatin for Educational Research, 1978.

Choppin, B. Testing the questions: The Rasch formula and item banking. In M. Raggett (Ed.) Assessment and testing of reading, Ward Lock, London, 1979.

Choppin, B. The use of latent-trait models in the measurement of cognitive abilities and skills. In D. Spearitt (Ed.) The improvement of measurement in education and psychology, Melbourne: ACER, 1982.

Divgi, D.R. A direct procedure for scaling tests with latent trait theory. Paper read at the Annual Meeting of the American Educational Research Association, Los Angeles, 1981.

Douglas, G.A. Conditional inference in a generic Rasch model. In D.Spearitt (Ed.), The improvement of measurement in education and psychology. Melbourne, ACER, 1982.

Engel, I. The differential effect of three different mathematics curricula on student's achievement through the use of sample-free scaling. MA thesis, Tel Aviv University, 1976.

Eysenck, H.J. The structure and measurement of intelligence. Berlin: Springer-Verlag, 1979.

Goldstein, H. Consequences of using the Rasch model for educational assessment. British Educational Research Journal, 1979, 5, 211-220.

Gustafsson, J.E. The Rasch model for dichotomous items. Research Report 63. Institute of Education, University of Goteberg, 1977.

Habermann, S. Maximum likelihood estimates in exponential response models. Annals of Statistics, 1977, 77, 815-841.

Maxwell, J.C. Matter and motion. London, 1876.

Mead, R.J. Analysis of fit to the Rasch model. Doctoral dissertation, University of Chicago, 1975.

Rasch, G. Probabilistic models for some intelligence and attainment tests. Copenhagen: Danmarks Paedagogiske Institut, 1960. (Reprinted by University of Chicago Press, 1980)

Rentz, R.R. & Bashaw, W.L. Equating reading tests with the Rasch model. Athens, Georgia: Educational Resource Laboratory, 1975.

Rentz, R.R. & Bashaw, W.L. The national reference scale for reading: An application of the Rasch model. Journal of Educational Measurement, 1977, 14, 161-180.

Skinner, B.F. A case history in scientific method. The American Psychologist, 1956, 11, 221-233.

Thurstone, L.L. The measurement of opinion. Journal of Abnormal and Social Psychology, 1928, 22, 415-430.

Vincent, D. personal communication, 1980.

Willmott, A. Controlling the examination system. Paper presented at the Schools Council Forum on Comparability of Public Examinations, London, 1979.

Willmott, A. & Fowles, D. The objective interpretation of test performance: The Rasch model applied. Windsor: NFER Publishing Co., Ltd., 1974.

Woodcock, R.W. Woodcock reading mastery tests. Circle Pines, Minnesota: American Guidance Service, 1974.

Wright, B.D. Sample-free test calibration and person measurement. In Proceedings of the 1967 invitational conference on testing problems. Princeton, N.J.: Educational Testing Service, 1968.

Wright, B.D. Solving measurement problems with the Rasch model. Journal of Educational Measurement, 1977, 14, 97-116.

Wright, B.D. Afterword. In G. Rasch Probabilistic models for some intelligence and attainment tests. University of Chicago Press (1980 edition).

Wright, B.D. & Douglas, G.A. Conditional versus unconditional procedures for sample free item analysis. Educational and Psychological Measurement, 1977, 37, 573-586.

Wright, B.D. & Masters, G. Rating Scale Analysis, Chicago: MESA Press, 1982.

Wright, B.D. & Mead, R.J. BICAL: Calibrating items with the Rasch model. Research Memorandum 23, Statistics Lab, Education Department, University of Chicago, 1976.

Wright, B.D. & Stone, M.H. Best Test Design, Chicago: MESA Press, 1979.

THE THREE-PARAMETER LOGISTIC MODEL[1]

Ronald K. Hambleton
University of Massachusetts, Amherst

1. Definition and Background

In a few words, item response theory postulates that (a) examinee test performance can be predicted (or explained) by a set of factors called traits, latent traits, or abilities, and (b) the relationship between examinee item performance and the set of traits assumed to be influencing item performance can be described by a monotonically increasing function called an item characteristic function. This function specifies that examinees with higher scores on the traits have higher expected probabilities for answering the item correctly than examinees with lower scores on the traits. In practice, it is common for users of item response theory to assume that there is one dominant factor or ability which explains performance. In the one-trait or one-dimensional model, the item characteristic function is called an item characteristic curve (ICC) and it provides the probability of examinees answering an item correctly for examinees at different points on the ability scale. In addition, it is common to assume that item characteristic curves are described by one-, two-, or three-parameters. The interpretation of these parameters will be described in section 3. In any successful application of item

[1] Laboratory of Psychometric and Evaluative Research Report No. 126. Amherst, MA: School of Education, University of Massachusetts, 1982.

response theory, parameter estimates are obtained to describe the test items, ability estimates are obtained to describe the performance of the examinees, and there is evidence that the chosen item response model, at least to an adequate degree, fits the test data set (Hambleton, Murray, & Simon, 1982).

Item response theory (or latent trait theory, or item characteristic curve theory as it is sometimes called) has become a very popular topic for research in the measurement field. There have been numerous published research studies, conference presentations, and diverse applications of the theory in the last several years (see for example, Hambleton et al., 1978; Lord, 1980; Weiss, 1980). Interest in item response models stems from two desirable features which are obtained when an item response model fits a test data set: Descriptors of test items (item statistics) are not dependent upon the choice of examinees from the population of examinees for whom the test items are intended, and the expected examinee ability scores do not depend upon the particular choice of items from the total pool of test items to which the item response model has been applied. Invariant item and examinee ability parameters, as they are called, are of immense value to measurement specialists.

Today, item response theory is being used by many of the large test publishers, state departments of education, and industrial and professional organizations, to construct both norm-referenced and criterion-referenced tests, to investigate item bias, to equate tests, and to report test score information. In fact, the various applications have been so successful that discussions of item response theory have shifted from a consideration of their advantages and disadvantages in relation to classical test models to consideration of such matters as model selection, parameter estimation, and the determination of model-data fit. Nevertheless, it would be misleading to convey the impression that issues and technology associated with item response theory are fully developed and without controversy. Still, considerable progress has been made since the seminal papers by Frederic Lord (1952, 1953). It would seem that item response model technology is more than adequate at this time to serve a variety of uses (see, for example, Lord 1980) and there are several computer programs available to carry out item response model analyses (see Hambleton & Cook, 1977).

The purposes of this paper are to address (1) the measurement philosophy underlying item response theory, (2) the assumptions underlying one of the more popular of the item response models, the three-parameter logistic model, (3) the strengths and weaknesses of the three-parameter model, and present gaps in our knowledge of the model, (4) several promising three-parameter model applications, (5) extensions and new applications of the model, and (6) several controversies.

2. Measurement Philosophy

There are many well-documented shortcomings of standard testing and measurement technology.[1] For one, the values of such useful item statistics as item difficulty and item discrimination depend on the

[1] "Standard testing and measurement technology" refers to commonly used methods and techniques for test design and analysis.

particular examinee samples in which they are obtained. The average level of ability and the range of ability scores in an examinee group influences the values of the item statistics, often substantially. This means that the item statistics are only useful when constructing. tests for examinee populations which are very similar to the sample of examinees in which the item statistics were obtained. Another shortcoming of standard testing technology is that comparisons of examinees on an ability measured by a set of test items comprising a test are limited to situations where examinees are administered the sàme (or parallel) test items. But, a próblem is that many achievement and aptitude tests are (typically) suitable for middle-ability students and so the tests do not provide very precise estimates of ability for either high- or low-ability examinees. Increased test score validity without any increase in test length can be obtained if the test difficulty is matched to the approximate ability level of each examinee. But, when several forms of a test which vary substantially in difficulty are used, the task then of comparing examinees becomes more complex because test scores, only, cannot be used. For example, two examinees who perform at a 50% level on two tests which differ substantially in difficulty cannot be considered equivalent in ability, but how different are they in ability? And, how can the ability levels of two examinees be compared when they receive different scores on tests which vary in their difficulty?

Another shortcoming of standard testing technology is that it provides no basis for determining what a particular examinee might do when confronted with a test item. Such information is necessary, for example, if a test designer desires to predict test score characteristics in one or more populations of examinees or to design tests with particular characteristics for certain populations of examinees. In addition to the three shortcomings of standard testing technology mentioned above, standard testing technology has failed to provide satisfactory solutions to many testing problems: For example, the design of tests, identification of biased items, and the equating of test scores. For these and other reasons, psychometricians have been investigating and developing more appropriate theories of mental measurements.

Item response theory purports to overcome the shortcomings of classical or standard measurement theory by providing an ability scale on which examinee abilities are independent of the particular choice of test items from the pool of test items over which the ability scale is defined. Ability estimates obtained from different item samples for an examinee will be the same except for measurement errors. This feature is obtained by incorporating information about the items (i.e., their statistics) into the ability estimation process. Also, item parameters are defined on the same ability scale. They are, in theory, independent of the particular choice of examinee samples drawn from the examinee pool for whom the item pool is intended although errors in item parameter estimation will be group dependent. More will be said about this point later. Again, item parameter invariance across samples of examinees differing in ability is achieved by incorporating information about examinee ability levels into the item parameter estimation process. Finally, by deriving standard errors associated with the ability estimates, another of the criticisms of the classical test model can be overcome.

In summary, the goal of item response theory is to provide both invariant item statistics and ability estimates. These features will be obtained when there is a reasonable fit between the chosen model and the data set. Through the estimation process, items and persons are placed on an ability scale in such a way that there is as close a relationship as possible between the expected examinee probabilities for success on test items obtained from the estimated item and ability parameters and the actual probabilities of performance for examinees positioned at each ability level. Item parameter estimates and examinee ability estimates are revised continually until the maximum agreement possible is obtained between predictions based on the ability and item parameter estimates and the actual test data.

The feature of item parameter invariance can be observed in Figure 1. In the upper part of the figure are three item characteristic curves (ICCs); in the lower part are two distributions of ability. When the chosen model fits the data set the same ICCs are obtained regardless of the distribution of ability in the sample of examinees used to estimate the item parameters. Notice that an ICC provides the probability of examinees at a given ability level answering each item correctly but the probability value does not depend on the number of examinees located at the ability level. The number of examinees at each ability level is different in the two distributions. But, the probability value is the same for examinees in each ability distribution or even in the combined distribution. Of course suitable item parameter estimation will require a heterogeneous distribution of examinees on the ability measured by the test.

It is possible that to some researchers the property of item invariance may seem surprising and unlikely to be obtained in practice, but it is a property which is obtained whenever we study, for example, the linear relationship (as reflected in a regression line) between two variables, X and Y. The hypothesis is made that a straight line can be used to connect the average Y scores conditional on the X scores. When, the hypothesis of a linear relationship is satisfied, the same linear regression line is expected regardless of the distribution of X scores in the sample drawn. Of course proper estimation of the line does require that a suitably heterogeneous group of examinees be chosen. The same situation arises in estimating the parameters for the item characteristic curves which are also regression lines (albeit, non-linear).

3. Assumptions

When fitting an item response model to a test data set, assumptions concerning three aspects of the data set are commonly made (Lord, 1980; Wright & Stone, 1979): These three assumptions will be introduced next.

Dimensionality. It is commonly assumed that only one ability is being measured by a set of items in a test. Of course, this assumption cannot be strictly met because there are always many cognitive, personality, and test-taking factors which impact on test performance, at least to some extent. These factors might include level of motivation, test anxiety, ability to work quickly, knowledge of the correct use of answer sheets, and other cognitive skills in addition to the dominant one measured by the set of test items. What

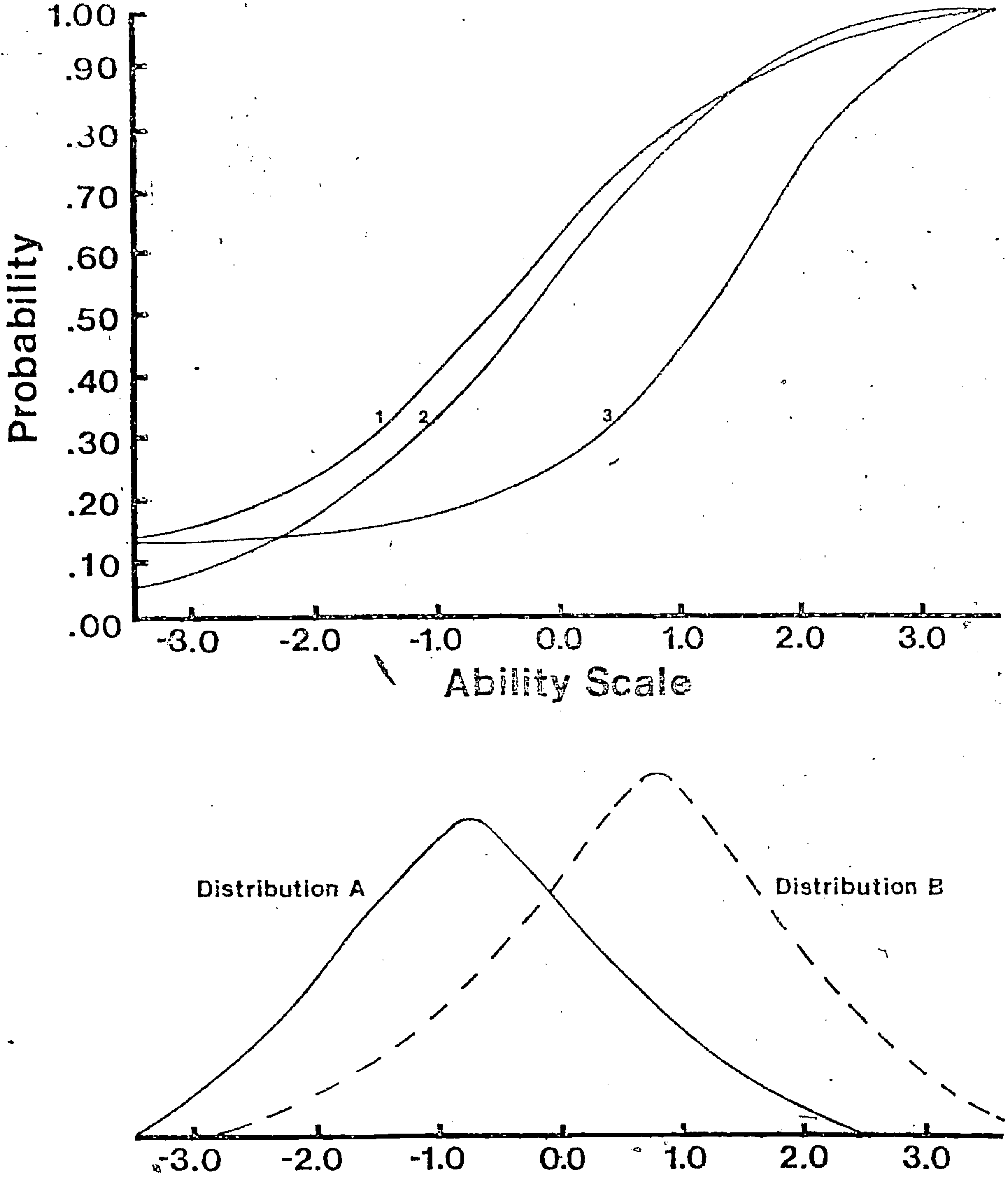


Figure 1. A diagram showing the independence of the shape of item characteristic curves from the underlying ability distribution.

is required for this assumption to be met adequately by a set of test data is a "dominant" component or factor which influences test performance. This dominant component or factor is referred to as the ability measured by the test. This is the ability on which examinees are being measured. All other contributing factors to test performance are defined as errors.

Item response models in which a single ability is presumed sufficient to explain or account for examinee performance are referred to as unidimensional models. Those models in which it is assumed that more than a single ability is necessary to account for examinee test performance are referred to as multi-dimensional models. These latter models are complex, and to date, not well-developed.

Principle of local independence. There is an equivalent assumption to the assumption of unidimensionality known as the assumption of the principle of local independence[1] (Lord & Novick, 1968; Lord, 1980). In words, the assumption requires that the probability of an examinee answering an item correctly (obtained from a one-dimensional model) is not influenced by his/her performance on other items in a test. When an examinee learns information from one test item which helps him or her on other test items the assumption is violated. What the assumption means then is that only the examinee's ability and the characteristics of the test item related to the dominant trait measured by the test influence performance.

Suppose we let u_j be the response of a randomly chosen examinee on items j (j=1, 2, ..., n), and u_j=1, if the examinee answers the

[1] Actually the equivalence only holds when the principle of local independence is defined in the one-dimensional case.

item correctly, and $u_j=0$, if the examinee answers the item incorrectly. Suppose also we let the symbols, P_j, and Q_j $(Q_j=1-P_j)$ denote the probability of the examinee answering the item correctly and incorrectly, respectively. The assumption of the principle of local independence in mathematical terms can then be stated in the following way:

$$\text{Prob } (U_1 = u_1, U_2 = u_2, \ldots, U_n = u_n)$$

$$= P_1^{u_1} Q_1^{1-u_1} P_2^{u_2} Q_2^{1-u_2} \ldots P_n^{u_n} \; Q_n^{1-u}$$

$$= \prod_{j=1}^{n} P_j^{u_j} Q_j^{1-u_j}$$

In words, the assumption of local independence in the one dimensional case requires that the probability of any response pattern occurring for an examinee is given by the product of probabilities associated with his/her successes and/or failures on the test items. The probabilities are obtained from a one-dimensional model.

Mathematical form of the item characteristic curves. An item characteristic curve is a mathematical function that relates the probability of success on an item to the ability measured by the set of items contained in the test. There is no concept comparable to the notion of an item characteristic curve in standard test technology. A primary distinction among different item response models is in the mathematical form of the corresponding item characteristic curves. It is up to the user to choose one of the many mathematical forms for the shape of the item characteristic curves. In doing so, an assumption about the items is being made which can be verified later by how well the chosen model "explains" the observed test results.

Each item characteristic curve for a particular item response model is a member of a family of curves of the same general form. The number of parameters required to describe the item characteristic curves in the family will depend on the particular item response model. With the three-parameter logistic model, statistics which correspond approximately to the notions of item difficulty and discrimination (used in standard testing technology), and the probability of low-ability examinees answering an item correctly, are used. The mathematical expression for the three-parameter logistic curve is:

$$(1) \qquad P_g(\theta) = c_g + (1-c_g)\frac{e^{Da_g(\theta - b_g)}}{1+e^{Da_g(\theta - b_g)}}, \quad g=1, 2, \ldots, n,$$

where:

$P_g(\theta)$ = the probability that an examinee with ability level θ answers item g correctly,

b_g = the item g difficulty parameter,

a_g = the item g discrimination parameter,

c_g = the lower asymptote of an ICC representing the probability of success on item g for low-ability examinees,

D = 1.7 (a scaling factor),

and

n = the number of items in the test.

The parameter c_g is the lower asymptote of the item characteristic curve and represents the probability of examinees with low ability correctly answering an item. The parameter c_g is included in the model to account for test response data at the low end of the ability continuum, where among other things, guessing is a factor in test performance. It is now common to refer to the parameter c_g as the pseudo-chance level parameter in the model.

Typically, c_g assumes values that are smaller than the value that would result if examinees of low ability were to guess randomly to the item. As Lord (1974) has noted, this phenomenon can probably be attributed to the ingenuity of item writers in developing "attractive" but incorrect choices. For this reason, c_g is no longer called the "guessing parameter". To obtain the two-parameter logistic model from the three-parameter logistic model, it must be assumed that the pseudo-chance level parameters have zero-values. This assumption is most plausible with free response items but it can often be approximately met when a test is not too difficult for the examinees. For example, this assumption may be met when competency tests are administered to students following effective instruction. Perhaps the most popular of the present item response models is the one-parameter logistic model (or commonly named as the "Rasch Model" after Georg Rasch the discoverer of the model). It can be obtained from the three-parameter logistic model by assuming that all items have pseudo-chance level parameters equal to zero and by assuming all items in the test are equally discriminating. Also, the one-parameter model, or Rasch model as it is commonly referred to, can be produced from a different set of measurement principles and assumptions. Readers are referred to Choppin (in this volume) for an alternate development of the Rasch model. The viability of these assumptions is discussed by Hambleton et al. (1978).

Item characteristic curves for the latent linear model[1] and the one-, two-, and three-parameter logistic models are shown in Figure 2. Readers are referred to Hambleton (1979), Lord (1980), and Wright and Stone (1979) for additional information about logistic test models.

4. Strengths, Weaknesses, and Gaps

The exploration of item response models and their application to eductional testing and measurement problems has been under study for about fifteen years now. Certainly there are many problems requiring resolution but enough is known about item response models to use them successfully in solving many testing problems (see Lord, 1980; Hambleton, 1983). Item response models, when they provide an accurate fit to a data set, and in theory, the three-parameter logistic model will fit a data set more accurately than a logistic model with fewer item parameters, can produce invariant item and ability parameters described earlier. Some of these promising applications will be described in the next two sections (also see, Hambleton, 1983).

On the negative side, the three-parameter model is based upon several strong assumptions. (Of course, the one- and two-parameter logistic models are based on even stronger assumptions.) When these assumptions are not met, at least to an approximate degree, desirable

[1] The item characteristic curves for the latent linear model are of the form:

$$P_g(\theta) = b_g + a_g \theta \cdot$$

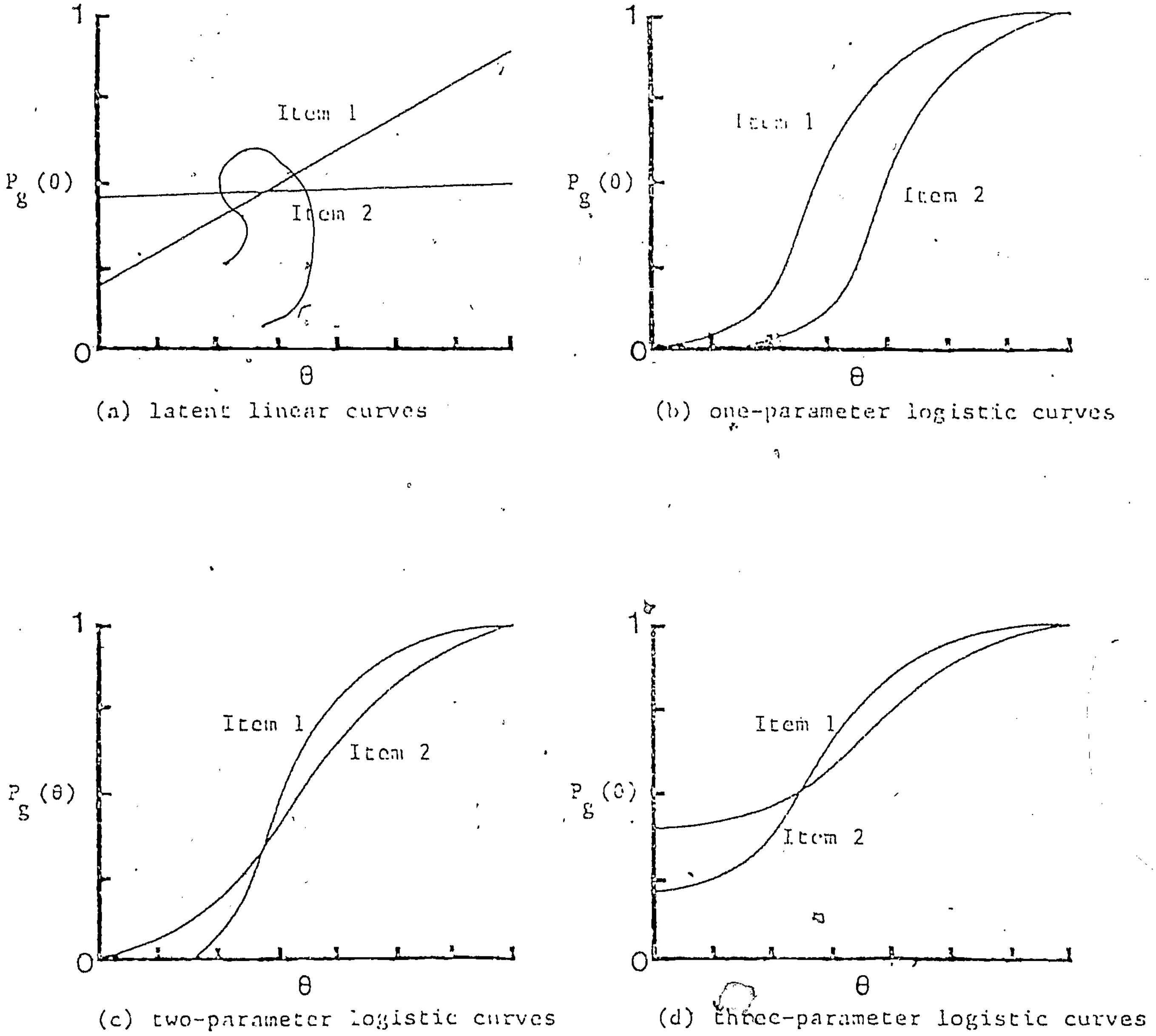


Figure 2. Examples of item characteristic curves.

features expected from applying the three-parameter model will not be obtained. Other weaknesses, presently, of the three-parameter model are (1) the need of rather large numbers of items and examinees for proper item parameter estimation, (2) the relatively high computer costs for obtaining item and ability parameter estimates, and (3) the difficulties inherent in interpreting a complex model to test practitioners.

On the first point, Lord (1980) suggested examinee sample sizes in excess of 2,000 are needed. Perhaps Lord is overly conservative in his figure but it does appear that sample sizes in excess of 600 or 700 are needed, and a disproportionate number of examinees near the lower end of the ability scale so that the c parameters can be estimated properly. Because of the required minimum sample sizes, small scale measurement problems (e.g., teacher-made tests) cannot properly be addressed with the three-parameter model. With respect to the second point, it is common to report high costs associated with using LOGIST although there is evidence that the LOGIST program will run substantially faster and cheaper on some computers. Hutten (1981) reported an average cost of $69 to run 25 data sets with 1,000 examinees and 40 test items on a CYBER 175 ($800/hour for CPU time). Finally, the untrained test developer will have difficulty working with three statistics per item but as CTB/McGraw-Hill has shown in building the latest version of the California Tests of Basic Skills, test editors can be trained to successfully use the additional information provided by the three-parameter model (Yen, 1983).

There is (at least) one practical shortcoming of the three-parameter model and its applications: There does seem to be a shortage of available computer programs to carry out a three-parameter logistic model analysis. The most readily available program is LOGIST, described by Wingersky (1983) and Wingersky, Barton, and Lord (1982). The most readily available version of this program runs on IBM equipment although there is evidence that the program may run substantially faster on other computers. Additional investigation of this finding is needed along with on-going studies to try and speed up the convergence of estimates. In addition, there may be other ways to improve the estimation process. Swaminathan and Gifford (1981) have obtained very promising results with Bayesian item and ability parameter estimates. Their results compare favorably with results from LOGIST and they can be obtained considerably faster and more cheaply than the same estimates obtained with LOGIST.

There are (at least) three areas in which we lack full understanding of item response models. First, additional robustness studies with the one- and two-parameter logistic models are needed and with respect to a number of promising applications. What is the practical utility of the three-parameter model in comparison to the one- and two-parameter models? Second, appropriate methods for testing model assumptions and determining the goodness of fit between a model and a data set are needed. Hambleton and his colleagues (Hambleton, 1980; Hambleton, Murray, & Simon, 1982) have made a promising start by organizing many of the present methods and developing several new ones. Much of their work involves the use of graphs, replications, residual analyses and cross validation procedures. More work along the same general lines would seem desirable. Third, there is a great need for persons to gain experiences with the three-parameter model and to share their new found knowledge and experiences with others.

5. Applications[1]

In this section, several promising applications of the three-parameter logistic model will be described briefly: Item banking, test development, criterion-referenced testing, item bias, and adaptive testing. Other applications of the three-parameter model are discussed by Hambleton et al. (1978), Lord (1980), and Hambleton (1983).

Item banking. The development of criterion-referenced testing technology has resulted in increased interest in item banking (Choppin, 1976). An item bank is a collection of test items, "stored" with known item characteristics. Depending on the intended purpose of the test, items with desired characteristics can be drawn from the bank and used to construct a test with known properties. Although classical item statistics (item difficulty and discrimination) have been employed for this purpose, they are of limited value for describing the items in a bank because these statistics are dependent on the particular group used in the item calibration process. Latent trait item parameters, however, do not have this limitation, and consequently are of much greater use in describing test items in an item bank (Choppin, 1976). The invariance property of the latent trait item parameters makes it possible to obtain item statistics that are comparable across dissimilar groups. Since the item parameters depend on the ability scale, it is not possible to directly compare

[1] Some of the material in this section is taken from and/or edited from a paper by Hambleton et al. (1978).

latent trait item parameters derived from differnt groups of examinees until the ability scales are equated in some way. Fortunately, the problem is not too hard to resolve since Lord and Novick (1968) have shown that the item parameters in the two groups are linearly related. Thus, if a subset of calibrated items is administered to both groups, the linear relationship between the estimates of the item parameters can be obtained by forming two separate bivariate plots, one establishing the relationship between the estimates of the item discrimination parameters for the two groups, and the second, the relationship between the estimates of the item difficulty parameters. Having established the linear relationship between item parameters common to the two groups, a prediction equation can then be used to predict item parameters for those items not administered to the first group. In this way, all item parameters can be equated to a common group of examinees and corresponding ability scale. One large test publishing company, the California Test Bureau/McGraw-Hill, presently customizes tests for school districts wih items calibrated using the three-parameter logistic model.

Test development. The three-parameter model is presently being used by a number of organizations in test development (e.g., CTB/McGraw-Hill, ETS). The three-parameter model provides the test developer with not only sample invariant item parameters but also with a powerful method of item selection (Birnbaum, 1968). This method involves the use of information curves, i.e., items are selected depending upon the amount of information they contribute to the total amount of information supplied by the test (Lord, 1980)[1]. One of the

[1] Readers are referred to Hambleton (1979) for an introduction to item and test information and efficiency curves.

useful features of item information curves is that the contribution of each item to the test information function can be determined without knowledge of the other items in the test. When standard testing technology is applied the situation is very different. The contribution of any item to such statistics as test reliability cannot be determined independently of the characteristics of all the other items in the test.

Lord (1977) outlined a procedure for use of item information curves to build a test to meet any desired set of specifications. The procedure employs a pool of calibrated items, with accompanying information curves, such as might be obtained from the item banking methods described earlier. The procedure outlined by Lord consists of the following steps:

1. Decide on the shape of the desired test information curve. Lord (1977) calls this the target information curve.
2. Select items with item information curves that will fill up the hard-to-fill areas under the target information curve.
3. After each item is added to the test, calculate the test information curve for the selected test items.
4. Continue selecting test items until the test information curve approximates the target information curve to a satisfactory degree.

An example of the application of this technique to the development of tests for differing ranges of ability (based on simulated data) is given by Hambleton (1979).

Criterion-referenced testing. A principal use of a criterion-referenced test is to estimate an examinee's level of mastery (or "ability") on an objective. Thus, a straightforward application of the three-parameter model would produce examinee ability scores. Among the advantages of th application would be that items could be sampled (for example, at random) from an item pool for each examinee, and all examinee ability estimates would be on a common scale. A potential problem with this application, however, concerns the estimation of ability with relatively short tests.

Since item parameters are invariant across groups of examinees, it would be possible to construct criterion-referenced tests to "discriminate" at different levels of the ability continuum. Then, a test developer might select an "easier" set of test items for a pretest than a posttest, and still be able to measure "examinee growth" by estimating examinee ability with the three-paramete model at each test occasion on the same ability scale. This cannot be done with classical approaches to test development and test score interpretation. If we had a good idea of the likely range of ability scores for the examinees, test items could be selected so as to maximize the test information in the region of ability for the examinees being tested. The optimum selection of test items would contribute substantially to the precision with which ability scores were estimated. In the case of criterion-referenced tests, it is common to observe substantially lower test performance on a pretest than on a posttest; therefore, the test constructor could select the easier test items from the domain of items measuring an objective for the pretest and more difficult items could be selected for the posttest. This would enable the test constructor to maximize the precision of measurement of each test in the region of ability where the examinees would most likely be located. Of course, if the assumption about the location of ability scores was not accurate, gains in precision of measurement would not be obtained.

The results reported in Tables 1 and 2 (from Hambleton, 1979) show clearly the advantages of "tailoring" a test to the ability level of a group. Of course, the potential improvements depend on the validity of a test developer's assumption about the examinee ability distribution. If he or she uses an incorrect prior distribution as a basis for designing a test, the resulting test will certainly not have the desired characteristics.

Item bias. Identifying biased items in a test usually involves comparing the performance of the subgroups of interest (e.g., Blacks, Hispanics, and Whites) on the test items. The problem that arises is that differences among the subgroups due to bias is confounded with any true differences in abilities among the subgroups. Needed is an item bias detection method that can control for true ability differences. Via a three-parameter model analysis, it is possible to compare corresponding item characteristic curves. At each ability level, independent of the proportion of examinees in each subgroup who are located at the ability level, the expected proportion of successes in each subgroup,obtained from the ICCs, can be compared. The ICCs estimated in each group, in theory, do not depend upon the underlying ability distributions. Any differences in the curves, beyond the

Table 1

Test Information Curves and Efficiency for Three Criterion-Referenced Test Designs From a Domain of Items of Equal Discrimination and Pseudo-chance Levels Equal to .20

Ability Level	Test Information Curves			Efficiency (Relative to the "Wide Range Form")		Change in Effective Test Length	
	"Wide Range Form"	"Easy Form"	"Difficult Form"	"Easy Form"	"Difficult Form"	"Easy Form"	"Difficult Form"
-3.0	.22	.36	.07	1.63	.31	63%	-69%
-2.0	.86	1.31	.36	1.53	.42	53%	-58%
-1.0	2.08	2.81	1.31	1.35	.63	35%	-37%
0.0	3.04	3.29	2.81	1.08	.92	8%	-8%
1.0	2.76	2.28	3.29	.82	1.19	-18%	19%
2.0	1.69	1.12	2.28	.66	1.35	-34%	35%
3.0	.79	.46	1.12	.59	1.42	-41%	42%

Table 2

Test Information Curves and Efficiency for Three Criterion-Referenced Test Designs From a Domain of Items with Varying Discrimination Indices and Pseudo-chance Levels Equal to .20

Ability Level	Test Information Curves			Efficiency (Relative to the "Wide Range Form")		Change in Effective Test Length	
	"Wide Range Form"	"Easy Form"	"Difficult Form"	"Easy Form"	"Difficult Form"	"Easy Form"	"Difficult Form"
-3.0	.24	.37	.08	1.58	.35	58%	-65%
-2.0	.86	1.27	.37	1.48	.44	48%	-56%
-1.0	2.02	2.71	1.27	1.35	.63	35%	-37%
0.0	2.94	3.18	2.71	1.08	.92	8%	-8%
1.0	2.65	2.16	3.18	.81	1.20	-19%	20%
2.0	1.59	1.06	2.16	.67	1.36	-33%	36%
3.0	.75	.46	1.06	.61	1.41	-39%	41%

usual sampling errors, can be attributed to differential subgroup responses to the items, i.e., bias. It is becoming routine practice for several large test publishers to investigate bias in test items with the aid of the three-parameter logistic model. Since the three-parameter model often provides a somewhat better fit to test data at the lower end of the ability continuum (Hambleton et al., 1982) than less general logistic models, the three-parameter model may be more useful than other logistic models for studying bias.

Adaptive testing. Possibly the first and most well-developed application of the three-parameter logistic model to date is adaptive testing (Lord, 1980; Weiss, 1980). In adaptive testing each examinee is administered a set of test items "tailored" or "adapted" to his/her ability level. Clearly, total test scores cannot provide an adequate basis upon which to compare examinees. Some examinees will be administered sets of test items which are substantially more difficult (or easier) than the test items administered to other examinees. By calibrating test items using the three-parameter logistic model in advance of the actual testing, and using the three-parameter model to estimate examinee ability levels, examinees can be compared even though the test items administered to different examinees may differ substantially in difficulty. Because of the ready availability of the computer, scoring difficulties associated with the use of the three-parameter model can be overcome easily.

The U.S. military is firmly committed to the use of adaptive testing with the three-parameter model in many of its testing programs. Presently a feasibility study is being conducted along with the preparation of plans for adaptive testing implementation and evaluation of the total adaptive testing system.

6. Possible Extensions/New Applications

Numerous researchers are presently addressing the development of new item response models. For example, Samejima (1979) is exploring the development of multidimensional models in which item options are ranked based on their relationship to ability, and characteristic curves are produced for each option. McDonald (1982) has provided a general formulation for generating a wide range of multidimensional linear and non-linear polychotomous item response models. Bock, Mislevy, and Woodson (1982) have described a two-parameter item response model which can handle continuous data and where the unit of analysis can be a group (e.g., the classroom or a school). This model will be especially useful in program evaluation investigations. A minor variation of the three-parameter model which appears to have some utility is a model in which a common value of the c parameter is used for all test items (Wingersky, 1983). This revised three-parameter model will receive some use in the coming years. A four-parameter logistic model has also been suggested (the fourth parameter is the upper asymptote) but it appears to have very limited practical usefulness. All of these new models can be viewed as modifications/extensions of the three-parameter logistic model and they will undoubtedly receive study from researchers in the coming years.

Because of the newness of the IRT area, all applications of the three-parameter model might legitimately be classified as new. For the purposes of this paper, "new applications" will be those which to date have not been published. Two new applications, then, of the three-parameter model to the problems of item selection (Hambleton & de Gruijter, 1983) and score prediction (Hambleton & Martois, 1983) will be described briefly next.

Item selection. Item response models appear useful to the problem of item selection because they lead to item statistics which are referenced to the same scale on which examinee abilities are defined. In addition, it should be noted that IRT provides a procedure for placing a cut-off score which is normally set on a proportion-correct scale defined over a domain of items on the same scale as the test items and the examinees (Lord, 1980). Therefore, the usefulness of a test item for measurement at any point on the ability scale can be assessed.

Hambleton and de Gruijter (1983) described a nine step procedure for selecting test items using three-parameter model item statistics, and via a computer simulation study showed the advantages, at least in the absence of errors associated with item parameter estimates, of item selection with the aid of IRT over a standard item selection procedure.

Test score predictions. The concept of item banking has attracted considerable interest in recent years from school districts, state departments of education, and test publishing companies. When item banks consist of test items which are technically sound and validly measure the objectives or competencies to which they are referenced, the task of producing high quality tests is made considerably easier. Item banks are most often used to construct criterion-referenced tests (CRTs) or mastery tests or competency tests, as they are sometimes called. What is not commonly available for use with these CRTs are derived scores such as percentiles. Derived scores are not always valued but on occasion they are required by school districts who receive federal funds (e.g., Title I) for they must evaluate their funded programs with national norms (e.g., percentile scores).

In theory, the problem faced by school districts who require information fur (1) diagnosing and monitoring student performance in relation to competencies and (2) normative scores for the comparison of examinees is easy to solve. Teachers can use their item banks to build classroom tests on an "as-needed" basis, and when the need arises, they can administer any necessary commercially available standardized norm-referenced tests. But this solution has problems: (1) the amount of testing time for students is increased, and (2) the financial costs of school testing programs is increased. On the other hand, when testing time is held constant, and norm-referenced tests are administered, there is less time available for instructionally relevant testing (i.e., CRTs). A more satisfactory solution would allow teachers to administer test items measuring objectives of interest in their instructional programs, and at the same time, allow for normative scores to be estimated from the test items which are administered. An often used solution of selecting a norm-referenced test to provide normative scores and criterion-referenced information through the interpretation of examinee performance on an item by item basis is not very suitable criterion-referenced measurement and will not insure that all competencies of interest are measured in the test.

Hambleton (1980) suggested a possible item response model solution to the problem of providing both instructonal information and normative information from a single test. A latent ability scale to which a large pool of test items are referenced can be very useful in obtaining normative scores from tests constructed by drawing items from the pool. A norms table can be prepared from the administration of a sample of items in the pool. Then the norms table can be used successfully with any tests which are constructed by drawing items from the pool. Local norms can be prepared by districts who build their own item banks. A test publishing company probably would prepare national norms for selected tests constructed from their item banks.

Hambleton and Martois (1983) recently finished a study in which it was found that both the one- and the three-parameter logistic models resulted in excellent predictions of how examinees performed on a norm-referenced test. Predictions were made from tests with items that were easier, comparable to, or harder than items in the normed test. Similar results were obtained in three subject areas at two grade levels. Further research along the same general lines seems highly desirable because of the importance of the problem area.

7. Controversies

Perhaps like any emerging area, item response theory has generated considerable controversy and strong emotional feelings in support of one model versus another. Much of the debate has centered on the choice between the one- and three-parameter logistic models. There has also been some controversy surrounding the utility of

Bayesian estimators (Samejima vesus Novick and Swaminathan) and the appropriateness of item response models for the analysis of aptitude versus achievement tests. On this latter point there is some feeling that items on achievement tests are instructionally sensitive and therefore item response model statistics will not be invariant in pre- and post-instructional groups.

With respect to the choice of the one- versus the three-parameter logistic model, a number of questions have arisen:

1. What is the effect of boundary constraints placed on item and ability parameter estimates obtained with LOGIST?

2. What is the practical utility of the three-parameter model? In most practical settings, won't the two models produce highly similar results?

3. What is the additional cost of running a three-parameter model analysis and is the practical utility of the gains that accrue worth the financial costs and the added complexity which results?

4. Since examinees can guess the answers to multiple-choice test items, the three-parameter model should be selected on the basis of this a priori consideration (Traub, 1983).

5. How well do the item response models fit any data sets? This point is in dispute because many of the present goodness of fit statistics have been found to be inappropriate (e.g., see papers by Wollenberg, 1980; Divgi, 1981).

These and other questions will undoubtedly be addressed in the coming years. Answers will contribute to our knowledge of the three-parameter logistic model and the situations in which the model should be used.

REFERENCES

Birnbaum, A. Some latent trait models and their use in inferring an examinee's ability. In F.M. Lord & M.R. Novick, Statistical theories of mental test scores. Reading, MA: Addison-Wesley, 1968.

Bock, R.D., Mislevy, R. & Woodson, C. The next stage in eductional assessment. Educational Researcher, 1982, 16, 4-11.

Choppin, B.H. Recent developments in item banking: A review. In D.N.M. deGruijter & L.J.Th. van der Kamp (Eds.), Advances in psychological and educational measurement. New York: Wiley, 1976.

Divgi, D.R. Does the Rasch model really work? Not if you look closely. Paper presented at the Annual Meeting of NCME, Los Angeles, 1981.

Hambleton, R.K. Latent trait models and their applications. In R. Traub (Ed.), Methodological developments: New directions for testing and measurement (No. 4). San Francisco: Jossey-Bass, 1979.

Hambleton, R.K. Latent ability scales, interpretations, and uses. In S. Mayo (Ed.), New directions for testing and measurement: Interpreting test performance. San Francisco: Jossey-Bass, 1980.

Hambleton, R.K. Applications of item response theory. Vancouver, BC: Educational Research Institute of British Columbia, 1983.

Hambleton, R.K., & Cook, L.L. Latent trait models and their use in the analysis of educational test data. Journal of Educatinal Measurement, 1977, 14, 75-96.

Hambleton, R.K., & de Gruijter, D.N.M. Application of item response models to criterion-referenced test item selection. Journal of Educational Measurement, 1983, 20, in press.

Hambleton, R.K., & Martois, J. Evaluation of a test score prediction system based upon item response model principles and procedures. In R.K. Hambleton (Ed.), Applications of item response theory. Vancouver, BC: Educational Research Institute of British Columbia, 1983.

Hambleton, R.K., Murray, L., & Simon, R. Applications of item response models to NAEP mathematics exercise results. Final Report. Submitted to the Educational Commission of the States, 1982.

Hambleton, R.K., Swaminathan, H., Cook, L.L., Eignor, D.R., & Gifford, J.A. Developments in latent trait theory: Models, technical issues, and applications. Review of Educational Research, 1978, 48, 467-510.

Hutten, L. Fitting the one- and three-parameter models to a variety of tests. Laboratory of Psychometric and Evaluative Research Report No. 116. Amherst, MA: School of Education, University of Massachusetts, 1981.

Lord, F.M. A theory of test scores. Psychometric Monograph No. 7, 1952.

Lord, F.M. An application of confidence intervals and of maximum likelihood to the estimation of an examinee's ability. Psychometrika, 1953, 18, 57-75.

Lord, F.M. Estimation of latent ability and item parameters when there are omitted responses. Psychometrika, 1974, 39, 247-264.

Lord, F.M. Practical applications of item characteristic curve theory. Journal of Educational Measurement, 1977, 14, 117-138.

Lord, F.M. Applications of item response theory to practical testing problems. Hillsdale, NJ: Erlbaum, 1980.

Lord, F.M., & Novick, M.R. Statistical theories of mental test scores. Reading, MA: Addison-Wesley, 1968.

McDonald R.P. Linear versus non-linear models in item response theory. Applied Psychological Measurement, 1982, 6, in press.

Samejima, F. A new family of models for the multiple-choice item. Office of Naval Research, Research Report 79-4, 1979.

Swaminathan, H., & Gifford, J.A. Bayesian estimation in the three-parameter logistic model. Laboratory of Psychometric and Evaluative Research Report No. 119. Amherst, MA: School of Education, University of Massachusetts, Amherst, 1981.

Traub, R. E. A priori considerations in choosing an item response model. In R. K. Hambleton (Ed.), Applications of Item Response Theory. Vancouver, BC: Educational Reserach Institute of British Columbia, 1983.

Weiss, D. (Ed.), Proceedings of the 1979 Computerized Adaptive Testing Conference. Minneapolis, MN: University of Minnesota, 1980.

Wingersky, M.S. LOGIST: A program for computing maximum likelihood procedures for logistic test models. In R.K. Hambleton (Ed.), Applications of item response theory. Vancouver, BC: Educational Research Institute of British Columbia, 1983.

7

Wingersky, M.S., Barton, M.A., & Lord, F.M. LOGIST user's guide. Princeton, NJ: Educational Testing Service, 1982.

Wollenberg, A.L. van den. On the Wright-Panchapakesen goodness of fit test for the Rasch model. Internal Report 80-MA-02. Nijmegen, The Netherlands: Katholieke Universiteit, Vakgroep Mathematische Psychologie, Psychologisch Laboratorium, 1980.

Wright, B.D., & Stone, M.H. Best test design. Chicago: MESA Press, 1979.

Yen, W. Use of the three-parameter model in constructing a standardized achievement test. In R.K. Hambleton (Ed.), Applications of item response theory. Vancouver, BC: Educational Research Institute of British Columbia, 1983.

MEASURING ACHIEVEMENT WITH LATENT STRUCTURE MODELS

Rand R. Wilcox
Center for the Study of Evaluation
University of California, Los Angeles

1. MEASUREMENT PHILOSOPHY

The basic assumption in latent class models designed to measure achievement is that an examinee can be described as knowing or not knowing the answer to a test item, and that inferences about an examinee's ability level should take this notion into account. The goals of an n-item test might be to determine how many of the items an examinee knows, which items are known or which are not known, or what proportion of items among a domain of items are known. The problem is that examinees might give the correct response when they do not know, or they might carelessly give the wrong response when they know. Latent class models are an attempt to measure and correct the effects of these errors when addressing a particular measurement problem. Even if some other model is ultimately preferred, such as a latent trait model, latent class models are potentially useful.

Currently it appears that correcting for guessing is more important than might have been expected. Moreover, assuming random guessing seems to be an unsatisfactory solution. Consider, for example, the problem of determining the length of a criterion-referenced test where the goal is to determine whether an examinee's percent correct true score or domain score, p, is above or below some known constant p_0. If p_0=.8 and n=29 items are used, the probability of correctly determining whether $p \geq p_0$ is at least .9 when $p \geq .9$ or $p \leq .7$, and when the binomial error model is assumed. If random guessing is assumed, nearly 200 items are needed (van den Brink and Koele, 1980), and if one allows for the possibility that guessing is not at random, over 2,600 items are required to attain the same level of accuracy (Wilcox, 1980). In some cases guessing might be nearly random, but there is empirical evidence that this is generally not the case (Coombs et al., 1956; Bliss, 1980; Cross and Frary, 1977; Wilcox, 1982a, 1982b).

Another way of describing the measurement philosophy of latent class models is that an examinee's test score is a function, in part, of the distractors that are used, and that it is important to take this effect into account. In the past this problem was ignored, probably because there were no reasonable ways of dealing with it, and because it was not clear just how serious this problem was. Now, however, there are several ways of measuring and correcting the effects of distractors. It might appear that some latent trait models deal with guessing, but in fact latent trait models ignore the errors that are of concern here. Thus, these errors might have a serious effect on how latent trait models are used and interpreted. Wainer and Wright (1980) as well as Mislevy and Bock (1982) examined certain aspects of how guessing affects latent trait models, but the type of guessing examined here is different.

2. THE MODELS AND THEIR ASSUMPTIONS

Generally latent class models are based on assumptions about how examinees behave when responding to an item, or how items are related to one another, or the manner in which tests are administered. While a general description of latent class models is possible, such a description is not given here. Instead attention is focused on those models that seem to have the most practical value.

A Latent Structure Model for Answer-Until-Correct Tests

This section assumes that an examinee responds to a multiple-choice test item according to an answer-until-correct (AUC) scoring procedure. This means that if an examinee chooses an incorrect response, another response is chosen, and this process continues until the correct response is identified.

AUC tests are easily administered in the classroom using especially designed answer sheets where the examinee erases a shield corresponding to a particular alternative. (These answer sheets are available commercially, for example, through Van Valkenburg, Nooger and Neville in New York, N.Y., and they are relatively inexpensive.) If the letter under the shield indicates an incorrect response, the examinee erases another shield, and this continues until tne correct shield is erased.

Consider a population of examinees, and let ζ_i be the proportion of the examinees who can eliminate i distractors from consideration. That is, because of partial information, some of the examinees will rule out some of the distractors without knowing the correct response. If there are t alternatives from which to choose, and if the examinee can eliminate t-1 distractors from consideration, the examinee is said to know the correct response. Thus, ζ_{t-1} is the probability that a randomly sampled examinee knows the correct response. Note that no distinction is made between examinees who can eliminate all the distractors via partial information and those that know. In other words, an examinee might choose the correct response, not because the correct answer is known, but because the test constructor was unable to produce at least one effective distractor. Thus, it is assumed that at least one effective distractors is being used, and presumably this problem can be minimized by choosing t to be reasonably large. Of course the crucial step is finding someone who can write effective distractors.

As alluded to earlier, it is assumed that among the examinees who do not know, some might be able to eliminate one or more distractors from consideration via partial information. It is further assumed that once these distractors are eliminated, the examinee guesses at random among the alternatives that remain. Hence, if p_i is the probability of a correct response on the $i^{\underline{th}}$ attempt of the item $(i=1,\ldots,t)$,

$$p_i = \sum_{j=0}^{t-i} \zeta_j/(t-j) \qquad (2.0)$$

For example, if t=3

$$p_1 = \zeta_0/3 + \zeta_1/2 + \zeta_2$$

$$p_2 = \zeta_0/3 + \zeta_1/2,$$

and

$$p_3 = \zeta_0/3 .$$

In general, the proportion of examinees who know the correct response is

$$\zeta_{t-1} = p_1 - p_2. \qquad (2.1)$$

The model implies that

$$p_1 \geq p_2 \geq \cdots \geq p_t, \qquad (2.2)$$

and this can be tested by applying results in Robertson (1978). Empirical investigations (Wilcox, 1982a, 1982b) suggest that (2.2) will usually hold.

The next section describes how one might proceed when (2.2) appears to be unreasonable.

For N randomly sampled examinees, let x_i be the number who get the correct response on the $i^{\underline{th}}$ attempt. Then the x_i's have a multinomial distribution give by $\binom{N}{x} p_1^{x_1} \ldots p_t^{x_t}$ where $\binom{N}{x} = N!/(x_1! \ldots x_t!)$, $\sum x_i = N$, $0 \leq p_i \leq 1$, and $\sum p_i = 1$. An unbiased maximum likelihood estimate of p_i is just x_i/N, and so

$$\hat{\zeta}_{t-1} = (x_1 - x_2)/N \tag{2.3}$$

is a maximum likelihood estimate of ζ_{t-1}, the proportion of examinees who know the correct response. Semantically, if we compute the proportion of examinees who get the item correct on the first attempt, and then subtract the proportion who get it right on the second attempt, we have an estimate of the probability that the typical examinee will know the answer.

Note that $\hat{\zeta}_{t-1}$ given by (2.3) can be negative, but ζ_{t-1} is positive when the model is assumed to be true. This can be corrected by simply estimating ζ_{t-1} to be zero when $\hat{\zeta}_{t-1} < 0$. From Barlow et al. (1972), a maximum likelihood estimate of ζ_{t-1} under the assumption that (2.2) holds can be had by applying the pool-adjacent-violators algorithm.

A Misinformation Model

The previous section assumed that the inequality in equation (2.2) is true, but experience indicates that occasionally this will not be the case. In this event a misinformation model may be appropriate. Of course for some items an investigator might suspect a misinformation model is needed before any test data is collected in which case the results in this section might be applied without testing (2.2).

As will soon become evident, there is no specific misinformation model, but rather a class of models that might be used. The choice from among these models will depend on what seems to be a reasonable assumption about how examinees behave. At the moment there are no empirical procedures to aid a test constructor when choosing from among the various misinformation models. So far, however, this does not seem to be a serious problem.

To better understand how to apply these models, consider the following test item.

> When a block of iron is heated until it is red hot, it gets bigger. If the iron weighs 20 lbs. at room temperature, how much will it weigh when red hot?
>
> 1) 19.8 lbs.   2) 20 lbs.   3) 20.1 lbs.   4) 20.5 lbs.
> 5) 20.61 lbs.

This item is similar to one investigated in Wilcox (1982b) where the examinees were approximately 14 years old. The point is that it seems reasonable to suspect that some examinees will choose from among the last three alternatives because they believe the iron weighs more when it expands. The goal then is to devise a model that takes this behavior into account.

In this section it is assumed that the examinees belong to one of three mutually exclusive groups: 1) they know the item, 2) they have misinformation, 3) or they do not know, do not have misinformation, and guess at random. For examinees with misinformation, it is also assumed that they will choose c specific incorrect alternatives before choosing the correct response. At the moment there is no empirical method for choosing c; this must be done based on what seems reasonable for the item being used. For example, in the item described above, c=3 would be considered. In some cases the resulting latent structure model can be checked with a goodness-of-fit test, but as will be seen this is not always the case.

For the population of examinees being tested, let ζ be the proportion of examinees who know, ν_1 be the proportion who do not know, do not have misinformation and guess at random, and let ν_2 be the proportion who have misinformation. If an AUC scoring procedure is used, and if p_i is defined as before, then for c=3 and t=5

$$p_1 = \zeta + \nu_1/5 \tag{2.4}$$

$$p_2 = \nu_1/5 \tag{2.5}$$

$$p_3 = \nu_1/5 \tag{2.6}$$

$$p_4 = \nu_2 + \nu_1/5 \tag{2.7}$$

$$p_5 = \nu_1/5 \tag{2.8}$$

Thus, $\zeta = p_1 - p_2$ as before and ζ is estimated with $(x_1-(x_2 + x_3 + x_5)/3)/N$. The model can be tested with the usual chi-squre test, and it gave a good fit to the data in Wilcox (1982b).

More generally, for arbitrary c,

$$p_1 = \zeta + \nu_1/t \tag{2.9}$$

$$p_{c+1} = \nu_2 + \nu_1/t \tag{2.10}$$

and

$$p_i = \nu_i/t,\ i \neq 1,\ c + 1\ . \tag{2.11}$$

Slight generalizations of the model may be possible. Suppose, for example, c=3 and t=5, as in equations (2.4)-(2.8), but for examinees with misinformation, let ν_3 be the proportion of examinees who choose the correct response once c=3 alternatives are eliminated. Then p_5 and p_4 take the more general form

$$p_4 = \nu_3\nu_2 + \nu_1/t \qquad (2.12)$$

and

$$p_5 = (1-\nu_3)\nu_2 + \nu_1/t \qquad (2.13)$$

Now, however, a goodness-of-fit test is no longer possible because there are zero degrees of freedom.

Equivalent and Hierarchically Related Items, and Related Latent Structure Models

In recent years, several investigators have proposed models based on the notion of equivalent or hierarchically related items. Two items are said to be equivalent if examinees know both or neither one. If in addition, there are examinees who know the first but not the second, the items are hierarchically related. As argued by Molenaar (1981), clearly there are situations where it may be difficult or impossible to generate eqivalent items. However, experience suggests that there are situations where one of these assumptions might be reasonable (e.g., Macready and Dayton, 1977; Harris and Pearlman, 1978; Harris et al., 1980).

It should be mentioned that in some instances a test consisting of hierarchically related items is considered to be desirable and the goal is to measure the extent to which a test has this property. Put another way, the goal is to determine the extent to which the items on a test form a Guttman scale. One such measure was proposed by Cliff (1977). (See also Harnisch and Linn, 1981, and the paper by MacArthur in this volume.)

The simplest model consists of two equivalent items, and it arises as follows. Let ζ be the proportion of examinees who know both items. In contrast to earlier sections, a conventional scoring procedure is used. That is, examinees get only one attempt at an item, and the item is scored either correct or incorrect. Let p_{ij} be the probability of the response pattern ij (i=0,1; j=0,1) where a 0 means incorrect, and a 1 means correct. This, p_{10} represents the probability of a correct-incorrect response for a randomly sampled examinee. If β_1 is the probability of correctly guessing the response to the first item when the randomly sampled examinee does not know, and if β_2 is the corresponding probability on the second item, and if local independence holds (i.e., given an examinee's latent state, the responses are independent) then

$$p_{11} = \zeta + (1-\zeta)\beta_1\beta_2$$

$$p_{10} = (1-\zeta)\beta_1(1-\beta_2)$$

$$p_{01} = (1-\zeta)\beta_2\,(1-\beta_1)$$

$$p_{00} = (1-\zeta)(1-\beta_1)(1-\beta_2).$$

Solving for ζ, β_1, and β_2 yields

$$\beta_1 = \frac{p_{10}}{p_{10} + p_{00}}$$

$$\beta_2 = \frac{p_{01}}{p_{01} + p_{00}}$$

and

$$\zeta = 1 - (p_{01} + p_{00})(p_{10} + p_{00})/p_{00} \cdot$$

If x_{ij} is the number of examinees who have an ij response pattern, the unbiased maximum likelihood estimate of p_{ij} is $\hat{p}_{ij} = x_{ij}/N$ where, as before, N is the number of randomly sampled examinees. Thus, ζ can be estimated.

An interesting feature of the equivalent item model is that it is possible to include additional errors at the items level such as Pr(incorrect|examinee knows) (Macready and Dayton, 1977). However, estimating the parameters usually requires iterative procedures that are typically implemented on a computer. Goodman (1979) describes one such procedure, and Macready and Dayton (1977) used the scoring method (cf. Kale, 1962).

Testing Whether Two Items are Equivalent

One way to check the assumption of equivalent items is to apply the usual goodness-of-fit test as illustrated by Macready and Dayton (1977). For some cases, such as the equivalent item model described above, this cannot be done because there are zero degrees of freedom.

An alternative and relatively simple test of whether two items are equivalent is possible using an answer-until-correct scoring procedure.

For a randomly sampled examinee let p_{ij} be the probability of a correct response on the ith of the first item and the jth attempt of the second. If two items are indeed equivalent, and if for example, t=3, it can be seen that

$$p_{12} = p_{21} = p_{22}$$

$$p_{13} = p_{23}$$

and $$p_{31} = p_{23} .$$

For recent results on testing these equalities, see Smith et al. (1979), and Wilcox (1982e).

Hartke (1978) describes another approach based on latent partition analysis, and an index proposed by Baker and Hubert (1977) might also be useful.

Hierarchically Related Items

Dayton and Macready (1976, 1980) describe very general latent structure models for handling hierarchically related items. Again these models can be used to measure guessing, and they have the advantage of including other errors at the item level such as ζ = Pr(incorrect|examinee knows). The model for AUC tests essentially sets $\zeta = 0$, but the practical implications of this have not been established.

As was the case for equivalent items, estimating the parameters in the model requires iterative techniques. In some instances simple (closed form) estimates exist (e.g., Wilcox, 1980b), but these models make certain assumptions that may be unreasonable in many situations.

3. STRENGTHS AND WEAKNESSES OF LATENT CLASS MODELS

Latent class models have three primary strengths. First, it now appears that one of two models can be used to explain the observed responses to a multiple-choice test item (Wilcox, 1982b). These models are an oversimplification of reality (as are all models), but they seem to give a good approximation of how examinees behave when taking a test. Of course future investigations might reveal that more complex models are really needed, but so far this does not appear to be the case.

The second strength is that many measurement problems can now be solved that were previously impossible to address. In particular, these models correct for guessing, or measure the effects of guessing which in turn improves the accuracy of tests and measurement techniques. Note that the nature of guessing in latent class models is different from the guessing parameter in latent trait models (Wilcox, 1982c).

Third, even if some other model is ultimately preferred, a latent class model may be useful, for example, when estimating the item parameters in a latent trait model.

A weakness of latent class models is that certain technical problems still need to be solved. These include better ways of scoring an n-item test, testing the model used in Wilcox (1982e), and finding a strong true-score model that is reasonable when the model in Wilcox (1982a) gives a poor fit to data. Also, some examinees may give an incorrect response when they know, but the seriousness of this problem is not well understood.

4. PRESENT AREAS OF APPLICATION

This section outlines some of the measurement problems that can now be solved with latent class models.

The Accuracy of an Item and the Effectiveness of Distractors

In addition to estimating the proportion of examinees who know the item, the latent structure models for AUC tests can be used to estimate the probability of correctly determining whether a typical examinee knows the item. More specifically, assume it is decided that an examinee knows the correct response if the correct answer is given on the first attempt (i.e., a conventional scoring procedure is used). For a randomly sampled examinee, the probability of correctly determining whether he/she knows is just $\tau = 1-p_2$ (Wilcox, 1981a), and this is estimated with $\hat{\tau} = 1-x_2/N$. Note that when (2.2) is assumed $0 \leq p_2 \leq \frac{1}{2}$, in which case $\frac{1}{2} \leq \tau \leq 1$.

The parameter τ is a function of two important quantities. The first is the proportion of examinees who know the answer, i.e., ζ_{t-1}, and the second is the effectiveness of the distractors among the examinees who do not know. To see this more clearly, note that

$$\tau = \zeta_{t-1} + \sum_{i=2}^{t} p_i \ . \tag{4.1}$$

When ζ_{t-1} is close to one the item accurately reflects the true latent state of the examinees because presumably examinees who know will choose the correct response on their first attempt. As ζ_{t-1} moves closer to zero, the accuracy depends more on the effectiveness of the distractors. Thus, it may be important to determine how well distractors are performing among the examinees who do not know.

It can be shown that the distractors are most effective when guessing is at random which corresponds to

$$p_2 = p_3 = \cdots = p_t \tag{4.2}$$

(Wilcox, 1981a). This suggests (4.2) be tested, and/or we estimate how "far away" the p_i values are from the ideal case where (4.2) holds.

Testing (2.3) can be accomplished by noting that the conditional distribution of $x_2,\ldots,x_t$ given x_1 is multinomial with parameters $N-x_1$ and $p_i/(1-p_1)$, i = 2, ...,t. Thus, the ususal chi-square test can be applied. That is, compute

$$X^2 = \sum_{i=2}^{t} \frac{(x_i-(N-x_1)/(t-1))^2}{(N-x_1)/(t-1)} \tag{4.3}$$

If X^2 is greater than or equal to the $100(1-\alpha)$ percentile of the chi-square distribution with t-2 degrees of freedom, reject the hypothesis that (4.2) holds. For recent results on using (4.3), see Chacko (1966), Smith et al. (1979), Wilcox (1982e).

Empirical results indicate that guessing will not be at random. Thus, a more interesting question might be to determine whether the distractors are "close" to the ideal situation where (4.2) holds. The first step in solving this problem is to choose a measure of how unequal the p_i values are $(i = 2,\ldots,t)$. Many such measures have been proposed which have similar properties (e.g., Marshall and Olkin, 1979; Bowman et al., 1971). One of these is the entropy function which was used by Wilcox (1982a), and another is Simpson's measure of diversity (Simpson, 1949) given by

$$\sum_{i=2}^{t} [p_i/(1-p_1)]^2$$

Writing (4.3) as

$$-(N-x_1) + \frac{t-1}{N-x_1} \sum_{i=2}^{t} x_i^2 \quad ,$$

it is seen that the usual maximum likelihood estimate of Simpson's measure of diversity, namely, $\sum_{i=2}^{t} (x_i/(N-x_1))^2$, is a simple linear transformation of X^2. Since X^2 is better known than Simpson's measure of diversity, X^2 will be used here.

It is helpful to note that the smallest possible value for X^2 is

$$L = \frac{t-1}{n-x_1} [(n-x_1)(2r+1) - (t-1)r(r+1)] - n+x_1 \qquad (4.4)$$

where r is the largest integer satisfying $r(t-1) \leq n-x_1$ (Dahiya, 1971).

The maximum value is

$$M = (n-x_1)(t-2) \qquad (4.5)$$

(Smith et al., 1979). The closer X^2 is to M, the more effective are the distractors. Since L and M are known, the relative extent to which X^2 is close to M can be determined. In particular,

$$E=(X^2-L)/(M-L)$$

measures the effectiveness of the distractors being used, where $0 \leq E \leq 1$. If E=0, the distractors are as effective as possible in determining whether an examinee knows the correct response. As E approaches 1, the distractors become less effective.

Comparing Two Items

If the AUC model is assumed, and if independent estimates of the p_i values for two items are available, it is possible to test the hypothesis that one of the items is at least as effective as the second by applying results in Robertson and Wright (1981). The null hypothesis of interest here is that $\sum_{i=2}^{k} p_i/(1-p_1) \geq \sum_{i=2}^{k} p_i'/(1-p_1')$, k=2,...,t-2 where p_i'

is the p_i value for the second item. Let τ_1 and τ_2 be the value of τ for two items. Another way of comparing two items is to test whether the first item is better than the second by testing whether $\tau_1 \geq \tau_2$. In effect this approach compares the overall effectiveness of the two items in terms of the population of examinees, while the approach previously described is to compare the effectiveness of the distractors among the examinees who do not know.

Characterizing Tests

Let τ_i be the value of τ for the i<u>th</u> item on an n-item test. A natural way of describing the accuracy of a test is to use $\tau_s = \sum_{i=1}^{n} \tau_i$. This is the expected number of correct decisions about whether a typical (randomly sampled) examinee knows the answer to the items on a test. If, for example, $\tau_s = 7$ and $n = 10$, then on the average, 7 correct decisions would be made about whether an examinee knows the answer to an item, but for 3 of the items it would be decided that the examinee knows when in fact he/she does not.

Estimating τ_s is easily accomplished using previous results. In particular, for a random sample of N examinees, let $x_{ij} = 0$ if the j<u>th</u> examinee gets the i<u>th</u> item correct on the <u>second</u> attempt; otherwise $x_{ij} = 1$. Then

$$\hat{\tau}_s = N^{-1} \sum_{i=1}^{n} \sum_{j=1}^{N} x_{ij}$$

is an unbiased estimate of τ_s.

The k Out of n Reliability of a Te.t

Once test data is available, the question arises as to how certain we can be that τ_s is large or small. That is, we want to estimate the $Pr(\hat{\tau}_s \geq \tau_0)$(cf. Tong, 1978). This problem is similar to one found in the engineering literature where the goal is to estimate the k out of n reliability of a system. Bounds on this probability can be estimated without assuming anything about $cov(x_{ij}, x_{i'j'})$ (Wilcox, 1982e). The procedure is outlined below.

Let $z_i=1$ if a correct decision is made about whether a randomly sampled examinee knows the i^{th} item on a test; otherwise $z_i=0$. For a randomly sampled examinee $Pr(z_i=1) = \tau_i$. Note that from previous results $Pr(z_i=1) = Pr(x_{ij}=1)$. The k out of n reliability of a test is defined to be

$$\rho_K = Pr(\sum z_i \geq K)$$

This is the probability that for a typical examinee, at least k correct decisions are made among the n items on a test. By a correct decision is meant the event of correctly determining whether the examinee knows an item. Knowing ρ_k yields additional and important information about the accuracy of a test. An estimate of ρ_k is not available unless $cov(z_i,z_j)=0$, or n, the number of items, is small. (See Wilcox, 1982g, 1982j.)

For any two items, let p_{km} be the probability that a randomly selected examinee chooses the correct response on the k^{th} attempt of the first item, and the m^{th} attempt of the second. (It is assumed that both items are administered according to an AUC scoring procedure.) Let $\kappa_{ij}(i=0,\ldots,t-1;\ j=0,\ldots,t-1)$ be the proportion of examinees who can eliminate i distractors on the first item and j distractors on the second. Then, under certain mild independence assumptions

$$p_{km} = \sum_{i=0}^{t-k} \sum_{j=0}^{t-m} \kappa_{ij}/[(t-i)(t-j)] .$$

The equation makes it possible to express the κ_{ij}'s in terms of the p_{km}'s which in turn makes it possible to estimate κ_{ij} for any i and j.

Next let ε be the probability that for both items, a correct decision is made about an examinee's latent state. It can be seen that

$$\varepsilon = \kappa_{t-1,t-1} + 1-p_{11}$$

and so ε can also be estimated.

For the i^th^ and j^th^ item on a test, let ε_{ij} be the value of ε, and define

$$S = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \varepsilon_{ij}$$

$$U_K = \tau_s - K$$

where τ_s was previously defined to be $\Sigma\tau_i$ and

$$V_K = (2S - K(K-1)/2).$$

Then from Sathe et al. (1980)

$$\rho_K \geq (2V_{K-1} - (K-2))\, U_{K-1}/[n(n-K+1)]$$

If $2V_{K-1} \leq (n+K-2)U_{K-1}$

$$\rho_K \geq \frac{2((K^*-1)U_{K-1} - V_{K-1}}{(K^*-K)(K^*-K+1)}$$

where $K^* + K - 3$ is the largest integer in $2V_{K-1}/U_{K-1}$. Two upper bounds are also available. The first is

$$\rho_K \leq 1 + ((n+K-1)U_K - 2V_K)/Kn$$

and the second is that if $2V_k \leq (K-1)U_K$;

$$\rho_K \leq 1 - 2\,\frac{(K^*-1)U_K - V_K}{(K-K^*)(K-K^*+1)}$$

where $K^* + K - 1$ is the largest integer in $2V_K/U_K$.

What these results mean is that we can estimate quantities that indicate whether ρ_k is large or small. For example, suppose the right side of the third to last inequality is estimated to be .9, and that $2V_{k-1} \leq (n+K-2)U_{k-1}$. This does not yield an exact estimate of ρ_k but it does say that ρ_k is estimated to be least .9. Thus, this would indicate that the overall test is fairly accurate. If, for example, the above inequalities indicate that $\rho_k \leq .95$ and $\rho_k \geq .1$, this does not give very useful information about whether ρ_k is reasonably large. If $\rho_k \leq .1$ we have a poor test.

Estimating the Proportion of Items an Examinee Knows

It is a simple matter to extend previous results to situations when a single examinee responds to items randomly sampled from some item domain. For example, let q_i be the probability of a correct response on the $i^{\underline{th}}$ attempt of a randomly sampled item. Let $\gamma_i(i=0, \ldots , t=1)$ be the proportion of items for which the examinee can eliminate i distractors. It is assumed that each item has at least one effective distractor, so γ_{t-1} is the proportion of items the examinee knows. It follows that

$$q_i = \sum_{j=0}^{t-i} \gamma_j/(t-i)$$

which is the same as equation (2.0) where p_i and ζ_i are replaced with q_i and γ_i. In fact, all previous results extend immediately to the present case.

Criterion-Referenced Tests

A common goal of a criterion-referenced test is to sort examinees into two categories. (See Hambleton et al., 1978a; Berk, 1980; and the 1980 special issue of Applied Psychological Measurement.) Frequently these categories are defined in terms of some true score, and here the true score of interest is γ_{t-1}, the proportion of items in an item domain that an examinee knows. The goal is to determine whether γ_{t-1} is larger or smaller than some predetermined constant, say γ'.

It is known that guessing can seriously affect the accuracy of a criterion-referenced test (van den Brink and Koele, 1980). Moreover, assuming random guessing can be highly unsatisfactory (Wilcox, 1980c). Another advantage of the AUC scoring model is that it substantially reduces this problem (Wilcox, in press, b). For some results on comparing γ_{t-1} to γ' when equivalent items are available, see Wilcox (1980a).

Sequential and Computerized Testing

In certain situations, such as in computerized testing, sequential procedures will be convenient to use. Some progress has been made in this area, but much remains to be done.

Suppose an examinee responds to items randomly sampled from an item domain and presented on a computer terminal. Further suppose the examinee responds according to an AUC scoring procedure. A typical sequential procedure for this situation is to continue sampling until there are n items for which the examinee gives a correct response on the first attempt. Let y_i $(i=1, \ldots , t)$ be the number of items for which the examinee requires i attempts to get the correct response. For the sequential procedure just described, sampling continues until $y_1 = n$, in which case the joint probability function of $y_2, \ldots , y_t$ is negative multinomial given by

$$f(y_2,\ldots,y_t|q_1,\ldots,q_t) = n\Gamma(y_0) \prod_{i=j}^{n} p_i^{y_i}/y_i!$$

where $y_0 = \sum_{i=1}^{t} y_i$, and for $i \geq 2$, $y_i = 0,1,\ldots$

The problem with the sequential procedure just described is that with positive probability, the number of sampled items will be too large for practical purposes. This might be an extremely rare event, but it is desirable to avoid this possibility all together. A solution to this problem is to use a closed sequential procedure where sampling continues until $y_1=n_1$, or $y_2=n_2$, etc. where $n_1,\ldots,n_t$ are positive integers chosen by the investigator. In this case the joint probability function of $y_1,\ldots,y_t$ is

$$\Gamma(y_0)(\sum y_i \quad I_{[y_i=n_i]}) \prod_{i=1}^{t} p_i^{y_i}/y_i!$$

where I is the usual indicator function given by

$$I_{[y_i=n_i]} = \{ \begin{matrix} 1, \text{if } y_i=n_i \\ 0, \text{ if otherwise} \end{matrix}$$

For the special case $n_1=n_2=\ldots=n$, the probability function becomes

$$n\Gamma(y_0) \prod_{i=1}^{t} p_i^{y_i}/y_i!$$

which has the same form as the negative multinomial except that for some j, $y_j=n$, and $0 \leq y_i \leq n-1$, $i \neq j$.

The maximum likelihood estimate of q_i is $\hat{q}_i = y_i/y_0$, so the maximum likelihood estimate of γ_{t-1}, the proportion of items an examinee knows, is $\hat{\gamma}_{t-1} = \hat{q}_1 - \hat{q}_2$ (Zehna, 1966). If the model is assumed to hold, $\hat{\gamma}_{t-1}$ may not be a maximum likelihood estimate. Instead one would estimate γ_{t-1} to be zero when $\hat{\gamma}_{t-1} \leq 0$; if the estimates of q_i $(i=1,\ldots,t)$ do not satisfy the inequality $q_1 \geq q_2 \geq \cdots \geq q_t$ apply the pool-adjacent-violators algorithm (Barlow et al., 1972).

Wilcox (in press) shows that if the goal is to compare γ_{t-1} to the known constant γ', as in criterion-referenced testing, and if $\gamma_{t-1} \geq \gamma'$ is decided if and only if $\hat{\gamma}_{t-1} \geq \gamma'$ the sequential and closed sequential procedures have the same level of accuracy. Moreover, it appears that the closed sequential procedures nearly always improves upon the more conventional fixed sample approach. More recently Wilcox (1982f) proposed two tests of $q_1=\ldots=q_t$, and methods of determining the moments of the distribution were also described.

A Strong True Score Model

Strong true score models attempt to relate a population of examinees to a domain of items. In many situations an item domain does not exist de facto, in which case strong true score models attempt to find a family of probability functions for describing the observed test scores of any examinee, and simultaneously to find a distribution that can be used to describe the examinees' true score.

Perhaps the best known model is the beta-binomial. If y is the number of correct responses from an examinee taking an n-item test, it is assumed that for a specific examinee, the probability function of y is

$$\binom{n}{y} q^y (1-q)^{n-y}$$

For the population of examinees, it is assumed that the distribution of q is given by

$$g(q) = \frac{\Gamma(r+s)}{\Gamma(r)\Gamma(s)} q^{r-1} (1-q)^{s-1}$$

where $r > 0$ and $s > 0$ are unknown parameters that are estimated with observed test scores. Apparently Keats (1951) was the first to consider this model in mental test theory.

The beta-binomial model has certain theoretical disadvantages, but experience suggests that it frequently gives good results with real data. A review of these results is given by Wilcox (1981d). However, the model does not always give a good fit to data, and some caution should be exercised (Keats, 1964). In the event of a poor fit, a gamma-Poisson model might be considered (Wilcox, 1981d).

When the beta-binomial is assumed, many measurement problems can be solved. These include equating tests by the equipercentile method, estimating the frequency of observed scores when a test is lengthened, and estimating the effects of selecting individuals on a fallible measure (Lord, 1965). Other applications include estimating the reliability of a criterion-referenced test (Huynh, 1976a), estimating the accuracy of a criterion-referenced test (Wilcox, 1977c), and determining passing scores (Huynh, 1976b).

A problem with the beta-binomial model is that it ignores guessing. Attempts to remedy this problem are summarized by Wilcox (1981d), but all of these solutions now appear to be unsatisfactory in most situations. This is unfortunate because it means that a slightly more complex model must be used. More recently, however, Wilcox (1982a, 1982b) proposed a generalization of the beta-binomial model that takes guessing into account, and which gives a reasonably good fit to data.

Some Miscellaneous Applications of Latent Structure Models

Several applications of latent structure models have already been described, and there are several other situations where they may be useful. For example, Ashler derives an expression for the biserial correlation coefficient that includes ζ_{t-1}, the proportion of examinees who know an item. Wilcox (1982g) discusses how to empirically determine the number of distractors needed on a multiple choice test item, and Knapp (1977) discusses a reliability coefficient based on the latent state point of view. (See also Frary, 1969.) Macready and Dayton (1977) illustrate how the models can be used to determine the number of equivalent items needed for measuring an instructional objective, and Emrick (1971) shows how the models might be used to determine passing scores. Note that Emrick's estimation procedure is incorrect (Wilcox and Harris, 1977), but this is easily remedied using the estimation procedures already mentioned; closed form estimates are given by van der Linden (1981).

5. POSSIBLE EXTENSIONS AND CONTROVERSIAL ISSUES

The AUC models assumed that examinees eliminate as many distractors as they can and then guess at random from among the alternatives that remain. A recent empirical investigation suggests that the random guessing portion of this assumption will usually give a reasonable approximation of reality (Wilcox, 1982k). No doubt there will be cases where this assumption is untenable in which case there are no guidelines on how to proceed.

A theoretical advantage of the latent structure model based on equivalent or hierarchically related items is that they included not only guessing, but errors such as Pr(incorrect response|examinee knows). The practical implications of this are not well understood.

Wilcox (1981a) mentions that under an item sampling model for AUC tests, an examinee with partial information can improve his/her test score by choosing a response, and if it is incorrect, deliberately choose another incorrect response. Thus, if $(y_1-y_2)/n$ is used to estimate ζ_{t-1}, the estimate would be higher for such an examinee because y_2 is lower. Four points should be made. First, this problem can be partially corrected by estimating the q_i's with the pool-adjacent-violators algorithm (Barlow et al., 1972, pp. 13-15). Second, if an examinee is acting as described, it is still possible to correct for guessing by applying the true score model proposed by Wilcox (1982a). If it gives a good fit to data, estimate ζ_{t-1} to be $q_1-(1-q_1)\xi(q_1)$. The third point is that there is no indication of how serious this problem might be. Finally, a new scoring procedure is being examined that might eliminate the problem.

It has been argued (e.g., Messick, 1975) that tests should be homogeneous in some sense. Frequently this means that at a minimum, a test should have a single factor. A sufficient condition for the best known latent trait models (see e.g., Lord, 1980; Wainer et al., 1980; Hambleton et al., 1978b; Choppin, this volume) is that this assumption be met (cf. McDonald, 1981). In general, the latent structure models described in this paper do not require this assumption. One exception is the equivalent item model. (See Harris and Pearlman, 1978.) The point is that in this paper, no stand on this issue is needed, i.e., it is irrelevant whether a test is homogeneous when applying, say, the answer-until-correct scoring procedure, or the corresponding strong true-score model.

Wainer and Wright (1980) and Mislevy and Bock (1982) have studied the effects of guessing on latent trait models, but these investigations do not take into account the results and type of guessing described here. If guessing proves to be a problem, perhaps latent class models can be of use when latent trait models are applied.

REFERENCES

Baker, F. B., & Hubert, L. J. Inference procedures for ordering theory. Journal of Educational Statistics, 1977, 2, 217-233.

Barlow, R., Bartholomew, D., Bremner, J., & Brunk, H. Statistical inference under order restrictions. New York: Wiley, 1972.

Berk, R. Criterion-referenced measurement. Baltimore: The Johns Hopkins University Press, 1980.

Bliss, L. B. A test of Lord's assumption regarding examinee guessing behavior on multiple-choice tests using elementary school students. Journal of Educational Measurement, 1980, 17, 147-153.

Bowman, K., Hutcheson, K., Odum, E., & Shenton, L. Comments on the distribution of indices of diversity. In G. Patil, E. Pielou, and W. Waters (Eds.) International Symposium on Statistical Ecology, Vol. 3. University Park: Pennsylvania State Press, 1971.

Chacko, V. J. Modified chi-square test for ordered alternatives. Sankhya, 1966, Ser. B, 28, 185-190.

Cliff, N. A theory of consistency of ordering generalizable to tailored testing. Psychometrika, 1977, 42, 375-399.

Coombs, C. H., Milholland, J. E., & Womer, F. B. The assessment of partial information. Educational and Psychological Measurement, 1956, 16, 13-27.

Cross, L. H., & Frary, R. B. An empirical test of Lord's theoretical results regarding formula-scoring of multiple-choice tests. Journal of Educational Measurement, 1977, 14, 313-321.

Dahiya, R. C. On the Pearson chi-squared goodness-of-fit test statistic. Biometrika, 1971, 58, 685-686.

Dayton, C. M., & Macready, G. B. A probabilistic model for validation of behavioral hierarchies. Psychometrika, 1976, 41, 189-204.

Dayton, C. M., & Macready, G. B. A scaling model with response errors and intrinsically unscalable respondents. Psychometrika, 1980, 45, 343-356.

Emrick, J. A. An evaluation model for mastery testing. Journal of Educational Measurement, 1971, 8, 321-326.

Frary, R. B. Reliability of multiple-choice test scores is not the proportion of variance which is true variance. Educational and Psychological Measurement, 1969, 29, 359-365.

Goodman, L. A. On the estimation of parameters in latent structure analysis. Psychometrika, 1979, 44, 123-128.

Hambleton, R. K., Swaminathan, H., Algina, J., & Coulson, D. B. Criterion-referenced testing and measurement: A review of technical issues and developments. Review of Educational Research, 1978, 48, 1-48. (a)

Hambleton, R. K., Swaminathan, H., Cook, L. L., Eignor, D. R., Gifford, J. A. Developments in latent trait theory: Models, technical issues, and application. Review of Educational Research, 1978, 48, 467-510.

Harnisch, D. L., & Linn, R. L. Analysis of item response patterns: Questionable test data and dissimilar curriculum practices. Journal of Educational Measurement, 1981, 18, 133-146.

Harris, C. W., Houang, R. T., Pearlman, A. P., & Barnett, B. Final report submitted to the National Institute of Education. Grant No. NIE-G-78-0085, Project No. 8-0244, 1980.

Harris, C. W., & Pearlman, A. An index for a domain of completion or short answer items. Journal of Educational Statistics, 1978, 3, 285-304.

Hartke, A. R. The use of latent partition analysis to identify homogeneity of an item population. Journal of Educational Measurement, 1978, 15, 43-47.

Huynh, H. On the reliability of decisions in domain-referenced testing. Journal of Educational Measurement, 1976, 13, 253-264. (a)

Huynh, H. Statistical consideration of mastery scores. Psychometrika, 1976, 41, 65-78. (b)

Kale, B. K. On the solution of likelihood equations by iteration processes. The multiparametric case. Biometrika, 1962, 49, 479-486.

Keats, J. A. A statistical theory of objective test scores. Melbourne: A.C.E.R., 1951.

Keats, J. A. Some generalizations of a theoretical distribution of mental test scores. Psychometrika, 1964, 29, 215-231.

Knapp, T. R. The reliability of a dichotomous test item: A correlationless approach. Journal of Educational Measurement, 1977, 14, 237-252.

Lord, F. M. A true-score theory, with applications. Psychometrika, 1965, 30, 239-270.

Lord, F. M. Applications of item response theory to practical testing problems. Hillsdale, New Jersey: Erlbaum, 1980.

Macready, G. B., & Dayton, C. M. The use of probabilistic models in the assessment of mastery. Journal of Educational Statistics, 1977, 2, 99-120.

Marshall, A. W., & Olkin, I. Inequalities: Theory of majorization and its applications. New York: Academic Press, 1979.

McDonald, R. P. The dimensionality of tests. British Journal of Mathematical and Statistical Psychology, 1981, 34, 100-117.

Messick, S. The standard problem: Meaning and values in measurement and evaluation. American Psychologist, 1975, 30, 955-966.

Mislevy, R. J., & Bock, R. D. Biweight estimates of latent ability. Educational and Psychological Measurement, 1982, 42, 725-737.

Molenaar, I. On Wilcox's latent structure model for guessing. British Journal of Mathematical and Statistical Psychology, 1981, 34, 79-89.

Robertson, T. Testing for and against an order restriction on multinomial parameters. Journal of the American Statistical Association, 1978, 73, 197-202.

Robertson, T., & Wright, F. T. Likelihood ratio tests for and against a stochastic ordering between multinomial populations. Annals of Statistics, 1981, 9, 1248-1257.

Sathe, Y. S., Pradhan, M., & Shah, S. P. Inequalities for the probability of the occurrence of at least m out of n events. Journal of Applied Probability, 1980, 17, 1127-1132.

Simpson, E. Measurement of diversity. Nature, 1949, 163, 688.

Smith, P. J., Rae, D. S., Manderscheid, R., & Silberg, S. Exact and approximate distributions of the chi-square statistic for equiprobability. Communications in Statistics--Simulation and Computation, 1979, B8, 131-149.

van den Brink, W. P., & Koele, P. Item sampling, guessing and decision-making in achievement testing. British Journal of Mathematical and Statistical Psychology, 1980, 33, 104-108.

van der Linden, W. Estimating the parameters of Emrick's mastery testing model. Applied Psychological Measurement, 1981, 5, 517-530.

Wainer, H., Morgan, A., & Gustafson, J. A review of estimation procedures for the Rasch model with an eye toward longish tests. Journal of Educational Statistics, 1980, 5, 35-64.

Wilcox, R. R. Estimating the likelihood of false-positive and false-negative decisions in mastery testing: An empirical Bayes approach. Journal of Educational Statistics, 1977, 2, 289-307. (c)

Wilcox, R. R. Determining the length of a criterion-referenced test. Applied Psychological Measurement, 1980, 4, 425-446. (a)

Wilcox, R. R. Some results and comments on using latent structure models to measure achievement. Educational and Psychological Measurement, 1980, 40, 645-658. (b)

Wilcox, R. R. An approach to measuring the achievement or proficiency of an examinee. Applied Psychological Measurement, 1980, 4, 241-251. (c)

Wilcox, R. R. Solving measurement problems with an answer-until-correct scoring procedure. Applied Psychological Measurement, 1981, 5, 399-414. (a)

Wilcox, R. R. A review of the beta-binomial model and its extensions. Journal of Educational Statistics, 1981, 6, 3-32. (d)

Wilcox, R. R. Some empirical and theoretical results on an answer-until-correct scoring procedure. British Journal of Mathematical and Statistical Psychology, 1982, 35, 57-70. (a)

Wilcox, R. R. Some new results on an answer-until-correct scoring procedure. Journal of Educational Measurement, 1982, 19, 67-74.

Wilcox, R. R. Approaches to measuring achievement with an emphasis on latent structure models. Technical Report, Center for the Study of Evaluation, University of California, Los Angeles, 1982. (c)

Wilcox, R. R. Bounds on the k out of n reliability of a test, and an exact test for hierarchically related items. Applied Psychological Measurement, 1982, 6, 327-336. (e)

Wilcox, R. R. On a closed sequential procedure for categorical data, and tests for equiprobable cells. British Journal of Mathematical and Statistical Psychology, 1982, to appear. (f)

Wilcox, R. R. Using results on k out of n system reliability to study and characterize tests. Educational and Psychological Measurement, 1982, 42, 153-165. (g)

Wilcox, R. R. An approximation of the k out of n reliability of a test, and a scoring procedure for determining which items an examinee knows. Center for the Study of Evaluation, University of California, Los Angeles, 1982. (j)

Wilcox, R. R. How do examinees behave when taking multiple-choice tests. *Applied Psychological Measurement*, 1982, to appear. (k)

Zehna, P. W. Invariance of maximum likelihood estimation. *Annals of Mathematical Statistics*, 1966, *37*, 744.

GENERALIZABILITY THEORY

Noreen Webb
University of California, Los Angeles

Definition and Focus

Generalizability theory evolved out of the recognition that the concept of undifferentiated eror in classical test theory provided too gross a characterization of the multiple sources of error in a measurement. The multidimensional nature of measurement error can be seen in how a test score is obtained. For example, one of many possible test forms might be admiistered on one of many possible occasions by one of many possible testers. Each of these choices--test form, occasion and tester--is a potential source of error. G-theory attempts to assess each source of error in order to characterize the measurement and improve its design.

A behavioral measurement, then, is a sample from a universe of admissible obsrvations characterized by one or more facets (e.g., test forms, occasions, testers)[1]. This universe is usually defined by the Cartesian product of the levels (called conditions in G-theory) of the facets. From this perspective, Cronbach et al. (1972, p. 15) say:

> The score on which the decision is to be based is only one of many scores that might serve the same purpose. The decision maker is almost never interested in the response given to the particular stimulus objects or questions, to the particular tester at the particular moment of testing. Some, at least, of these conditions of measurement could be altered without making the score any less acceptable to the decision maker. That is to say, there is a universe of observations, any of which would have yielded a usable basis

[1] Introduction to G-theory are provided by Brennan (1977a, 1979a) Brennan and Kane (1980), Cronbach et al. (1972), Erlich and Shavelson (1976b) Gillmore (1979) Cardinet and Tourneur (1978), Huysamen (1980), Shavelson and Webb (1981), Tourneur (1978), Tourneur and Cardinet (1977), Van der Kamp (1976), and Wiggins (1973).

> for the decision. The ideal datum on which to base the decision would be something like the person's mean score over all acceptable observations, which we shall call his "universe score." The investigator uses the observed score or some function of it as if it were the universe score. That is, he generalizes from sample to universe. The question of "reliability" thus resolves into a question of accuracy of of generalization or generalizability.

Since different measurements may represent different universes, G-theory speaks of universe scores rather than true scores, acknowledging that there are different universes to which decision makers may generalize. Likewise the theory speaks of generalizability coefficients rather than the reliability coefficient realizing that the value of the coefficient may change as definitons of universes change.

G-theory distinguishes a decision (D) study from a generalizability (G) study. This distinction recognizes that certain studies are associated with the development of a measurement procedure (G studies) while other studies then apply the procedure (D studies). Although the decision-maker must begin to plan the D study before conducting the G study, the results of the G study will guide the specification of the D study. In planning the D study, the decision maker (a) defines the universe of generalization and (b) specifies his proposed interpretation of a measurement. These plans determine (c) the questions to be asked of the G study data in order to optimize the measurement design. Each of these points is considered in turn.

(a) G-theory recognizes that the universe of admissible observations encompassed by a G study may be broader than the universe to which a decision maker wishes to generalize. That is, the decision maker proposes to generalize to a universe comprised of some subset of the facets in the G study. The universe is called the universe of generalization. It may be defined by reducing the universe of admissible observations, i.e. by reducing the levels of a facet (creating a fixed facet; cf. fixed factor in ANOVA) by selecting and thereby controlling one level of a facet, or by ignoring a facet. All three alternatives have consequences for the estimation of the components of error variance that enter into the observed score variance.

(b) G-theory recognizes that decision makers use the same test score in different ways. For example, some interpretations may focus on individual differenes (i.e., relative or comparative decisions), some may use the observed score as an estimate of a person's universe score (absolute decisions; cf. criterion-referenced interpretations), while still others may use the observed score in a regression estimate of the universe score (cf.Kelley's, 1947, regression estimate of true scores). There is a different error associated with each of these proposed interpretations.

To illustrate the distinction between relative and absolute decisions, suppose that a decision is to be made using scores on an objective test of arithmetic. As an example of a relative decision, a decision-maker might want to channel the top 20 percent of the scorers into an above-average academic track (regardless of their actual scores). In this case, if all items on the test rank students in the same way, even if some items are more difficult than others, it would not matter to a student which items he or she received. The same students would be sc ected for the accelerated track whether the test consists of easy items or difficult items. In more formal terms the variation in item means would not be a part of error. As an example of an absolute decision, a decision-maker might want to select for accelerated placement all students who answer correctly 75 percent or more of the items on the test. In this case, the variation in item means would contribute to error. Even if all items rank students in the same way, a test composed of easy items would place more students into the accelerated program than a test composed of difficult items.

(c) Ordinarily, the universe of admissible observations in a G study is defined as broadly as possible within practical and theoretical constraints. In most cases Cronbach et al. recommend using a crossed G study design so that all sources of error and interactions among sources of error can be estimated. (It should be noted, however, that a nested G study is sometimes useful because it provides more degrees of freedom for some estimates of sources of error.) The design of D studies, on the other hand, can vary widely and include crossed partially nested, and completely nested designs. Often, in D studies, nested designs are used for convenience, to reduce costs, for increasing sample size, orfor a combination of these reasons. All facets in the D study design may be random or only some may be random.

Development of the Model

Scores and variance components. In G-theory a person's score is decomposed into a component for the universe score (μ_p) and one or more error components. To illustrate this decomposition, we consider the simplest case for podagogical purposes--a one facet, p x i (person by, say, item) design. (The object of measurement, here persons, is not a source of error and, therefore, is not a facet.) The presentation readily generalizes to more complex designs. In the p x i design with generalization over all admissible items taken from an indefinitely large univese the score for a particular person (p) on a particular form (i) is:

$$
\begin{aligned}
X_{pi} &= \mu && \text{(grand mean)} \\
&+ \mu_p - \mu && \text{(person effect)} \\
&+ \mu_i - \mu && \text{(item effect)} \\
&+ X_{pi} - \mu_p - \mu_i + \mu && \text{(residual)}
\end{aligned} \tag{1}
$$

Since this design is crossed all persons receive the same items. Except for the grand mean, each score component has a distribution. Considering all persons in the population, there is a distribution of $\mu_p - \mu$ with mean zero and variance $\xi(\mu_p - \mu)^2 = \sigma_p^2$ which is called the universe-score variance and is analogous to the true-score variance of classical theory. Similarly, the component for item has mean zero and variance $\xi(\mu_i - \mu)^2 = \sigma_i^2$ which indicates the variance of constant errors associated with items while the residual component has mean zero and variance $\sigma_{pi,e}^2$ which indicates the person x item interaction confounded with residual error, since there is one observation per cell. The collection of observed scores. X_{pi} has a variance of $\sigma^2_{X_{pi}} = \xi (X_{pi} - \mu)^2$ which equals the sum of the variance components:

$$\sigma^2_{X_{pi}} = \sigma^2_p + \sigma^2_i + \sigma^2_{pi,e} \tag{2}$$

G-theory focuses on these variance components They are estimated by means of a generalizability (G) study. The relative magnitudes of the components provide information about particular sources of error influencing a measurement. It is convenient to estimate variance components from an ANOVA of sample data. Numerical estimates of the variance components are obtained by setting the expected mean squares equal to the observed mean squares and solving the set of simultaneous equations as shown in Table 1.

Table 1

Estimates of Variance Components for a One Facet p x i Design

Source of Variation	Mean Square	Expected Mean Square*	Estimated Varianced Component
Person (p)	MS_p	$\sigma^2_{pi,e} + n_i\sigma^2_p$	$\hat{\sigma}^2_p = (MS_p - MS_{res})/n_i$
Item (i)	MS_i	$\sigma^2_{pi,e} + n_p\sigma^2_i$	$\hat{\sigma}^2_i = (MS_i - MS_{res})/n_p$
pi,e	MS_{res}	$\sigma^2_{pi,e}$	$\hat{\sigma}^2_{pi,e} = MS_{res}$

*n_i = number of items; n_p = number of persons.

Estimation of error. Not only do the magnitudes of the variance components show the importance of each source of error in the measurement, they can be used to estimate the total error for relative and absolute decisions. For relative decisions, the error in a p x i design is defined as:

$$(3) \qquad \delta_{pI} = (X_{pI} - \mu_I) - (\mu_p - \mu) ,$$

where I indicates that an average has been taken over the levels of facet i under which p was observed. The variance of the errors for relative decisions is:

$$(4) \qquad \sigma^2_\delta = \sigma^2_{pI} = \sigma^2_{pi,e}/n_i' ,$$

where n_i' indicates the number of conditions of facet i to be sampled in a D study. Notice that (a) $\sigma^2_{pi,e}/n_i'$ is the standard eror of the mean of a person's scores averaged over the levels of i (items in our example). And (b) the magnitude of the error is under the control of the decision maker in the D study. In order to reduce σ^2_δ, n_i' may be increased. This is analogous to the Spearman-Brown prophecy formula in classical theory and the standard error of the mean in sampling theory.

For absolute decisions, the error is defined as:

$$(5) \qquad \Delta_{pI} = X_{pI} - \mu_p$$

The variance of these errors in a p x i design is:

$$(6) \qquad \sigma^2_\Delta = \sigma^2_I + \sigma^2_{pI} = \sigma^2_i/n_i + \sigma^2_{pi,e}/n_i$$

In contrast to σ^2_δ, σ^2_Δ includes the variance of constant errors associated with facet i (σ^2_i). This arises because, in absolute decisions, the difficulty of the particular items that a person receives will influence his observed score and, hence, the decision maker's estimate of his universe score. For relative decisions, however, the effect of item is constant for all persons and so does not influence the rank ordering of them (see Erlich & Shavelson, 1976b).

Finally, for decisions based on the regression estimate of a person's universe score, error (of estimate) is defined as:

$$(7) \qquad \varepsilon_p = \hat{\mu}_p - \mu_p ,$$

where $\hat{\mu}_p$ is the regression estimate of a person's universe score, . The estimation procedure for the variance of errors of estimate may be found in Cronbach et al. (1972, p. 97ff).

The variance components from a crossed p x i G study design can also be used to estimate error in a nested D study design with items nested within persons (we write i:p to denote nesting). So, the effect of the constant errors associated with facet i is confounded with the effect associated with the person by i-facet interaction (pi,e). Hence,

$$\sigma^2_{X_{pI}} = \sigma^2_p + \sigma^2_{I,pI,e} = \sigma^2_p + \sigma^2_\Delta \tag{8}$$

Note that, for a completely nested design, $\sigma^2_\delta = \sigma^2_\Delta$.

Generalizability coefficients. While stressing the importance of variance components and errors such as σ^2_δ , generalizability theory also provides a coefficient analogous to the reliability coefficient in classical theory. A generalizability (G) coefficient can be estimated for each of a variety of D study designs using the estimates of variance components and error produced by the G study. A decision-maker can then use the estimated G coefficients to choose among the D study designs. For the one-facet case described here, generalizability coefficients can be estimated for crossed or nested D study designs with any number of items. For designs with more than one facet, there are many D study designs possible each with an estimated G coefficient.

The generalizatility (G) coefficient, $\xi\rho^2$, for relative decisions is defined as the ratio of the universe-score variance to the expected observed-score variance, i.e., an intraclass correlation:

$$\xi\rho^2 = \frac{\sigma^2_p}{\xi\sigma^2(X)} = \frac{\sigma^2_p}{\sigma^2_p + \sigma^2_\delta} \tag{9}$$

The expected observed-score variance is used in G-theory because the theory assumes only random sampling of the levels of facets and so the observed-score variance may change from one application of the design to another. Sample estimates of the parameters in (9) are used to estimate the G coefficient:

$$(9a) \qquad \xi\hat{\rho}^2 = \frac{\hat{\sigma}^2_p}{\hat{\sigma}^2_p + \hat{\sigma}^2_\delta} .$$

$\xi\hat{\rho}^2$ is a biased but consistent estimator of $\xi\rho^2$.

For absolute decisions a generalizability coefficient can be defined in an analogous manner:

$$(10) \qquad \xi\rho^2 = \frac{\sigma^2_p}{\sigma^2_p + \sigma^2_\Delta} \quad \text{and}$$

$$(10a) \qquad \xi\hat{\rho}^2 = \frac{\sigma^2_p}{\hat{\sigma}^2_p + \hat{\sigma}^2_\Delta}$$

Finally, note that, for completely nested designs regardless of whether relative or absolute decisions are to be made, error variance is defined as σ^2_Δ and so (10) provides the generalizability coefficient for such designs.

A two-faceted example. A study of the dependability of measures of mathematics achievement illustrates the theory's treatment of multifaceted measurement error. In designing a generalizability (G) study, the decision-maker specifies possible sources of error in the measurement of mathematics achievement. Variablility across test items is clearly a possible source of error. Furthermore students may obtain different scores on multiple occasions even though no learning has taken place between occasions, so occasions is a possible source of error. (It is assumed that true ability is constant from one occasion to the next. Therefore, a time interval between occasions must be selected that is short enought to prevent true changes from taking place--learning or maturation--but is long enough to prevent students' memory of the test from influencing their scores.) Another source of error might be item format, such as multiple choice, true-false, or open-answer (student fills in the correct answer): Students' scores might differ across item formats. For the present illustration, the item and occasion sources of error will be considered.

In the generalizability study, thirty tenth-grade students (p) were administered a twenty-item (i) test on two occasions (j). In differentiating students with respect to mathematics achievement, errors in the measurement may arise from inconsistencies associated with items, occasions, and other unidentified sources. G-theory incorporates these potential sources of error into a measurement model and estimates the components of variance associated with each source of variation in the 30 x 20 x 2 (p x i x j) design.

Table 2 enumerates the sources of variation and presents the estimated variance components for the mathematics test.

Table 2

Generalizability of Measures of Mathematics Achievement

Source of Variation	Estimated Variance Components		
	$n_i'=1, n_j'=1$	$n_i'=10, n_j'=1$	$n_i'=10, n_j'=2$
Students (P)	7.55	7.55	7.55
Items (I)	1.73	.17	.17
Occasions (J)	.96	.96	.48
PI	5.42	.54	.54
PJ	.71	.71	.36
IJ	.50	.05	.02
Residual (PIJ,e)	4.88	.49	.25
$\hat{\sigma}^2_\delta$	11.01	1.74	1.15
G coefficient for relative decisions	.39	.81	.87
$\hat{\sigma}^2_\Delta$	14.20	2.92	1.82
G coefficient for absolute decisions	.35	.72	.81

The first column shows that three estimated variance components are large relative to the other components. The first, for students (σ^2_p) is analogous to true score variance in classical test theory and is expected to be large. The second, the student by item interaction (σ^2_{pi})

represents one source of measurement error and is due to the tendency of different items to rank students differently. The third is the residual term representing the three-way interaction between students, items, and occasions and unidentified sources of measurement error ($\sigma^2_{pij,e}$). The small components associated wth occasions (the J, PJ, IJ components) suggest that the occasion of testing introduces little variablility into the measurement of mathematics achievement. Average student performance over items is similar across occasions ($\hat{\sigma}^2_j$); students are ranked nearly the same across occasions ($\hat{\sigma}^2_{pj}$); and item means are ordered nearly the same across occasions ($\hat{\sigma}^2_{ij}$). The optimal D study design then, will include multiple test items but few ocasions.

Table 2 also gives estimated variance components, error, and generalizability coefficients for three D study designs: one item and one occasion, ten items and one occasion, and ten items and two occasions. Information is presented for both relative and absolute decisions. As described earlier, a relative decision might be to select the top 20 percent of the scorers for a special program. The variance components contibuting to error in this case include the components for all interactions with persons: PI, PJ, and PIJ,e. These are the only components that influence the rank ordering of students. An absolute decision might be to select all students who obtain a score of 75 percent correct or better. The error in this case consists of all components except that for students: I, J, PI, JP, IJ, and PIJ,e. All of these components influence students' absolute level of performance. As the estimates of error and generalizability coefficients in Table 2 indicate, administering a ten-item test on one occasion would substantially reduce error over a single item. Increasing the number of occasions to two would reduce error by only a small amount. The small reduction in error may not warrant the extra time and expense involved in administering the test twice.

Typically, several D study designs will yield the same level of generalizability. For a decision-maker who desires a generalizability coefficient (relative decision) of .87, for example, there are at least two D study designs to choose from. As indicated in Table 2, ten items administered on two occasions would be expected to produce this level of generalizability. Alternatively, 25 items administered on one occasion would also produce this result. The decision-maker must balance cost considerations to choose the appropriate D study design. When items are difficult and expensive to produce, the former design may be more practical. When items are fairly easy to generate (as is probably the case in tests of mathematics achievement), the latter design may be preferable.

Assumptions

Lack of restrictions. Before discussing the assumptions underlying the generalizability model and procedures, it is instructive to describe which assumptions and restrictions occurring in other measurement theories (for example, classical theory) are not held in generalizability theory. First, generalizability theory avoids the classical assumption of parallelism: equal means, variances and intercorrelations among conditions of a facet (for example, item scores). The lack of these assumptions has implications for the interpretation of the results of G and D studies. One cannot assume that conditions sampled within a facet are equivalent. For example, one cannot assume that items sampled for a study have the same means, variances and intercorrelations. Furthermore, conditions sampled across studies cannot be assumed to be equivalent. For example, the items selected for the G study may not have the same level of difficulty as those selected for the D study. Moreover, the items in one D study may not be equivalent to those selected for another D study. The differences among conditions and between sets of conditions may be due to characteristics of examinees as well as characteristics of items.

To deal with the difficulty that one set of conditions sampled in a D study (for example, items or occasions) may not be equivalent to each other or to another set, Cronbach et al. (1972) discuss an item-sampling design proposed by Lord and Novick (1968). In this plan, a large sample of persons is subdivided at random into three or more subsamples. In the G study, each subsample would be observed under the set of coditions to be sampled in the D study and one additional condition. The additional condition would be different for each subsample. Each subsample, then, would be observed under identical conditions plus one different condition. A comparison of the results (variance component estimates) across subsamples would reveal how well the set of conditions to be sampled in the D study represent the universe of conditions. If the results across subsamples are similar, then one can confidently generalize the results of the D study to the conditions in the universe of generalization. If the results are different across subsamples, one must be very cautious in generalizing beyond the condi ions (for example, items) sampled in the D study.

Second, the generalizability model makes no assumptions about the distributions underlying the measurements obtained in the G and D studies, or of the universe scores. Little is known, however, about the effects of different underlying distributions of scores on the estimates of variance components and the efficiencies of the estimators. It should be noted that generalizability theory does make assumptions about the distributions underlying variance component estimation (see next section).

Third, there is no restriction about the kinds of conditions that can be defined as facets. Any source of variation can be defined as a facet including, for example, test item, test form, item format, occasion of testing, and test administrator. Generalizability theory may be the only way to disentangle the effects of these sources of variation. Item-response models are not able to deal with the effects of administrator variation, for example.

Random sampling. One of the few assumptions of generalizability theory is random sampling of persons and conditions (for random facets). Although this assumption is considerably weaker than the assumption of classical theory that conditions are strictly parallel (equal means, variances, correlations), it has often raised objections from those who maintain that measurements rarely consist of random samples from well-defined universes of generalization (for example, Loevinger, 1965; Rozeboom, 1966; Gillmore, 1979). As Kane (1982, p. 30) points out, "The effects of unintended departures from the random sampling assumption cannot be evaluated accurately, and therefore the interpretation of G-study results must always be somewhat tentative."

Brennan (1981) sets a more optimistic tone by suggesting that the universe of generalization need not be undifferentiated (as, for example, a universe of test items), but may be structured such that the assumption of random sampling is more acceptable (for example, sampling from categories representing different item or content specifications).

Lord and Novick (1968, p. 235) also provide support for the random sampling assumption, which is relevant for generalizability theory:

> A possible objection to the item-sampling model (for example, see Loevinger, 1965) is that one does not ordinarily build tests by drawing items at random from a pool. There is, however, a similar and equally strong objection to classical test theory: Classical theory requires test forms that are strictly parallel, and yet no one has ever produced two strictly parallel forms for any ordinary paper-and-pencil test. Classical test theory is to be considered a useful idealization of situations encountered with actual mental tests. The assumption of random sampling of items may be considered in the same way. Further, even if the items of a particular test have not actually been drawn at random, we can still make certain interesting projections: We can conceive an item population from which the items of the test might have been randomly drawn and then consider the score the examinee would be expected to achieve over this population. The abundant information available on such expected scores enhances their natural interest to the examinee.

Infinite universe. Related to random sampling assumption described above is the assumption for random facets that the number of conditions in the universe of admissible conditions be indefinitely large. When the universe (of admissible observations or of generalization) is finite, the analysis and interpretation need to be adjusted, depending upon the relationships among the number of conditions sampled in the G study, the number of conditions in the universe of admissible observations, and the number of conditions in the universe of generalizaton. The universe of admissible observatins comprises all possible combinations of conditions represented in the G study. The universe of generalization consists of those combinations of conditions over which the decision-maker wishes to generalize. Although the two universes may be the same, the universe of generalization often will be smaller (fewer facets) than the universe of admissible observations. For example, a G study with items, test administrators, and occasions as facets may show little variability due to test administrators and occasions but substantial variability due to items. For the D study, then, the decision-maker may decide to use one test administrator and administer the test on only one occasion but use multiple items. The universe of admissible observations would have three facets; the universe of generalization would have one facet. Cronbach et al. (1972) consider several possibilities of finite universes and describe the implications for analysis. As Cronbach et al. point out, the intermediate cases in which a subset of a finite universe of conditions is sampled can be complex.

In most applications, the decision-maker's choice is between random sampling from an indefinitely large universe (random facet) or inclusion of all of a finite set of conditions (fixed facet). In the latter case, Shavelson and Webb (1981) recommend that the decision-maker examine the variablility of the conditions of the fixed facet. If the variability is small, the scores can be averaged over conditons of the fixed facet. When the variability is large, however, each condition should be treated separately or the scores should should be treated as a profile. Whenever there is a question about the magnitude of the variability, it may be most reasonable to present the results for each condition separately as well as the average over the conditions of the facet. This recommendation applies to the D study as well as to the G study.

Variance components. Generalizability theory assumes that the distributions underlying variance components are normal and that variance components cannot be negative. Analyses of non-normal distributions of variance components by Scheffe (1959; see Cronbach et al., 1972, p. 52) suggest that departures from normality can have a large effect on the "trustworthiness" of the confidence interval around the variance component.

Negative estimates of variance components can arise as a result of sampling variability or model misspecification. For example, a random-effect model may not be valid (Nelder, 1954). Cronbach et al. (1972) suggest that zero be substituted for negative estimates and substituted in any expected mean square equation containing that component. As Scheffe (1959) and others have pointed out, the zero estimates and modified estimates for other effects are biased. The greater the number of facets in the design (particularly for crossed designs), the greater the potential for a large number of biased estimates of variance components.

The problem of negative estimates of variance components is not insurmountable, however. Cronbach et al. (1972) suggest the use of a Bayesian approach, which not only provides a solution to the problem of negative estimates, but also provides estimates of variance components that are interpretable with respect to the sample data, not to repeated sampling. Fyans' (1977; see also Box & Tiao, 1973; Davis, 1974; Hill, 1965, 1967, 1970; Novick et al., 1971) strategy for obtaining Bayesian estimates constrains the estimates to be greater than or equal to zero. The resulting estimates are biased, however.

Limitations of the Procedures

The two major limitations of the procedures of generalizability theory to be discussed here are the need for extensive data for reliable estimates of variance components, and the difficulties of estimation in unbalanced designs. It should be noted that these limitations are not weaknesses in the theory but are difficulties arising in practice.

Sampling variability of estimated variance components. Since G-theory emphasizes the estimation and interpretation of variance components, their sampling variability is of great importance, albeit seldom addressed. Two issues arise: a comparison of sampling variability of variance components for different effects in a design, and the magnitude of sampling errors in studies with moderate numbers of observations.

Concerning the first issue, a comparison of sampling variances for different effects in a G-theory design suggests that the sampling estimates of the universe score variance may be less stable than estimates of components of error variance. This result derives from an inspection of general formulas for sampling variances of estimated variance components (see Smith, 1978). In fully crossed designs, at least, the formulas for sampling variability of estimated variance components for main effects contain more components, and (for moderate numbers of persons and conditions) can be expected to yield a larger sampling variance estimate, than the formulas for higher-order interaction effects. An illustration of this result for a two-facet, crossed (p x i x j), random model design comes from Smith (1978, Figure 1). The variance of the estimated variance component for persons (the universe score variance) is

$$\mathrm{var}(\hat{\sigma}^2_p) = \frac{2}{(n_p-1)}\left[(\sigma^2_p + \frac{\sigma^2_{pj}}{n_j} + \frac{\sigma^2_{pi}}{n_i} + \frac{\sigma^2_{res}}{n_i n_j})^2 + \frac{1}{(n_j-1)}(\frac{\sigma^2_{pj}}{n_i} + \frac{\sigma^2_{res}}{n_i n_j})^2 + \frac{1}{(n_i-1)}(\frac{\sigma^2_{pi}}{n_i} + \frac{\sigma^2_{res}}{n_i n_j})^2 + \frac{1}{(n_i-1)(n_j-1)}(\frac{\sigma^2_{res}}{n_i n_j})^2\right]$$

while the variance of the estimated component for the residual is

$$\mathrm{var}(\hat{\sigma}^2_{res}) = \frac{2}{(n_p-1)(n_i-1)(n_j-1)}\,\sigma^4_{res}\ .$$

In general, the sampling errors are expected to be greater for designs with greater numbers of facets than for designs with few facets, thus producing a trade-off between band width and fidelity.

The second issue concerns the magnitude of sampling errors of estimated variance components. Monte carlo simulations conducted by Smith (1978, 1980), Calkins et al. (1978), and Leone and Nelson (1966) for a variety of crossed and nested designs produced large sampling errors for small and moderate numbers of persons and conditions. Smith, for example, found that "(a) the sampling errors of variance components are much greater for multifaceted universes than for single faceted universes; (b) for $\hat{\sigma}^2_p$ the sampling errors were large unless the total number of observations $(n_p n_i n_j)$ was at least 800; (c) stable estimates of σ^2_i and σ^2_j required at least eight levels of each facet; and (d) some nested designs produced more stable estimates than did crossed designs" (Shavelson & Webb, 1981, p. 141). Smith's results pose a serious problem for the interpretation of results in the moderately sized designs typically used. The requirements of large numbers of conditions and large numbers of total observations for stable estimates of variance components are rarely met in most G and D studies.

Woodward and Joe (1973) and Smith (1978) recommended that measurements be allocated in the D study in specific ways to minimize sampling variability. For example, in a p x i x j design, they recommended using equal numbers of conditions of facets i and j when $\hat{\sigma}^2_{res}$ increases relative to $\hat{\sigma}^2_{pi}$ and $\hat{\sigma}^2_{pj}$, and making the numbers of conditions of facets i and j proportional to $\hat{\sigma}^2_{pi}/\hat{\sigma}^2_{pj}$

when $\hat{\sigma}^2_{res}$ decreases relative to $\hat{\sigma}^2_{pi}$ and $\hat{\sigma}^2_{pj}$. These decisions are based on the results of the G study.

To deal with the requirement of large numbers of observations, Smith (1980) also proposed the use of several small G studies with many conditions of a few facets, each estimating part of a complex G study, instead of one large G study with a few conditions of many facets. As Shavelson and Webb (1981) point out, however, there is a question of how well the restricted universes of the several small G studies represent the universe of the single, large G study.

Unbalanced designs. A major difficulty with the ANOVA approach to estimating variance components arises in unbalanced designs, in which there are unequal numbers of observations in its subclassifications. An example which occurs in many tests is an unequal number of items across subtests. Another example is students nested within classes where class size varies. The primary difficulty with unbalanced data is computational complexity. The usual rules for deriving expected values of mean squares (Cornfield & Tukey, 1956) do not apply to unbalanced designs. Although computer programs have been developed to estimate variance components in unbalanced designs, they require large storage capacities and, therefore, may be prohibitively expensive in many cases. (For descriptions of the computer programs, see Brennan et al., 1980; Llabre, 1978, 1980; Rao, 1971, 1972.)

Strengths and Weaknesses of the Model

The major strength of generalizability theory is its ability to assess sources of error in the measurement and, consequently, to design optimal decision-making studies. This ability affects not only a specific decision-maker's study but, as Cronbach et al. (1972, p. 384) point out, it can help evaluate existing testing practices:

> Application of generalizability theory should operate ultimately to increase the accuracy of test interpretations. It will make interpretation more cautious as the inadequate generalizability of a procedure becomes recognized, and it will encourage the development of procedures more suitable for generalized interpretation.

The weak assumptions afford the decision-maker great flexibility in designing generalizability and decision studies, and in defining relevant universes of interest. At the same time, however, the lack of assumptions leaves several questions unanswered. One is the lack of guidelines about the reasonableness of data. For example, the effects of outliers or influential observations on the estimates are not well known.

5

Present Areas of Application

Reliability. As was described in the first section of this paper, a primary goal of G-theory is to design measurement procedures that minimize error variability, and thereby maximize reliability, while at the same time allowing the decision-maker to generalize over a broad range of testing situations. Generalizability theory has been applied to a variety of areas in the behavioral sciences to study the dependability of measures of the behavior of schizophrenic patients (e.g., Mariotto & Farrell, 1979), assertion in the elderly (Edinberg et al., 1977), free-recall in children (Peng & Farr, 1976), depth and duration of sleep (Coates et al., 1979), behavior of teachers (Erlich

& Shavelson, 1978), dentists' sensitivity toward patients (Gershen, 1976), educational attainment (Cardinet et al., 1976), job satisfaction using Spanish and English forms (Katerberg et al., 1977), student ratings of instruction (Gillmore et al., 1978), and heterosexual social anxiety (Farrell et al., 1979).

Linked conditions and multivariate estimation. Educational and psychological measurements often provide multiple scores which may be interpreted as profiles (for example, patterns of scores on the Wechsler Intelligence Scale for Children are used to place students in special education programs) or composites (for example, the Comprehensive Test of Basic Skills). Although the most common procedures used to assess reliability focus on the separate scores or on the composite, neither method assesses the linkage or error covariation among the multiple scores. For example, subtest scores from the same test battery are "linked" by virtue of occurring on the same test form and on the same occasion. Information about the covariation among scores is important for designing an optimal D study, and permitting the decision-maker to determine the composite with maximum generalizability. For these purposes, a multivariate analysis is more appropriate (see Cronbach et al., 1972; Shavelson & Webb, 1981; Travers, 1969; Webb & Shavelson, 1981).

In extending G-theory's notion of multifaceted error variance to multivariate designs, subtest scores, for example, would be treated not as a facet of measurement but as a vector of outcome scores. While univariate G-theory focuses on variance components, multivariate G-theory focuses on matrices of variance and covariance components.

The matrix of variances and covariances among observed scores is decomposed into matrices of components of variance and covariance. The expected mean square and cross-product equations from a multivariate analysis of variance are solved in analogous fashion to their univariate counterparts. For example, the decompositio of the variance-covariance matrix of observed scores in a one-facet, crossed design with two dependent variables (for example, the grammar and paragraph comprehension subtests in a language arts battery) is:

$$\underset{\text{(observed scores)}}{\begin{bmatrix} \sigma^2({}_1X_{pi}) & \sigma({}_1X_{pi},{}_2X_{pg}) \\ \sigma({}_1X_{pi},{}_2X_{pg}) & \sigma^2({}_2X_{pg}) \end{bmatrix}} = \underset{\text{(persons)}}{\begin{bmatrix} \sigma^2({}_1p) & \sigma({}_1p,{}_2p) \\ \sigma({}_1p,{}_2p) & \sigma^2({}_2p) \end{bmatrix}}$$

$$+ \underset{\text{(conditions)}}{\begin{bmatrix} \sigma^2({}_1i) & \sigma({}_1i,{}_2g) \\ \sigma({}_1i,{}_2g) & \sigma^2({}_2g) \end{bmatrix}}$$

$$+ \underset{\text{(residual)}}{\begin{bmatrix} \sigma^2({}_1pi,e) & \sigma({}_1pi,e,{}_2pg,e) \\ \sigma({}_1pi,e,{}_2pg,e) & \sigma^2({}_2pg,e) \end{bmatrix}}$$

where $_1X_{pi}$ = score on variable 1 for person p observed under condition i,

$_2X_{pg}$ = score on variable 2 for person p observed under condition g, and

$_1p$ = abbreviated for $_1\mu_p$: the universe score on variable 1 for person p.

In the above equation, the term $\sigma(_1p,_2p)$ is the covariance between universe scores on variables 1 and 2 (grammar and paragraph comprehension). The term $\sigma(_1i,_2g)$ is the covariance between scores on the two variables due to the condition of observation. Facet i may be the same as facet g, for example, when the grammar and paragraph comprehension scores are obtained from the same test form (on the same occasion). The term $\sigma(_1pi,e;_2pg,e)$ is the covariance due to unsystematic error.

The matrices of variance and covariance components provide essential information for deciding whether multiple scores in a battery should be treated as a profile or a composite as opposed to separate scores. The matrix of covariance components for universe scores particularly shows whether it is reasonable to consider the scores as representing an underlying dimension, in which case a profile or a composite are reasonable. Small covariance components relative to the variance components suggest that the scores are not related and that a composite of the scores would not be interpretable.

Although the components of variance and covariance are of primary importance and interest, a decision-maker may find it useful to obtain the dimensions of scores (composites) with maximum generalizability. The multivariate extension of the univariate generalizability coefficient was developed by Joe and Woodward (1976). From a random effects multivariate analysis of variance, the canonical variates are determined to maximize the ratio of universe-score variation to univers-score plus error variation. For the two-facet fully crossed design, Joe and Woodward's multivariate coefficient for relative decision is

$$\rho 2 = \frac{\underline{a}'V_p\underline{a}}{\underline{a}'V_p\underline{a} + \frac{\underline{a}'V_{pi}\underline{a}}{n_i'} + \frac{\underline{a}'V_{pj}\underline{a}}{n_j'} + \frac{\underline{a}'V_e\underline{a}}{n_i'n_j'}}$$

where $\underline{V}$ = a matrix of variance and covariance components estimated from mean square matrices,

n_i' and n_j' = the number of conditions of facets i and j in a D study, and

$\underline{a}$ = the vector of canonical coefficients that maximizes the ratio of between-person to between-person plus within-person variance component matrices.

There is a set of canonical coefficients ($\underline{a}_s$) for each characteristic root in the above equation. Each set of canonical coefficients defines a composite of scores. By definition, the first composite is the most reliable while the last composite is the least reliable. This procedure, then, produces the most generalizable composite of subtest scores, for example, that takes into account the linkage among the scores.

An application of multivariate generalizability theory to arithmetic achievement (reported in Webb, Shavelson, & Maddahian, 1982) will be used as an illustration. Three subtests representing basic computational skills (addition/subtraction, multiplication, and division) were selected from the mathematics battery at grade five from the Beginning Teacher Evaluation Study (BTES), a research program designed to identify effectie teaching behavior in elementary school reading and mathematics. A sample of 127 students completed the three mathematics subtests on two occasions. The design of the multivariate study, then, had one facet (occasions) crossed with persons.

Table 3 presents the matrices of components of variance and covariance for the three effects in the design: persons, occasions, and the residual. The subtantial components of covariance for persons (which is the universe-score component matrix) shows that the three subtests are substantially related and that it is reasonable to form a composite of the scores. The non-zero components of covariance for the residual show that the tendency for students to be ranked ordered differently across occasions (interaction between persons and occasions) is consistent across subtests.

The dimensions of mathematical skill that have maximum generalizability are presented in Table 4. When the generalizability of mathematics scores was estimated for a single occasion, one dimension with generalizability coefficient exceeding .60 emerged from the analysis. This dimension is a general composite heavily weighted by division. The analysis with two occasions produced two dimensions with generalizability coefficients exceeding .60. The first is the general composite described above; the second is a contrast between addition/subtraction and division.

Table 3

Estimated Variance and Covariance Components for Multivariate Generalizability Study of Basic Skills ($\underline{n}_o$=1)

Source of Variation		Addition/Subtraction (1)	Multiplication (2)	Division (3)
Persons (P)	(1)	2.27		
	(2)	2.08	5.64	
	(3)	1.07	2.41	3.60
Occasions(0)	(1)	.00		
	(2)	-.12	1.27	
	(3)	-.04	.49	.17
P0,e	(1)	2.34		
	(2)	.84	5.84	
	(3)	.00	.28	1.74

Table 4

Canonical Variates for Multivariate Generalizability Study of Basic Skills

	$\underline{n}_o$ = 1			$\underline{n}_o$ = 2		
	I	II	III	I	II	III
(1) Addition/Subtraction	.11	-.36	-.34	.11	-.42	-.42
(2) Multiplication	.07	-.11	.31	.07	-.13	.38
(3) Division	.35	.28	-.12	.37	.33	-.15
Coefficient of Generalizability ($\hat{\rho}^2$)	.71	.44	.33	.83	.61	.50

New Areas of Application

This section includes areas that have been developed but rarely applied in practice, including test design and estimation of universe scores and profiles, as well as areas that need to be developed, including estimation of phenomena that change over time and the effects of underlying score distributions on estimation and sampling variability of estimators.

Test design. Generalizability theory can be used in designing tests: for example, providing information on variability among subtests, items within subtests, and item formats. Any of these characteristics of tests can be used to define the universes of admissible observations and generalization and can be included as facets in G and D studies. Complexly structured tests can even be considered, as in the case of unequal numbers of items for different subtests in a test battery. A straightforward way to deal with this case is to consider subtest as fixed, and to perform separate G analyses (with items as a facet) for each subtest. Conditions of the testing situation, as opposed to the test itself can also be taken into account, such as occasion, examiner, and scorer.

Estimation of universe scores and profiles. A contribution of generalizability theory is the estimation of point estimates of universe scores and of score profiles. Cronbach et al. (1972, p. 103) present an estimation equation (based on Kelley, 1947) for a point estimate of the universe score which is shown to be more reliable than observed scores:

$$\hat{\mu}_p = (\hat{\xi}\rho^2)\ X_{pI} + (1-\hat{\xi}\rho^2)\ X_{PI}$$

Although this procedure could be repeated for each subtest in a test battery, thus producing a universe score profile, it would not take full advantage of the relationships among the subtests.

Cronbach et al. (1972, p. 313-314) show how the correlations among variables in a test battery can be taken into account to produce a more dependable profile of universe scores. Basically, the regression equation for a particular score in the profile includes not only the observed scores on that variable (as in the above equation) but also the observed scores for all other scores in the set. The set of multiple regression equation equations produces a profile of estimated universe scores for each person. This profile is more reliable (and usually flatter) than that based on univariate regression equations. In an example using data from the Differential Aptitude Tests (DAT), Cronbach et al. (1972) reported reductions in error variance as large as 42 percent when all subtests were used as predictors compared to error variances from single predictors. Such universe score profiles are useful for guidance decisions and diagnostic purposes. It is important to note further that the regression methods outlined here may produce not only flatter profiles than observed scores, but sometimes will invert relationships in an observed-score profile. The important implication for counseling and research is that observed profiles and those estimated from univariate regressions may be much further from the true profiles than multivariate estimates.

Changing phenomena vs. steady state phenomena. All of the discussion thus far has assumed that the phenomenon being studied remains constant over observations. The problem is very complex, however, when the universe score changes over time, as is the case in maturation studies (e.g., Bayley, 1968). This problem is particularly acute in testing situations which assume no change in true ability or knowledge across testing situations but in which sufficient time elapses that true changes do appear. A further complication is that the growth patterns of different individuals over time may not be equivalent. A few inroads into this area are the work of Bryk (1980) and Maddahian (1982).

Underlying score distributions. The lack of knowledge about the impact of varying underlying score distributions on the estimation and sampling variability of univariate parameters, including universe score estimates, variance components, and generalizability coefficients, and multivariate parameters, including universe score profile estimation, components of covariance, multivariate generalizability coefficients, and canonical coefficients, clearly presents an area in need of development. Issues needing to be addressed include bias and efficiency of the estimators.

REFERENCES

Bayley, N. Behavioral correlates of mental growth: Birth to thirty-six years. American Psychologist, 1968, 23, 1-17.

Box, G.E.P. & Tiao, G.C. Bayesian Inference in Statistical Analysis. Reading, MA: Addison-Wesley, 1973.

Brennan, R.L. Generalizability analyses: Principles and procedures. ACT Tech. Bulletin, No. 26, Iowa City, IO: American College Testing Program, September, 1977a.

Brennan, R.L. Handbook for Gapid: A Fortaran IV computer program for generalizability analyses with single vacet designs. ACT Tech. Report No. 34, Iowa City, IO: American College Testing Program, October, 1979a.

Brennan, R.L., Jarjoura, D., & Deaton, E.L. Interpreting and estimating variance components in generalizability theory: An overview. Paper presented at the Annual Meeting of the American Educational Research Association, Boston, April 1980.

Brennan, R.L. & Kane, M.T.Generalizability theory: A review of basic concepts, issues and procedures. In R.E. Traub (Ed.), New Directions in Testing and Mesurement. San Francisco: Jossey-Bass, 1979.

Bryk, A.S., Strenio, J.F., & Weisberg, H.I. A method for estimating treatment effects when individuals are growing. Journal of Educatinal Statistics, 1980, 5, 5-34.

Calkins, D.S., Erlich, O., Marston, P.T., & Malitz, D. An empirical investigation of the distributions of generalizability coefficients and various estimates for an application of generalizability theory. Paper presented at the Annual Meeting of the American Educational Research Association, Toronto, March 1978.

Cardinet, J. & Tourneur, Y. Le calcul de marges d'erreurs dans la theorie de la generalizabilite. Neuchatel (Suisse): Institut Romand de Recherches et de Documentation Pedagogiques, 1978.

Cardinet, J., Tourneur, Y., & Allal, L. The generalizability of surveys of educational outcomes. In D.N.M. deGruijter & L.J.Th. van der Kamp (Eds.), Advances in Psychological and Educational Measurement, New York: Wiley, pp. 185-198, 1976a.

Cardinet, J., Tourneur, Y. & Allal, L. The symmetry of generalizability theory: Applications to educational measurement. Journal of Educational Measurement, 1976b, 13, 119-135.

Coates, T.J., Rosekind, M.R., Strossen, R.J., Thoresen, C.E., & Kirmil-Gray, K. Sleep recordings in the laboratory and home: A compaative analysis. Psychophysiology, 1979, 16, 339-347.

Cornfield, J. & Tukey, J.W. Average values of mean squares in factorials. Annals of Mathematical Statistics, 1956, 27, 907-949.

Cronbach, L.J., Gleser, G.C., Nanda, H., & Rajaratnam, N. The Dependability of Behavioral Measurements. New York: Wiley, 1972.

Davis, C. Bayesian inference in two way models: An approach to generalizability. Unpublished doctoral dissertation, University of Iowa, 1974.

Edinberg, M.A., Karoly, P., & Gleser, G.C. Assessing assertion in the elderly: An application of the behavioral-analytic model of competence. Journal of Clinical Psychology, 1977, 33, 869-874.

Erlich, O. & Shavelson, R. The application of generalizability theory to the study of teaching. Tech. Report 76-9-1, Beginning Teacher Evaluation Study, Far West Laboratory, September 1976b.

Erlich, O. & Shavelson, R.J. The search for correlations between measures of teacher behavior and student achievement: Measurement problem, conceptualization problem or both? Journal of Educational Measurement, 1978, 15, 77-89.

Farrell, A.D., Marco, J.J., Conger, A.J., & Wallander, J.L. Self-ratings and judges ratings of heterosexual social anxiety: A generalizability study. Journal of Consulting and Clinical Psychology, 1979, 47, 164-175.

Fyans, L.J., Jr. A new multiple level approach to cross-cultural psychological research. Unpublished doctoral dissertation, University of Illinois at Urbana-Champaign, 1977.

Gillmore, G.M. An Introduction to Generalizability Theory as a Contributor to Evaluation Research. Seattle, WA: Educational Assessment Center, University of Washington, 1979.

Gillmore, G.M., Kane, M.T., & Naccarato, R.W. The generalizability of student ratings of instruction. Journal of Educational Measurement, 1978, 15, 1-15.

Hill, B.M. Inference about variance components in the one-way model. Journal of the American Statistical Association, 1965, 60, 806-825.

Hill, B.M. correlated errors in the random model. Journal of the American Statistical Association, 1967, 62, 1387-1400.

Hill, B.M. Some contrasts between Bayesian and classical influence in the analysis of variance and in the testing of models. In D.L. Meyer & R.O Collier, Jr. (Eds.) Bayesian Statistics. Itasca, IL: F.E. Peacock, 1970.

Huysamen, G.K. Psychological Test Theory. Durbanville, South Africa: Uitgewery Bouschendal Distributor, 1980.

Joe, G.N. & Woodward, J.A. Some developments in multivariate generalizability. Psychometrika, 1976, 41, 205-217.

Katerberg, R., Smith, F.J., & Hoy, S. Language time and person effects on attitude scale translations. Journal of Applied Psychology, 1977, 62, 385-391.

Kelley, T.L. Fundamentals of Statistics. Cambridge, MA: Harvard University Press, 1947.

Leone, F.C. & Nelson, L.S. Sampling distributions of variance components--I. Empirical studies of balanced nested design. Technometrics, 1966, 8, 457-468.

Llabre, M.M. An application of generalizability theory to the assessment of writing ability. Unpublished doctoral dissertation, University of Florida, 1978.

Llabre, M.M. Estimating variance components with unbalanced designs in generalizability theory. Paper presented at the Annual Meeting of the American Educational Research Association, Boston, April 1980.

Loevinger, J. Person and population as psychometric concepts. Psychological Review, 1965, 72, 143-155.

Lord, F.M. & Novick, M. Statistical Theories of Mental Test Scores. Addison-Wesley, 1968.

Mariotto, M.J. & Farrell, A.D. Comparability of the absolute level of ratings on the inpatient multidimensional psychiatric scale within a homogeneous group of raters. Journal of Consulting and Clinical Psychology, 1979, 47, 59-64.

Nelder, J.A. The interpretation of negative components of variance. Biometrika, 1954, 41, 554-558.

Novick, M.R., Jackson, P.H., &n Thayer, D.T. Bayesian inference and the classical test theory mdel: Reliability and true scores. Psychometrika, 1971, 36, 261-288.

Peng, S.S. & Farr, S.D. Generalizability of free-recall measurements. Multivariate Behavioral Research, 1976, 11, 287-296.

Rao, C.R. Minimum variance quadratic unbiased estimation of variance components. Journal of Multivariate Analysis, 1971, 1, 445-456.

Rao, C.R. Estimation of variance and covariance components in linear models. Journal of the American Statistical Association, 1972, 67, 112-115.

Rozeboom, W.W. Foundations of the Theory of Prediction. Dorsey Press, Homewood, Ill., 1966.

Scheffe, H. The Analysis of Variance, New York: Wiley, 1959.

Shavelson, R.J. & Webb, N.M. Generalizability theory: 1973-1980. British Journal of Mathematical and Statistical Psychology, 1981, 34, 133-166.

Smith, P. Sampling errors of variance components in small sample multifacet generalizability studies. Journal of Educational Statistics, 1978, 3, 319-346.

Smith, P.L. Some approaches to determining the stability of estimated variance components. Paper presented at teh Annual Meeting of the American Educational Research Association, Boston, April 1980.

Tourneur, Y. Les objectifs du domaine cognitif, 2me partie--theorie des tests. Ministere de l'Education Nationale et de la Culture Francaise, Universite de l'Etat a mons, Faculte des Sciences Psycho-Pedagogiques, 1978.

Tourneur, Y. & Cardinet, J. Analyse de variance et theorie de la generalizabilite: Guide pour la realisation des calculs (Doc. 790.803/CT/9). Universite de l'Etat a Mons, 1979.

Travers, K.J. Correction for attenuation: A generalizability approach using components of covariance. Unpublished manuscript. University of Illinois, 1969.

van der Kamp, L.I.Th. Generalizability and educational measurement. In D.N.M. deGruijter & L.J.Th. van der Kamp (Eds.) Advances in Psychological and Educational Measurement. New York: Wiley, 1976.

Webb, N.M. & Shavelson, R.J. Multivariate generalizability of general educational development ratings. Journal of Educational Measurement, 1981, 18, 13-22.

Webb, N.M., Shavelson, R.J., & Maddahian, E. Multivariate generalizability theory. New Directions in Testing and Measurement: Generalizability Theory, in press, 1982.

Wiggins, J.S. Peresonality and Prediction: Principles of Personality Assessment. Reading, MA: Addison-Wesley, 1973.

Woodward, J.A. & Joe, G.W. Maximizing the coefficient of generalizability in multi-facet decision studies. Psychometrika, 1973, 38, 173-181.

ANALYSIS OF READING COMPREHENSION DATA*

The data set used in this analysis is taken from the 1971 survey of reading achievement in the United States carried out in conjunction with the International Association for Educational Achievement's Study of Reading Comprehension in 15 Countries (Thorndike, 1973). The total sample consisted of 5,479 fourth grade students drawn from a probability sample of 239 schools scattered across the United States (Wolf, 1977). Each of the selected students was asked to complete a variety of tests and questionnaires designed to establish the relative influence of various external factors to the development of reading achievement and an interest in reading.

The international research program called for the administration of essentially the same tests (though translated into different languages) to comparable samples of students in each country. The "between country" variation in background factors, school organization, parental expectation and involvement, cultural importance of written communication, etc., offered a unique opportunity to use the natural laboratory to investigate their respective influences. It was necessary in such a research study, however, to develop the me··-urement instruments with great care. They not only had to be of high psychometric quality, but also had to be capable of translation into a range of languages so as to yield comparable, relevant, and fair measures of achievement in all the participating countries. For this

* This chapter was compiled by David McArthur from contributions by Bruce Choppin, David McArthur, Raymond Moy, and Noreen Webb.

reason, the tests do not appear "familiar" in content or style to those regularly in use in any one country, but they were judged to be accessible enough to the average student in each country to yield an appropriately valid measure of achievement.

Two separate reading comprehension tests were administered. Each consisted of short reading passages of between 100 and 200 words, followed by a group of multiple-choice questions the answers to which could be found in the passage. The first section consisted of four reading passages and a total of 21 items. The second section had five reading passages and 24 items. Treated together for this analysis, they yield a multiple-choice test of reading comprehension containing 45 items (these are listed in Appendix I).

In order to perform a fair comparison of the different mathematical models for measuring achievement, it was decided to limit the analysis to samples of 1,000 students drawn from the master set. As a back-up and to estimate the stability of the parameters obtained, some analyses were repeated on a second, non-overlapping, sample of 1,000 students. Four approaches were applied to the 45 items of the Reading Comprehension Test for these samples of 1,000 cases: S-P analysis, Rasch analysis, Generalizability analysis, & 3 parameter latent trait analysis. Each is taken in turn below.

S-P Analysis

The S-P technique produced item p-values, person total scores, caution indices for both items and persons, the pair of curves (S &

P), the overall index of ordering and agreement with a perfect Guttman scale, and rank positions for both items and persons.

Average difficulty is p=0.532 with a range of 0.864 to 0.167 (of which the three most difficult items are answered correctly no better than chance). D*, the indicator of hypothetical misfit, is 0.506, a fairly high value. The average caution index for items (C_j*) is 0.250, ranging from 0.101 to 0.395. Eight of the items have caution indices exceeding 0.333.

In decreasing order of severity, these are items 16, 39, 31, 20, 43, 44, 7, and 42. The range of caution indices (C_i*) for respondents is from 0.038 to 0.730, with only three persons achieving below 0.050 but twenty-seven achieving above 0.500. There is a strong negative correlation (r= -0.45) between the item difficulties and their caution indices. According to this solution, the test appears to contain a moderate number of items poorly suited to this sample. Many correct responses are likely to be the result of chance guessing, and fully one-fifth of the items are exceptionally poor at discriminating between ability levels.

When those items with the highest caution indices are dropped altogether from the S-P analysis, the entire matrix and all associated indices for the items that remain and for all of the respondents are recalculated. While the truncated test on average is less difficult, there is little comparable decrease in the overall index of misfit. The number of respondents with elevated caution indices is exactly twice that of the first analysis, with the interesting finding that a proportion of that increase is to be found in the top-scoring 10% of the sample. It seems that when some items are removed because evidence shows that responses to them are generally not in correspondence with student ability, the S-P approach then penalizes some of the upper-ability students. This occurs when a student manages to get most of the included items correct, and most of the excluded items wrong, but also had one or two additional wrong ansers. In the analysis of the full set of items, those last one or two wrong answers do not cause the caution index to be all that out of line, but in the truncated set, those wrong answers can contribute heavily. For those students at the opposite end of the ability scale, both the first and second analyses show a sizeable number of high caution indices and very few low caution indices.

The low ability students are not measured well by this test, according to the S-P analysis, and generally there is an unanticipated large number of wrong answers by those whose overall ability level would have led one to expect success. The same findings proved true when the second sample of 1000 cases was analyzed, and also were obtained when the two sections comprising the 45 item test were analyzed separately.

Rasch Model Analysis

Computations using the same data set made by a Rasch model item analysis are as follows. For the complete set of 45 items that make up the two tests, the range of item difficulty is 18 wits (or about 4 logits). This a fairly is typical value for a classroom achievement test (which of course this was not!). The test was constructed to meet the needs of an international project and was designed to be effective in a broad spectrum of some 20 countries. As a result it appears not to be matched exactly to this particular sample of students in the USA. Although the easiest item in the test would have been "difficult" for fewer than one percent of the sample, the most difficult item (number 31) would have appeared quite easy to about 25 percent. For this particular group of students, the test could theoretically have been improved by the inclusion of one or two more difficult items.

In general the fit to the Rasch model was quite good. The worst fitting items were (in order of misfit) 16, 39, 43, 20, 31, 7, and 44. These are all comparatively difficult items. The analysis was repeated eliminating these items (and item 32) and the overall fit improved considerably. However, it should be stressed that only items 39 and 16 were sufficiently poor to the rejected by the usual Rasch item analysis criteria for fit.

It would appear that the inclusion of more difficult items as suggested ten lines above, would likely not have improved the test overall because of misfit due to guessing. the analysis emphasizes the seriousness of guessing on a four-way multiple-choice test.

There was a clear tendency for item discrimination to be related to item difficulty. The easiest items on the test discriminated well and the harder items comparatively poorly. All the misfitting items were among the poor discriminators. When the analysis was repeated omitting the eight poorest fitting items, the trend linking discrimination to difficulty remained. Even though the most difficult items on this test are not really very difficult for most of the sample of students, it would appear that guessing was very widespread. This would account for the overall relationship between difficulty and discrimination. An index of item discrimination deduced from the measure of misfit to the Rasch model correlated 0.967 with Sato's Caution Index suggesting that these two are measuring essentially the same thing (fit to a Guttman model).

To check the stability of the estimation of item difficulty the analysis run on the first 1000 cases in the data set and reported above was repeated on the second 1000. The results showed a high degree of stability. The conventional p-values of the items on the two separate samples of students correlated 0.982, while the delta values resulting from the Rasch scaling analysis correlated 0.984.

Each of the two sections of the test was composed of four clusters of items each relating to a short reading passage. These clusters vary little among themselves in terms of item characteristics although it may be noted that the first passage in each section (Tailor birds and Insects) are easier than those that follow them, and the final cluster on the record section (Musk Ox) is somewhat less discriminating than the average.

A check was made to see if the items operated differently for boys and girls. In general no major discrepancies were discovered although a few differences in individual item difficulty did reach significance. For example, items 7, 12, 24, 27, and 35 were relatively easier for the girls while items 16, 32, 33, and 44 were significantly easier for the boys. When the clusters were examined further small, but significant, trends were noted. The passages about "seals" and "the poet" were somewhat easier for the girls, while the passage about "eskimos" slightly favored the boys.

Generalizability Analysis

Generalizability analyses were performed to assess the magnitude of the sources of variation in the data set. The sources of variation include sex, persons, sections (first vs. second), passages (coded E in the tables), and items. The variation for persons is considered here to be the universe score variance (true score variance). All of the other sources of variation are considered error. For all of the analyses except that which includes sex, five items were selected at random from each passage to make a balanced design. For the analysis of sex, an equal number of boys an girls was selected.

Four designs of the basic data set were analyzed:

(1) Persons x Sections x Passages (Sections) x Items (Passages(Sections))

(2) Persons x Sections x Passages x Items (Passages)

(This design assumes that the same passages appeared in both sections and is probably not defensible. It was included to help disentangle the passage x section interaction in design (1).)

(3) Persons x Sections x Items (Sections)

(This design ignores passage as a source of variation.)

(4) Persons x Sections x Items

(This design assumes that each section has the same items and is probably not defensible. It was included to help disentangle the item x section interaction in the above design.)

An additional design was included to assess the effects of sex:

(5) Sex x Persons(Sex) x Sections x Passage(Sections) x Items(Passages(Sections)).

(This analysis is essentially the same design (1) with the additional stratification by sex.)

Table 1 gives the variance cc nponents for the five designs. These variance components are estimates for one section, one essay, and one item. The variance component for sections is zero, indicating that students performed equally well on both sections of the test. The persons x sections (PS) interaction is also low, indicating that students are ranked equally on both sections of the test.

In two sections, passages and items have nontrivial variation, even if low. Some passages are easier than other passages and some items are easier than other items. The variance components relating to items are the highest. Further, there is some tendency for items to rank students differently. To the extent that the section x item interaction can be interpreted, the position of item difficulties within one sect.on does not correspond to the other section. In other words, while the early items in the first section may be the easiest in that section, the early items in the second section may not be the easiest items in that section.

The large residual component in all designs suggests that there may be other sources of variation in test scores that have not been accounted for in the above designs.

Table 2 gives the generalizability coefficients for a variety of decision study designs. The coefficients were computed for absolute decisions: taking into account the absolute level of performance as well as relative rankings among students. All sources of variation other than that for persons, therefore, contribute to error. These G coefficients are considerably lower than those for relative decisions which include only the sources of variation interacting with persons (e.g., PS, PE(S), etc.).

The G coefficients for designs (1) and (2) are similar, as are those for designs (3) and (4). Increasing the number of items within each essay beyond 3 or 4 items has little impact on reliability, particularly, particularly when there are several passages in a section. Further, the total number of items seems to have the most impact of reliability; it does not matter how they are distributed

Table 1

Variance Components from Generalizability Analyses

P x S x E(S) x I(E(S))

Source	2	%
P	.031	12.4
S	.000	0.0
E(S)	.006	2.4
I(SE)	.022	8.8
PS	.000	0.0
PE(S)	.005	2.0
PI(SE),e	.187	74.5

P x S x E x I(E))

Source	2	%
P	.031	12.4
S	.000	0.0
E	.005	2.0
I(E)	.007	2.8
PS	.000	0.0
PE	.000	0.0
SE	.001	.4
PI(E)	.004	1.6
SI(E)	.015	6.0
PSE	.005	2.0
PSI(E),e	.182	72.8

P x S x I(S)

Source	2	%
P	.031	12.4
S	.000	0.0
I(S)	.027	10.8
PS	.001	0.4
PI(S),e	.191	76.4

P x S x I

Source	2	%
P	.031	12.4
S	.000	0.0
I	.011	4.4
PS	.001	0.4
PI	.005	2.0
SI	.016	6.4
PSI	.186	74.4

X x P(X) x S x E(S) x I(E(S))

Source	2	%
X	.000	0.0
S	.000	0.0
P(X)	.031	12.4
E(S)	.007	2.8
XS	.000	0.0
I(SE)	.022	8.8
PS(X)	.000	0.0
XE(S)	.000	0.0
PE(XS)	.005	2.0
XI(SE)	.000	0.0
PI(XSE),e	.186	74.4

P = Persons
X = Sex
S = Section (First vs. Second)
E = Passage
I = Item

Table 2

Generalizability Coefficients for Absolute Decisions

P x S x E(S) x I(ES)

No. Of Sections = 1

# of Passages	2	3	4
# of Items			
2	.34827	.44488	.51653
3	.43304	.53389	.60425
4	.49306	.59324	.66032
5	.53777	.63563	.69926

No. Of Sections = 2

# of Passages	2	3	4
# of Items			
2	.51661	.61580	.68120
3	.60437	.69613	.75331
4	.66046	.74469	.79541
5	.69941	.77723	.82301

P x S x I(S)

No. Of Sections = 1

# of Passages	2	3	4
# of Items			
2	.35798	.45310	.52251
3	.45310	.55063	.61704
4	.52251	.61704	.67841
5	.57540	.66518	.72146

No. Of Sections = 2

# of Passages	2	3	4
# of Items			
2	.52723	.62363	.68638
3	.62363	.71020	.76317
4	.68638	.76317	.80839
5	.73048	.79892	.83819

P x S x E x I(E)

No. Of Sections = 1

# of Passages	2	3	4
# of Items			
2	.34496	.44056	.51142
3	.42885	.52860	.59816
4	.44821	.58728	.65359
5	.53243	.62918	.69206

No. Of Sections = 2

# of Passages	2	3	4
# of Items			
2	.49016	.58983	.65659
3	.57439	.66848	.72811
4	.62839	.66848	.72811
5	.66595	.74829	.79761

P x S x I

No. Of Sections = 1

# of Passages	2	3	4
# of Items			
2	.35393	.44750	.51567
3	.44750	.54325	.60833
4	.51567	.60833	.66838
5	.56754	.65544	.71046

No. Of Sections = 2

# of Passages	2	3	4
# of Items			
2	.50586	.60239	.66592
3	.60239	.69019	.74444
4	.66599	.74444	.79107
5	.71091	.78128	.82197

across passages. For example, four passages with two items each has about the same reliability as two passages with four items each. The same result holds for sections; it does not matter how items are distributed across sections. For example, in design (1), one section with four passages with two items each has a G coefficient of .52; one section with two passages with four items each has a G coefficient of .52. All of the above combinations have eight items total. Similar combinations with a total of 16 items have G coefficient ranging from .66 to .68.

The final analysis examined sex as a source of variation. The component for sex was zero, indicating that boys and girls showed equal mean performance. Furthermore, the inclusion of sex did not affect any other component. In other words, items, passages and sections ranked boys and girls similarly. This finding seems to conflict somewhat with the finding in the Rasch analysis that some items ranked boys and girls differently.

Three-Parameter Latent Trait Analysis

With the introduction of an improved version of the LOGIST computer program for estimating the parameters in latent trait models, its use for examining test behavior is likely to become more widesp. ead. However, a problem remains in the evaluation of the results, as the parameters derived by the program are likely to be unstable. The problem is to identify the sources of instability and to assess their relative effects on the parameter estimates. The three sources of instability are:

1) Non-unidimensionality of the item responses,

2) Mis-specification of the item response model, and

3) Inadequacies of the estimation procedures.

Of these three sources, non-unidimensionality has the most serious impact for test users. Under this circumstance, items cannot be characterized as having uniquely identified parameters and examinee abilities estimated from any derived item parameters are left undefined as well. As an end result, one might be in no better position than if original raw number correct scores is used. In fact, one's position could be worse, in fact, if the test user were to act as if the ability estimates were item-free and sample-free.

If the sources of instability are due to model mis-specification or estimation inadequacies, and not due to non-unidimensionality, then one can speak of true values for both item and ability parameters which are only being inaccurately estimated. In this case, increased stability may be obtained through relatively straightforward fixes, such as going from a one-parameter model to a three-parameter, or increasing sample sizes. However, more complicated solutions may be needed, such as the development of a new model with different types of parameters.

Without the presence of external criteria it is difficult to separate out the various sources of instability; however, it is possible to gather circumstantial evidence that may enable one to deduce their relative effects. Under ideal circumstances, both item and examinee parameters should be estimable and stable regardless of the item and the examinees used in the estimation procedure. Therefore, one would expect that item parameters estimated from two separate runs on independent samples of examinees should correlate very highly with one another. Likewise examinee abilities estimated for independent subsets of items but calibrated to the same latent trait scale should also correlate very highly with one another. If these high correlations are maintained across nonrandom samples of items and examinees, one can place considerably more confidence in the parameter estimate.

With the Reading Comprehension Test data, the stability of item parameter estimates was investigated across independent random samples using different sample sizes in item sets. Table 3 contains the correlations for each of the three item parameters using different sample sizes. The correlations are between the item parameter estimates as they were derived from separate random samples of examinees. Thus for the 45-item Reading Comprehension Test, the Logist program produced 45 difficulty parameters for a sample of 1,000 examinees. Another Logist run was made with another sample of 1,000 examinees, and again it produced 45 difficulty parameters. The correlation between these two sets of difficulty parameters appears in Table 3 in the row labeled b. Similarly, correlations were produced for the discrimination and guessing parameters a and c.

Table 3

Stability Correlation of Item Parameter
Based on Sample Sizes of 1,000 and 500

	N = 1,000	N = 500
a	.72	.70
b	.97	.95
c	.64	.35

Table 4

Stability Correlation of Item Parameter for
Odd and Even Item Sets Based on Sample Sizes of 1,000

	Odd Item (N = 23)	Even Items (N = 22)
a	.62	.68
b	.97	.96
c	.35	.82

Table 5

Stability Correlation of Item Parameter for
Guessable and Non-Guessable Item Sets Based on
Sample Sizes of 1,000

	Guessable (N = 14)	Non-Guessable (N = 24)
a	.93	.38
b	.97	.91
c	.82	.25

The diffi~~culty~~ parameter has the highest correlation (.9699), discrimination is next highest (.7225), and the guessing parameter is lowest (.6448). In order to investigate the effect of sample sizes on the stability of estimates similar correlations were produced with sample sizes of 500. Both a and b parameters maintained the same magnitudes (.9546 and .7027 respectively), but the correlation for the guessing parameter drops considerably (to .3502). This suggests the importance of sample size in the estimation of the c parameter; however, the discrimination parameter correlations of .72 and .70 also indicate room for improvement.

Besides the effect of examinee sample sizes, the number of items being estimated may also have an effect on the stability of the estimation procedures. Because Logist utilizes maximum likelihood estimate procedures, the estimates are likely to be biased, especially when the total number of examinees by items observations are limited (Andersen, 1973). Table 4 illustrates the effect of reducing the number of items by half. Using sample sizes of 1,000, the correlations were calculated for odd items and again for even items. The stability of the difficulty parameters remains high (.97 and .96 for the odd and even item sets respectively), but the stability of the discrimination parameters drops. Surprisingly, however, the c parameter stability goes up considerably for the even items but falls for the odd items. This appears to suggest that the stability of the item parameters independent of sample sizes has a lot to do with the types of items included in the analysis. In other words, the unidimensionality of the items in the Reading Comprehension Test is questionable.

Pursuing this line of reasoning, it was felt that the 45 items could be classified in some way to produce more homogeneous item sets. Because the influence of guessing has received quite a lot of attention in the application of the three-parameter model, one method of classifying the items is on the basis of their guessability, that is, the likelihood of getting an item correct without possessing the requisite knowledge. In order to classify the item as guessable, the 45 reading items without their corresponding reading passages were presented to eight adult college-educated subjects. Guessable items were judged to be those for which seven of the eight subjects were able to answer correctly without having read the passages, while non-guessable items were those which two or fewer subjects were able to get correct.

In all, 14 items were classified as guessable, and 24 were classified as non-guessable. The resulting item correlations from the IEA examinees are based on sample sizes of 1,000 and are presented in Table 5. The stability of parameter estimates goes up for all three parameters for the guessable items and goes down for the non-guessable items. The stability correlations for the discrimination parameter goes up considerably for the guessable items (to a respectable .93), and the correlation for the c parameter also goes up (to .82). For the non-guessable items, the a and c parameters go down (to .38 and .25, respectively) which seem to indicate that the non-guessable items are non-unidimensional and that the non-unidimensionality is responsible for most of the instability of the item estimates.

The strategy used in the preceding three-parameter analysis was principally one of deduction from available correlational evidence without the use of external validating criteria. The general conclusion for the Reading Comprehension Test data is that the 45 items are not unidimensional and that such non-unidimensionality considerably affects the stability of Logist estimates. It should be noted, in particular, that this non-unidimensionality would not have been detected through the estimation of difficulty parameters alone as would be produced by the Rasch analysis.

The results of the three-parameter study also seemed to provide some evidence for the nature of the reading test behavior of the set of examinees. It seems that much of what is called reading ability depends on what the student brings to the reading situation, i.e., his or her own experiences with and exposure to particular topics. This may underly the higher stability of the parameter estimates for the guessable items as contrasted with the non-guessable items. The non-unidimensionality of the latter should not be too surprising since examinées, presumably, must read the passages before they select an answer, and their subsequent ability to respond correctly to the item is probably a function of several of reading compmrehension and test-taking strategies.

REFERENCES

Andersen, E.B. Conditional inference for multiple choice questionnaires. British Journal of Mathematical and Statistical Psychology, 1973, 26, 31-44.

Thorndike, R.L. Reading comprehension education in fifteen countries: An empirical study. New York: Wiley, 1973.

Wolf, R.M. Achievement in America. New York, Teachers College Press, 1977.

SUMMARY PAPER

J. Ward Keesling

I. Introduction: What should a measurement model provide?

A. An assessment of the fit of the model

B. Parameter estimates that capture information of importance about the elements of the model (e.g., person and item characteristics)

1. The estimated parameters for persons are the "measurements" in the model
2. The estimated parameters characterizing items should provide insight about the items (e.g., their difficulty levels) and permit more sophisticated construction and interpretation of tests
3. The special case of the multiple-choice item. The need for parameters to characterize distractors.

C. Estimates of the precision of the parameter estimates—to help us understand the latter statistics.

D. Overview of the chapter

II. Evaluation of the models given the above criteria

A. Logistic models

B. S-P model

C. G-Theory (is this really a measurement model?)

D. AUC models

III. An examination of the salience of the models to three types of use

A. Assessing pupil progress in a classroom

B. The norm-referenced evaluation

C. The domain-referenced evaluation

(For each, discuss the utility of the information in the various models, vs the cost of obtaining it. Attend especially to the potential of item banks.)

IV. Implications of microcomputer technology

(Review III, with a view to how technology could help/hinder)

V. Summary