FOUNDATIONS OF A NEW TEST THEORY

Robert J. Mislevy

This research was sponsored in part by the Cognitive Science Program, Cognitive and Neural Sciences Division, Office of Naval Research, under Contract No. N00014-88-K-0304. P&T 4421552

Robert J. Mislevy, Principal Investigator
Educational Testing Service
Princeton, New Jersey

October 1989

Reproduction in whole or in part is permitted for any purpose of the United States Government. Approved for public release; distribution unlimited.
Foundations of a New Test Theory (Unclassified)

Robert J. Mislevy

Technical

13a. Date of Report (Year, Month, Day) October 1989

13c. Source of Funding Numbers

13d. Program Element No.

13e. Project No.

13f. Task No.

13g. Work Unit No.

1. Title (Include Security Classification)

2. Security Classification of This Page Unclassified

3. Distribution/Availability of Report

4. Performinig Organization Report Numbers

5. Monitoring Organization Report Numbers

6. NAME OF PERFORMING ORGANIZATION

7. NAME OF MONITORING ORGANIZATION

8. NAME OF SPONSORING ORGANIZATION

9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER

10. SOURCE OF FUNDING NUMBERS

11. Title (Include Security Classification)

12. Personal Author(s)

13. Type of Report

14. Date Covered

15. Supplementary Notation

16. Subject Terms (Continue on reverse if necessary and identify by block number)

17. Classification Code

18. Abstract (Continue on reverse if necessary and identify by block number)

19. Distribution/Availability of Abstract

20. Security Classification of This Page Unclassified

21. Security Classification of Abstract Unclassified

22. Name of Responsible Individual

23. Office Telephone (include Area Code)

DD Form 1473, JUN 86
Foundations of a New Test Theory

Robert J. Mislevy

Educational Testing Service

October 1989

To appear in N. Frederiksen, R. Mislevy, and I. Bejar (Eds.), Test Theory for a New Generation of Tests, Lawrence Erlbaum and Associates. This work was supported in part by Contract No. N00014-88-K-0304, R&T 4421552, from the Cognitive Science Program, Cognitive and Neural Sciences Division, Office of Naval Research. I am grateful to Henry Braun, Norman Frederiksen, and Kentaro Yamamoto for comments on earlier drafts.
Foundations of a New Test Theory

Abstract

It is only a slight exaggeration to describe the test theory that dominates educational measurement today as the application of twentieth century statistics to nineteenth century psychology. Sophisticated estimation procedures, new techniques for missing-data problems, and theoretical advances into latent-variable modeling have appeared—all applied with psychological models that explain problem-solving ability in terms of a single, continuous variable. This caricature suffices for many practical prediction and selection problems because it expresses patterns in data that are pertinent to the decisions that must be made. It falls short for placement and instruction problems based on students' internal representations of systems, problem-solving strategies, or reconfigurations of knowledge as they learn. Such applications demand different caricatures of ability—more realistic ones that can express patterns suggested by recent developments in cognitive and educational psychology. The application of modern statistical methods with modern psychological models constitutes the foundation of a new test theory.

Key Words: Cognitive psychology
Educational measurement
Item response theory
Psychometrics
Test theory
Introduction

Educational measurement faces a crisis today that would appear to threaten its very foundations. The essential problem is that the view of human abilities implicit in standard test theory—item response theory as well as classical true-score theory—is incompatible with the view rapidly emerging from cognitive and educational psychology. Learners increase their competence not by simply accumulating new facts and skills, but by reconfiguring their knowledge structures, by automating procedures and chunking information to reduce memory loads, and by developing strategies and models that tell them when and how facts and skills are relevant. The types of observations and the patterns in data that reflect the ways that students think, perform, and learn cannot be accommodated by traditional models and methods. To some it would seem to some that psychometrics has little to offer in the quest to apply this new knowledge to the practical educational problems of the individual, the classroom, or the nation (Hunt and MacLeod, 1978).

I concur that the standard methods of test theory do not suffice for solving problems cast in the framework of what we are learning about how people acquire knowledge and competence, but I cannot agree that psychometrics has nothing to offer.

Standard test theory evolved as the application of statistical theory with a simple model of ability that suits the decision-making environment of most mass educational systems. Broader educational options, based on insights into the nature of learning and supported by more powerful technologies, demand a broader range of models of capabilities—still simple compared to the realities of cognition, but capturing patterns that inform a broader range of alternatives. A new test theory can be brought about by applying to well-chosen cognitive models the same general principles of statistical inference that led to standard test theory when applied to the simple model.
The first half of this paper sketches the evolution of standard test theory, highlighting the challenges that spurred each new advance. The challenges that cognitive and educational psychology present today are then discussed, and a framework for responding to that challenge is outlined. Directions for needed development are exemplified with current work.

The Early Context of Educational Decisions

The kinds of decisions that shaped the evolution of classical test theory were nearly universal in education at the beginning of this century, and dominate practice yet today. They were born of the constraints educators encountered as they launched their campaign to provide education on a broader scale than had ever been attempted hitherto:

"...the demand for tests arose during the period when school attendance was made compulsory and when higher education was developing its strengths. Educators faced the unprecedented dilemma of dealing with the range and diversity of abilities and backgrounds that individuals bring to schooling. They needed ways of determining which children and youths would be able to profit from some form of instruction as given in ordinary school and college practices as designed essentially for the majority of the population." (Glaser, 1981, p. 924).

Educators were confronted with selection or placement decisions for large numbers of students. Resources limited the information they could gather about each student, constrained the number of options they could offer, and precluded tailoring programs to individual students once a decision was made.

A first example is selecting applicants into a college that presents the same material in the same way to all students. There is only one treatment, and the alternatives are to accept or reject. The admissions officer would prefer to accept those who are likely to succeed. When resources permit more than one decision option, the usual generalization of the accept/reject paradigm is to offer a sequence of alternatives, each more demanding than the next. Placing high school freshmen into
academic tracks is an example of this latter type. Problems of selection into a single program and of placement into a single sequence are both decisions about "linearly ordered options."

Exposing a diverse group of students to a uniform educational treatment typically produces a distribution of outcomes (Bloom, 1976). An individual's degree of success depends on how his or her unique skills, knowledge, and interests match up with the equally multifaceted requirements of the treatment.

At costs substantially lower than personal interviews or performance samples, responses to multiple-choice test items provide information about certain aspects of this matchup. What is necessary is that each item tap some of the skills required for success. Even though a single item might require only a few of the relevant skills and offer little information in its own right, a tendency to provide correct answers over a large number of items supports some degree of prediction of success (Green, 1978). If all candidates are administered the same items, and one wishes to predict success in linearly-ordered options, their number-correct scores can be used (Dawes and Corrigan, 1974). Even though the several students at a given score level possess different constellations of skills, abilities, and backgrounds, making the same decision for all of them among the available alternatives is often about as well as can be done with the available data.

Once the test and the linearly-ordered options are specified, making decisions from test performances requires nothing more complicated than adding up numbers of correct responses. Two different tests constructed for the same decision, however, invariably line up examinees differently as they draw upon different particular skills from the myriad of those potentially informative. Additional statistical machinery is required to guide one in constructing tests and evaluating their quality. Classical test theory was a first response to these needs.
Classical Test Theory

Charles Spearman (1904a, 1904b, 1907, 1910, 1913) is credited with the central idea of classical test theory (CTT): a test score can be viewed as the sum of two components, a "true" score and a random "error" term. Two similar ("parallel") tests are considered to reflect the same true score, but disagree about an examinee's observed scores because of the error components—the variance of which can, under the assumptions of CTT, be driven to zero by just making the tests long enough. Ideally decisions would be based on true scores; in practice they must be based on observed scores. "Reliability," the degree to which the unobservable true scores account for the variance in observed scores, gauges the accuracy with which a test lines up a group of examinees—a reasonable criterion for the quality of a test if it is assumed that the items tap appropriate skills and scores will be used to decide among linearly ordered options.

Upon these notions was founded a practicable testing methodology. Reliability became a paramount measure of the quality of a test, although of course reliability had to be complemented with validity measures such as the correlation between test scores and subsequent performance. Validity studies had less influence on test construction, however, because they arrive too late in the process—only after the test has been administered and examinees have been followed over time. To obtain high reliability, one uses items that would be answered correctly by about half the examinees, for example, and avoids items that would have low correlations with the total test scores.

Note that these dicta could guide test construction solely from counts and patterns of right and wrong responses to candidate test items—ignoring both the content of the items and the contemplated decision alternatives. Of course good test construction does consider the knowledge, skill, and strategy requirements of items. The point is that these considerations lie outside the realm of the classical test theory. Test developers
use them independently of, sometimes in contradiction to, what test theory tells them.

Building upon Spearman's foundation, psychometricians developed a vast armamentarium of techniques for building and using tests (Gulliksen, 1950), such as approximating reliability from the internal consistency of items within a test (Kuder and Richardson, 1937) and estimating validity without knowing subsequent performances of rejected examinees (Kelley, 1923). Over time, a rigorous axiomatic foundation was laid for statistical inference under the aegis of CTT (Lord, 1959; Novick, 1966; Lord and Novick, 1968). The simple partitioning of observed scores into true and error components was generalized to multiple sources of variation from items, persons, and observational settings, and the full power of analysis of variance was brought to bear upon decision-making problems using test scores (Gronbach, Gleser, Nanda, and Rajaratnam, 1972; Lord and Novick, 1968).

A source of dissatisfaction with CTT early on was that its characterizations of examinees, such as total score and percentile rank, and of items, such as percent-correct and item-test correlation, are confounded descriptions of the particular items that constitute a test and a particular group of examinees who takes it (Wright, 1968). If one test consists of easier items than a second otherwise similar test, examinees' scores on the two tests are not directly comparable and score distributions have different shapes. If a test is administered to groups of examinees that differ in proficiency, item percents-correct and item-test correlations differ. When many tests could be constructed for the same purpose, differing perhaps in difficulty or length, should not there be a way to characterize examinees independently of the test they took, and items independently of the examinees who took them?

In attitude measurement, where agreements to a topic are analogous to correct answers to test questions, L.L. Thurstone (1928) expressed the following desideratum: "If a scale is to be regarded as valid, the scale values of the statements should not
be affected by the opinions of the people whose responses help to construct it." Thurstone (1925) and E.L. Thorndike (Thorndike et al., 1926) pioneered efforts to relate test scores to psychological traits, using item percent-correct and assumptions about distributions of traits to transform scores from different tests onto the same scale.

Thurstone and Thorndike scaling, despite allusions to an underlying trait, remained essentially theories for scores, albeit transformed (with the aid of untestable assumptions) to permit comparisons across nonparallel tests. Psychological traits per se appear as explicit parameters in the models of Ferguson (1950), Lawley (1943), and Tucker (1946). These researchers studied test construction problems within CTT by making an assumption beyond those of CTT proper; namely, that aside from random factors, item responses were driven by an unobservable ability variable. A second generation of test theory began to take form as attention shifted from test scores as the object of inference, to unobservable variables hypothesized to have produced them.

**Item Response Theory**

Item response theory (IRT), or "latent trait theory," as it was called then, appears as a test theory in its own right in the work of Frederic Lord (1953) and Georg Rasch (1960). Like classical test theory, IRT concerns examinees' overall proficiency in a domain of tasks. But while CTT makes no statement about the mechanisms that give rise to performance, IRT posits a single, unobservable, proficiency variable.1

At the heart of IRT is a mathematical model for the probability that a given person will respond correctly to a given

---

1 If classical test theory offers a statistical model for test scores without a psychological model, Guttman's (1944) scaling techniques offer a psychological model without a statistical model. Important in the reconceptualization of the scaling of test scores, a Guttman scale can be viewed as the limiting case in IRT in which each item is perfectly informative about whether an examinee's ability lies above or below a specific point on an ability continuum.
item, a function of that person's proficiency parameter and one or more parameters for the item. The item's parameters express properties such as difficulty or sensitivity to proficiency. The item response, rather than the test score, is the fundamental unit of observation. If an IRT model holds, responses to any subset of items support inferences on the same scale of measurement.

This conceptualization opens the door to solving many practical testing problems that were difficult under CTT, such as:

- **Test construction** (Birnbaum, 1968; Theunissen, 1985). If item parameters are available for a collection of items, tests can be constructed for optimal performance in specific applications, such as minimizing classification errors.

- **Adaptive testing** (Lord, 1980, Chapter 10; Weiss, 1984). An adaptive testing scheme selects the best item to administer next to an examinee, based on the amount of information that various available items would provide and a provisional estimate of the examinee's proficiency from responses to items given thus far.

- **Educational assessment** (Bock, Mislevy, and Woodson, 1982; Choppin, 1976; Messick, Beaton, and Lord, 1983). Assessments gauge proficiencies at the level of populations rather than individuals, to evaluate programs and monitor trends. IRT makes it possible to establish a stable measurement scale while allowing assessment instruments to evolve over time.

This work assumed, for the most part, that the IRT model was known and correct, and that true values or accurate estimates of item parameters were available. Current IRT research emphasizes integrating IRT into the general framework of statistical inference, and acquiring an understanding of just when and how IRT models are appropriate.

**Statistical Inference in Item Response Theory**

Early applications of IRT were designed more to demonstrate its potential than to solve actual measurement problems. Data were gathered with tests written according to CTT dicta; the same long tests were administered to many examinees, and each item had
passed CTT quality checks. Illustrative purposes were served adequately by rough estimation procedures that treat point estimates of examinee- and item-parameters as if they were the parameters themselves, ignoring the uncertainty associated with the estimates. These approximations break down when IRT is applied beyond the usual limits of CTT testing, as when examinees are presented only, say, fifteen items in adaptive testing or five in educational assessments (Mislevy, 1988). In response, IRT researchers have turned to two active lines of research in statistics: missing data methods and Bayesian estimation.

Missing data methods are relevant because a latent variable such as an IRT examinee proficiency parameter can be viewed as a datum whose value is missing for everyone. General results on estimating parameters when some data are missing, such as Dempster, Laird, and Rubin's (1977) EM algorithm, have led to methods of item parameter estimation that are at once rigorous and efficient (e.g., Eock and Aitkin, 1980; Tsutakawa, 1984). Results on statistical information in missing data problems yield insights into the uncertainty structures of IRT parameters (Mislevy and Sheehan, in press; Mislevy and Wu, 1988) and offer ways of increasing accuracy by exploiting collateral information about items and examinees (Mislevy, 1987, 1988a).

The Bayesian perspective confronts uncertainty head on, expressing what is known about parameters as probability distributions. When these distributions are concentrated, the expedient of using point estimates as if they were the true parameters can give acceptable results in subsequent analyses. But when the distributions are diffuse, one must propagate the uncertainty into subsequent analyses to obtain correct inferences. Statistical reasoning along these lines was proposed as far back as 1927 by Kelley (1927), and championed by Novick in the 1970's (e.g., Novick and Jackson, 1974), but only now are the ideas gaining currency. In this framework, one can determine when the standard, simpler, approximations suffice, but use (admittedly more complex) correct analyses when they don't. For examples in

The Question of Model Fit

But of course the IRT model is never exactly correct. A single variable that accounts for all nonrandomness in examinees' responses is not a serious representation of cognition, but a caricature that can solve applied problems when it captures the patterns that are salient to the job. The pattern that CTT and IRT can capture is examinees' tendencies to give correct responses, which can usefully inform decisions about linearly ordered alternatives. IRT was a practical advance beyond CTT because it provides information about overall proficiencies in more flexible ways. It was a conceptual advance because it provides a framework for detecting anomalies in the "overall proficiency" paradigm.

This can be illustrated with Rasch's (1960) model for right/wrong items, supposing for convenience all examinees are presented the same test. Under CTT, all examinees with a given total score would be treated alike. Under the Rasch model, all examinees with the same score would receive the same ability estimate and might also be treated alike--depending on an analysis of model fit. Combining an examinee's proficiency estimate with an item's difficulty estimate, the Rasch model states how likely a correct response would be if the single-proficiency conception of ability were true. The items that high scorers missed should usually be easy ones, and the items low scorers got right should be easy ones. Finding that these

---

2 Under other IRT models such as the 2- and 3-parameter logistic models, examinees with the same total score need not receive exactly the same ability estimate, but usually receive similar estimates. Correlations between total scores and IRT estimates in typical educational tests are usually above .95, and few decisions would be made differently with any IRT model, or, if everyone has taken the same test, even with CTT.
patterns hold supports making the same decisions about people with the same scores, because, to an approximation, they got the same items right and the same ones wrong. Total scores, and thus Rasch ability estimates, convey nearly everything these data have to say about comparing these examinees.

To the extent that high scoring examinees miss items that are generally easy and low scoring examinees get hard ones right, neither total scores nor IRT ability estimates may be capturing all the systematic information in the data. Analyses of an individual's unexpected responses can reveal misconceptions or atypical patterns of learning (Mead, 1976; Smith, 1986; Tatsuoka, 1983). To understand these patterns one must look beyond the simple universe of the IRT model--to the content of the items, the structure of the learning area, the pedagogy of the discipline, and the psychology of the problem solving tasks the items demand.

Now, patterns in responses other than overall level proficiency can have educational and psychological meaning, but yet hold no salience for a particular decision. If overall proficiency in a domain of items suffices for a particular decision, as can be the case with linearly ordered educational options, cross-current patterns constitute data variation that need not be explicated. This is the essence of statistical modeling: expressing the patterns that are dominant and meaningful in terms of model parameters, and allowing for departures from these patterns in terms of distributions of residuals. But if the decision does depend on the cross-current patterns, in addition to or instead of overall proficiency, neither CTT nor standard IRT may be the right tool for the job.

The issue of model fit, then, is more pragmatic than statistical, since lack of fit must be judged in practice by the nature and the magnitude of the errors it causes. An IRT model might be satisfactory for selecting honors math students, for example, if people with similar scores have similar chances of success—even though examinees with similar scores have different profiles of skills and knowledge. The profile differences could
be modeled as "noise" without harm for the selection decision—but probably not for advising individual examinees which topics to study to maximally increase their scores.

Measuring learning is one application where IRT models can fail, because they accommodate only a highly constrained type of change: an examinee's chances of success on all items must increase or decrease by exactly the same amount (i.e., an appropriate metric). A single IRT model applied to pretest and posttest data cannot reveal how different students learn different topics to different degrees—patterns that could be at the crux of an instructional decision.

**Testing and Learning**

Good "macro-level" decisions to place students into appropriate educational programs are important in increasing the quality of education, but they are not sufficient. Tracking students as they progress opens the door to finer grained "micro-level" decisions to enhance learning along the way. Good decision-making at this level requires an inferential framework built around an understanding of how students learn.

A picture of a learner that is consistent with standard test theory is that of a collector of facts and skills, adding each to his repertoire more or less independently of others. Recent developments in psychology sketch a markedly different picture, reflecting the astounding capabilities and the surprising limitations of the mind—lightning fast recognition of stored patterns and creative applications of heuristic strategies, on the one hand; yet with short term memory capacities of only about seven elements and an inability to perform more than one attention-demanding task at a time. **Performance** is to be understood through the availability of well-practiced procedures that no longer demand high levels of attention ("automaticity"); strategies by which actions are selected, monitored, and, when necessary, switched ("metacognitive skills"); and the mental structures that relate facts and skills ("schema"). **Learning** is
to be understood through the automatization of procedures; the
acquisition and enhancement of metacognitive skills; and the
construction, revision, and replacement of schema.

Comparing the performances of novices and experts offers
insights into the nature of performance and learning. A first,
unsurprising, difference is that experts command more facts and
concepts than novices, and have richer interconnections among
them. Interconnections overcome limitations of short term memory;
while the novice may work with seven distinct elements, the expert
works with seven constellations that embody relationships among
many elements ("chunking"). Moreover, experts often organize
their knowledge in schemata possessing not simply more
connections, but qualitatively different ones. The advanced
concept: that college physics students acquire, for example, can
be organized around informal associations or naive misconceptions
(Caramazza, McCloskey, and Green, 1981). These novices tackle
physics problems in less effective ways than expert physicists,
whose more appropriate schemata lead them to the crux of the
matter (Chi, Feltovich, and Glaser, 1981). Experts also differ
from novices by having automatized, through study and practice,
procedures that were once slow and attention consuming, allowing
them to focus on novel aspects of a problem, look from different
perspectives, and more efficiently monitor and guide their efforts
as they work (Lesgold and Perfetti, 1978).

The challenge to education is to discover what experiences
help a learner with a given configuration of propositions, skills,
and connections to reconfigure that knowledge into a more powerful
arrangement. Vosniadou and Brewer (1987) point to Socratic
dialogue and analogy as mechanisms that facilitate such learning.
To apply them effectively, one must take into account not simply
target configurations, such as the expert's model, but the
individual learners' current configurations. The challenge to
test theory is to provide models and methods to assess knowledge,
and to guide instruction, as seen in this new light.
To what extent can standard test theory meet this challenge? Recall that standard test theory characterizes performance only as to overall level of proficiency, and learning only as to change in overall proficiency. Cronbach and Furby (1970) note the inadequacy of such measures of change when applied with conventional broad range educational tests:

Even when test scores \( X \) and \( Y \) are determined by the same operation (e.g., scores under the same CTT or IRT model), they often do not represent the same psychological processes (Lord, 1958). At different stages of practice or development different processes contribute to performance of a task. Nor is this merely a matter of increased complexity: some processes drop out, some remain but contribute nothing to individual differences within an age group, some are replaced by qualitatively different processes. (p. 76).

Standard test scores can be connected more closely with cognition if they summarize performance over only tasks that are very homogeneous in their requirements (Glaser, 1963). And this specificity marked the criterion referenced testing movement of the 1960's and 1970's. Merely defining testing areas very narrowly, however, is not sufficient to make test scores instructionally relevant (Glaser, 1981). A list of scores in narrowly defined areas ignores the interconnections among scores induced by the knowledge, skills, and strategies they tap in pairs, in triples, or in hierarchies of the specific behaviors—yet it is at just this level that instructional relevance must be sought.

**New Tests, New Test Theory**

A learner's state of competence at a given point in time is a complex constellation of facts and concepts, and the networks that interconnect them: of automatized procedures and conscious heuristics, and their relationships to knowledge patterns that signal their relevance: of perspectives and strategies, and the management capabilities by which the learner focuses his efforts.
There is no hope of providing a complete description of such a state. Neither is there a need to. The new pedagogy need merely(!) identify communalities among states of competence that can be linked to instructional actions that facilitate changes to preferable states. Distinctions need not be made among all possible states, but only among classes of states with different instructional implications. The new tests to inform instructional decisions need merely(!) present tasks that learners in the different states are likely to carry out in observably different ways. Not only correctly as opposed to incorrectly, but at what speed, with what intermediate products, or with which incorrect response; not simply as independent pieces of information from distinct items, but in patterns of similarity, dissimilarity, or independence across tasks that probe knowledge structures and problem-solving strategies. The new test theory need merely(!) provide models whose parameters are capable of expressing the salient patterns, and inferential procedures upon which to base instructional decisions in the presence of uncertainty.

Foundations of the new pedagogy are to be found in the union of analyses of key concepts in a substantive area, research into the cognitive psychology of the area, and detailed observations of learners as they progress. Greeno (1976) argues that the tools and the perspectives of cognitive and educational psychology have developed to a point at which they can be used to generate instructional objectives in this manner. He provides detailed illustrations in three substantive domains at increasing levels of complexity and sophistication: fourth-grade fractions, high school geometry, and college level auditory psychophysics.

Foundations of the new theory of test construction are similarly to be found in educational and cognitive psychology (Embretson, 1985a; Messick, 1984). Standard vocabulary items suffice to ascertain the breadth of a learner's familiarity with concepts in a substantive area, but tasks based on analogies probe the interconnections among concepts. Speed of response is more informative than correctness about the automaticity of procedures.
hence a better guide to assigning additional practice on a currently conscious process. Designing appropriate measures demands familiarity with the substantive field, not just about the knowledge structures of the expert but about the incomplete or inaccurate structures novices often use. To see how the requisite cognitive and substantive analyses might be carried out, and how tasks that differentiate among learners at different states of competence might then be constructed, the reader is referred to Curlee and Glaser (1983) on reading achievement and Marshall (1985) on "story problems" in arithmetic.

Foundations of the new test theory are to be found in the general principles that led to the development of item response theory. The examinee will be characterized by parameters that express tendencies to act in accordance with the various continuous levels or discrete states in simplified models of cognition. Tasks will be characterized by parameters indicate the extent to which they tap different aspects of knowledge structures, procedures, or strategies. As in IRT, individual differences among examinees that are not salient to the decision will be modeled as random—not as a psychologically tenable position, but as a practically useful expedient.

Beyond "Low-to-High Proficiency"

The breadth of problems to which standard test theoretic models have been usefully employed, despite their limited low-to-high conception of proficiency, suggests a certain robustness of modeling. It is not necessary that models account for all possible ways students might approach a test, but it is necessary that they can capture instructionally relevant patterns. A test must be designed to highlight the pertinent patterns, and analyzed with a model capable of expressing them.

The idea of building test items around cognitive principles can be traced back at least as far as to Guttman's facet design tests (Guttman, 1970). Guttman worked out analytic methods for analyzing data from such tests within the framework of classical
test theory. Scheiblechner (1972) and Fischer (1973), with their "linear logistic test model" expressed item difficulty parameters in the Rasch IRT model as functions of psychologically salient features of test items, but still characterized examinees in terms of overall proficiency. More recently, test theory models built around patterns other than overall proficiency have begun to appear in the psychometric literature.

"Tectonic plate" models. Increasing competence in a substantive area need not be reflected as uniformly increasing chances of success on all tasks. Patterns of smooth increase may be observed for certain people on certain sets of tasks, in certain phases of development; standard test theory will give good summaries of change in these neighborhoods. Discontinuous patterns of change begin to appear as the scope of tasks becomes broader, as the range of development becomes greater, and as the range of experiences of examinees becomes more diverse. "Tectonic plate" models generalize IRT by allowing for a limited number of predetermined, theory-driven, discontinuities in item response patterns. In tectonic plate geological models, points within a given land mass, or plate, maintain their relative positions, but the plates move with respect to one another. In tectonic plate psychometric models, items tapping the same set of skills maintain their difficulties relative to one another, but the difficulties of the groups of items change with respect to other groups as learners acquire new skills or concepts.

Wilson's (1985, 1989) "Saltus" model extends the Rasch IRT model to development with discontinuous jumps. An example is Siegler's (1981) rule-learning analysis of balance-beam tasks, where students can increase their competence either by using the rules they know more effectively (continuous change) or by learning new rules (discontinuous change). Sometimes students who learn a new rule begin to miss a type of problem they used to get right, because their previous, less complete, set of rules gave the right answers for the wrong reasons. This pattern flouts standard test theory. The Saltus model assumes that each examinee
is in one of a number of unobservable stages of development.

-ems are classified so that all items in a class have the same relationship to developmental stages. One set of item parameters expresses relative difficulties among items within item classes, which, like Rasch item difficulty parameters, are the same for people in all stages. A second set of parameters quantifies patterns that the Rasch model cannot express: differences in relative difficulties between item classes for people in different stages, such as the difficulty reversals mentioned above. Saltus is effectively a mixture of standard Rasch models.

Mislevy and Verhelst (in press) have discussed mixture models more generally, listing assumptions, laying out general models, and suggesting estimation procedures. They emphasize situations in which different subjects follow different strategies, pointing out that instructional decisions can depend on how students solve problems, not just how many they solve. The salient features of items are those that can differentiate among users of different strategies, mental models, or conceptions about key relationships. An examinee is characterized by the probabilities that she employed the various alternative strategies, and a conditional estimate of proficiency under each. Measurement with such a model can indicate change that is either quantitative (e.g., the examinee employed Strategy A on both occasions, but more effectively at the second) or qualitative (e.g., she used Strategy A before instruction but Strategy B afterwards).

Latent class models. Although models with continuous latent variables have dominated educational measurement, Lazarsfeld (1950) introduced models with categorical latent variables nearly half a century ago. Most educational applications of latent class models have been in "mastery" testing: one attempts to infer an examinee's unobservable state—master or nonmaster—on the basis of observable responses (Macready and Dayton, 1977, 1980). In the more recent "binary skills" models (Haertel, 1984), examinees are classified in terms of which of a set of skills they possess. This "true" classification is unobservable. Items are classified
according to which of the skills they require for solution. This classification is known. Ideally, an examinee responds correctly to only and exactly those items that require skills he or she possesses. The stochastic parameters of the model reflect departures from this ideal.

Except in the special case of mastery testing, computational constraints have limited applications of latent class models to no more than about ten items until recently. Information about skill profiles in groups can be gleaned from such data, but individuals' skills could not be inferred accurately. Improved computational procedures have opened the door to applications with 50 or 60 items (e.g., Paulson, 1986; Yamamoto, 1987), and work with structurally similar models in expert systems holds promise of handling much larger problems (Lauritzen and Spiegelhalter, 1988). Progress in this direction is vital to educational applications, since these inferences demand more data than low-to-high proficiency inferences. Moreover, adaptive testing, which made IRT measurement more efficient, will be able to make latent class measurement practicable (Dayton and Macready, 1989; Falmagne and Doignon, 1988).

Componential models. The models described above were introduced with right/wrong test items, which, if constructed carefully, yield response patterns that differentiate examinees who tackle them in different ways. Richer information can be accumulated if it is possible to track intermediate products of solution. Consider, for example, a situation in which the binary skills model applies. Inferences about skill profiles can be stronger if one can be see which subtasks were attempted and their outcomes: overall correctness can result from one sequence of correct operations or another, or a fortuitous mixture of correct and incorrect operations; overall incorrectness can be caused by a poor plan of attack, or a flawed execution of a good plan. Early implementations of these ideas have been worked out by Embretson (1983, 1985b) and Samejima (1983).
All of the models discussed above—tectonic plate, latent class, and componential models—exhibit the same cardinal feature: they support inferences about proficiencies other than just low-to-high ability because, and only because, the user specifies theoretically salient patterns of response other than just less-to-more correct answers. Current implementations require expertise in statistics as well as in the substantive area. Test theory researchers must embed these approaches in generally applicable computer routines, or shells, so that a broader range of users can put them into practice in the substantive areas.\(^3\)

Beyond Right/Wrong, Multiple-Choice Items

Currently IRT is used almost exclusively to draw inferences about a low-to-high proficiency variable from responses to multiple-choice test items. The preceding section discussed how, even with multiple-choice data, one can found inferences upon radically different conceptions of proficiency. Inferences can be made yet stronger, and decision-making more efficient, if different kinds of data can be collected.

We have mentioned the possibility of exploiting the identity of incorrect responses to multiple-choice items, for when particular misconceptions are probed in more than one item and we wish to infer how an examinee is approaching tasks. IRT models that distinguish among incorrect alternatives have been discussed by Bock (1972), Masters (1982), Samejima (1979), and Thissen and

\(^3\) Similar diffusion processes have already occurred in two area related to test theory. The first is IRT itself. In the 1960's, only a handful of mathematically talented researchers could use IRT; now IRT is widely used by practitioners by virtue of production programs such as LOGIST (Wingersky, Barton, and Lord, 1982), BILOG (Mislevy and Bock, 1983), and BICAL (Wright, Mead, and Bell, 1980). The second area is that of linear structural relationships among variables with measurement error. Proposing such a model and solving the equations was once practically grounds for a Nobel prize in economics; now anyone with access to the LISREL computer program (Joreskog and Sorbom, 1986) can routinely carry out analyses undreamed of a few decades ago.
Steinberg (1984). These papers show how to connect observations more complex than right/wrong to the standard psychological model of low-to-high proficiency. The same machinery for the observational aspect of modeling can be used when the psychological aspect is an alternative cognitive model. Embretsen (1983, 1985b) and Masters (Masters and Mislevy, 1984) have taken some initial steps in this direction.

Because data collected on computers can provide response time routinely, response latency can also be exploited. Response latencies are particularly pertinent to inferences about automaticity: a correct answer arrived at through a laborious conscious process can have different instructional implications than the same response obtained through automatized processes. Response latencies can also be used in conjunction with correctness to design items that differentiate among examinees who use different strategies. Many quantitative items in the SAT, for example, can be solved either by a "brute force" calculation or by a simple calculation if a key relationship is recognized; "correct and fast" suggests the insightful solution. Scheiblechner (1983) and Thissen (1983) show how to use response times to measure low-to-high proficiency. Their methods of linking observed responses to expected responses could be applied with an alternative cognitive model for expected responses.

Beyond Tester-Controlled Observational Settings

Traditional educational tests present small, closed-form problems, isolated and packaged more neatly than the problems people encounter in life. Real-world tasks require one to recognize a problem space, to plan strategies, to take initial steps, and gather additional information and, observing preliminary results, to determine which direction to proceed. Controlling the observational setting in testing to some degree is probably unavoidable in a decision making system applied routinely to many learners. Controlled simulation tasks strike a compromise between the rigid, tester-controlled observational setting of...
traditional tests and the wholly unstructured observation of performance in natural settings.

The most work in this area has been carried out in the arena of medical education in the form of "patient management problems," or PMPs (Assmann, Hixon, and Kacmarek, 1979). A simulated patient (through a written or oral dialogue, or as a live actor or a computer model) presents the examinee with initial symptoms; the examinee requests tests, considers their results, prescribes treatments, and monitors their effects, generally attempting to identify and treat the initially unknown disease. Despite their appeal as evocators of critical problem-solving skills, PMPs do not seem to provide reliable data from the perspective of standard test theoretic techniques (McGuire, 1985). For the same amount of testing time, reliability coefficients of PMP scores prove disappointingly low compared with multiple-choice tests.

A possible explanation of this result is that standard test theory analyses of PMP data are not looking for the right patterns. They look at simple additive combinations of single outcomes, rather than relationships that might suggest associations among facts in examinees' schema, or indicate the use of effective or ineffective problem-solving strategies. A distinct stream of medical research, however, does address these relationships: "expert systems" that help health care workers with diagnostic problems (e.g., Pope, 1981; Shortliffe et al., 1976).

An expert system representation of a health care area is build around associations among unobservable disease states, observable symptoms and test results, and outcomes of treatments. Some expert systems express these associations through "fuzzy logic" (Zadeh, 1983) or "belief functions" (Shafer, 1976), but others that use conditional probabilities (Spiegelhalter, 1986) are extensions of the latent class models discussed above. In an educational setting, associations would be delineated among substantive concepts, strategies, observable outcomes, and prescribed instruction (Clancey, 1988).
There are two levels at which expert systems could be implemented in educational settings. The first appears more amenable to end-of-course or macro-level decision-making, while the second seems better suited to an ongoing instructional system.

In the first, simpler, approach, an expert system is built only for a "correct" model. An examinee's responses are evaluated in terms of their efficacy at each decision point as compared with the best possible action given present information. If scores were also available from a standard multiple-choice test of knowledge, one could distinguish performance problems caused by strategic errors from those caused by knowledge deficiencies.

In the second, more ambitious, approach, not only would a correct expert system be built, but examinees' possibly "inexpert systems" would be inferred. Perhaps the best known example of this type is Anderson's (Anderson and Reiser, 1985) computer programming tutor. Although more individualized instructional prescriptions can be made in this way, inferring even selected aspects of examinees' schema and strategies requires far more data than does comparing performance to a fixed expert model. A successful system of this type would probably require a more constrained problem space and more extensive interactions of the learner with the simulation.

Conclusion

Einstein's theory of relativity revolutionized physics, but it extended rather than supplanted Newton's laws of motion. Classical mechanics still works just fine, thank you, for building bridges, planning billiards shots, and figuring out how to stand up from an overstuffed easy chair. And as long as educators are called upon to make the macro-level, linearly-ordered decisions that engendered standard test theory, standard test theory will continue to be useful, and will continue to be used. Recent developments in technology, however, provide opportunities for decision making at the micro-level more frequently and for larger numbers of students than ever before; recent developments in
education and psychology give us conceptions of competence and learning that can be used to guide these decisions.

Researchers in education and psychology have begun to lay the theoretical groundwork to link testing with the cognitive processes of learning. Meanwhile, researchers in measurement and statistics have made breakthroughs in inferential procedures for the models of standard test theory. To inform modern educational decisions requires drawing together the insights from these two strands of research—the twin foundations of a new test theory.
References


Clancey, W.J. (1988). The role of qualitative models in instruction. In J. Self (Ed.), *Artificial intelligence and*
human learning: Intelligent computer-aided instruction.
London: Chapman and Hall.


Hunt, E., and MacLeod, C.M. (1978). The sentence-verification paradigm: A case study of two conflicting approaches to individual differences. *Intelligence, 2*, 129-144.


advise physicians regarding antimicrobial therapy. Computers in Biomedical Research. 6, 544-560.


Wright, B.D., Mead, R.J., and Bell, S.R. (1980). BICAL: Calibrating items with the Rasch model. *Research Memorandum*

Dr. Stanley Collyer  
Office of Naval Technology  
Code 082  
800 N. Quincy Street  
Arlington, VA 22217-5000

Dr. Hans F. Crombag  
Faculty of Law  
University of Limburg  
P.O. Box 616  
Maastricht  
The NETHERLANDS 6200 MD

Dr. Timothy Davey  
American College Testing Program  
P.O. Box 168  
Iowa City, IA 52243

Dr. C. M. Dayton  
Department of Measurement  
Statistics & Evaluation  
College of Education  
University of Maryland  
College Park, MD 20742

Dr. Ralph J. DeAyala  
Measurement, Statistics,  
and Evaluation  
Benjamin Bldg., Rm. 4112  
University of Maryland  
College Park, MD 20742

Dr. Dattprasad Divgi  
Center for Naval Analysis  
4401 Ford Avenue  
P.O. Box 16268  
Alexandria, VA 22302-0268

Dr. Henrik Dong  
Bell Communications Research  
B Corporate Place  
FYA-1K226  
Piscataway, NJ 08854

Dr. Fritz Drasgow  
University of Illinois  
Department of Psychology  
603 E. Daniel St.  
Champaign, IL 61820

Defense Technical  
Information Center  
Cameron Station, Bldg 5  
Alexandria, VA 22314  
Attn: TC  
(12 copies)

Dr. Stephen Dunbar  
224B Lindquist Center  
for Measurement  
University of Iowa  
Iowa City, IA 52242

Dr. James A. Earles  
Air Force Human Resources Lab  
Brooks AFB, TX 78235

Dr. Kent Eaton  
Army Research Institute  
5001 Eisenhower Avenue  
Alexandria, VA 22333

Dr. Susan Embretson  
University of Kansas  
Psychology Department  
426 Fraser  
Lawrence, KS 66045

Dr. George Englehard, Jr.  
Division of Educational Studies  
Emory University  
210 Fishburne Bldg.  
Atlanta, GA 30322

Dr. Benjamin A. Fairbank  
Performance Metrics, Inc.  
5825 Callaghan  
Suite 225  
San Antonio, TX 78228

Dr. P-A. Federico  
Code 51  
NPRDC  
San Diego, CA 92152-6800

Dr. Leonard Feldt  
Lindquist Center  
for Measurement  
University of Iowa  
Iowa City, IA 52242
Dr. Richard L. Ferguson
American College Testing
P.O. Box 148
Iowa City, IA 52243

Dr. Gerhard Fischer
Liebigasse 5/3
A 1010 Vienna
AUSTRIA

Dr. Myron Fischl
U.S. Army Headquarters
DAPE-MRR
The Pentagon
Washington, DC 20310-0300

Prof. Donald Fitzgerald
University of New England
Department of Psychology
Armidale, New South Wales 2351
AUSTRALIA

Mr. Paul Foley
Navy Personnel R&D Center
San Diego, CA 92152-6800

Dr. Alfred R. Fregly
AFOSR/NL, Bldg. 410
Bolling AFB, DC 20332-6448

Dr. Robert D. Gibbons
Illinois State Psychiatric Inst.
Rm 529W
1601 W. Taylor Street
Chicago, IL 60612

Dr. Janice Gifford
University of Massachusetts
School of Education
Amherst, MA 01003

Dr. Robert Glaser
Learning Research & Development Center
University of Pittsburgh
3333 O'Hara Street
Pittsburgh, PA 15260

Dr. Bert Green
Johns Hopkins University
Department of Psychology
Charles & 34th Street
Baltimore, MD 21218

DORNIER GMBH
P.O. Box 1420
D-7990 Friedrichshafen 1
WEST GERMANY

Prof. Edward Haertel
School of Education
Stanford University
Stanford, CA 94305

Dr. Ronald K. Hambleton
University of Massachusetts
Laboratory of Psychometric
and Evaluative Research
Hills South, Room 152
Amherst, MA 01003

Dr. Delwyn Harnisch
University of Illinois
51 Gerty Drive
Champaign, IL 61820

Dr. Grant Henning
Senior Research Scientist
Division of Measurement
Research and Services
Educational Testing Service
Princeton, NJ 08541

Ms. Rebecca Hetter
Navy Personnel R&D Center
Code 63
San Diego, CA 92152-6800

Dr. Paul W. Holland
Educational Testing Service, 21-T
Rosedale Road
Princeton, NJ 08541

Prof. Lutz F. Hornke
Institut fur Psychologie
RWTH Aachen
Jaegerstrasse 17/19
D-5100 Aachen
WEST GERMANY

Dr. Paul Horst
677 G Street, #184
Chula Vista, CA 92010
Mr. Dick Hoshaw
O''-135
Arlington Annex
Room 2834
Washington, DC 20350

Dr. Lloyd Humphreys
University of Illinois
Department of Psychology
603 East Daniel Street
Champaign, IL 61820

Dr. Steven Hunka
3-104 Educ. N.
University of Alberta
Edmonton, Alberta
CANADA T6G 2G5

Dr. Huynh Huynh
College of Education
Univ. of South Carolina
Columbia, SC 29208

Dr. Robert Jannarone
University of South Carolina
Columbia, SC 29208

Dr. Douglas H. Jones
Thatcher Jones Associates
P.O. Box 6640
10 Trafalgar Court
Lawrenceville, NJ 08648

Dr. Brian Junker
University of Illinois
Department of Statistics
101 Illini Hall
725 South Wright St.
Champaign, IL 61820

Dr. Milton S. Katz
European Science Coordination
Office
U.S. Army Research Institute
Box 65
FPO New York 0951J-1500

Prof. John A. Keats
Department of Psychology
University of Newcastle
N.S.W. 2308
AUSTRALIA

Dr. G. Gage Kingsbury
Portland Public Schools
Research and Evaluation Department
501 North Dixon Street
P. O. Box 3107
Portland, OR 97209-3107

Dr. William Koch
Box 7246, Meas. and Eval. Ctr.
University of Texas-Austin
Austin, TX 78703

Dr. Leonard Kroeker
Navy Personnel R&D Center
Code 62
San Diego, CA 92152-6800

Dr. Jerry Lehnus
Defense Manpower Data Center
Suite 400
1600 Wilson Blvd
Rosslyn, VA 22209

Dr. Thomas Leonard
University of Wisconsin
Department of Statistics
1210 West Dayton Street
Madison, WI 53705

Dr. Michael Levine
Educational Psychology
210 Education Bldg.
University of Illinois
Champaign, IL 61801

Dr. Charles Lewis
Educational Testing Service
Princeton, NJ 08541-0001

Dr. Robert L. Linn
Campus Box 249
University of Colorado
Boulder, CO 80309-0249

Dr. Robert Lockman
Center for Naval Analysis
4401 Ford Avenue
P.O. Box 16266
Alexandria, VA 22302-0268

Dr. Frederic M. Lord
Educational Testing Service
Princeton, NJ 08541
<table>
<thead>
<tr>
<th>Name</th>
<th>Address</th>
</tr>
</thead>
<tbody>
<tr>
<td>Assistant for MPT Research,</td>
<td>Mr. Steve Reiss</td>
</tr>
<tr>
<td>Development and Studies</td>
<td>N660 Elliott Hall</td>
</tr>
<tr>
<td>OP 0187</td>
<td>University of Minnesota</td>
</tr>
<tr>
<td>Washington, DC 20370</td>
<td>75 E. River Road</td>
</tr>
<tr>
<td>Dr. Judith Orasanu</td>
<td>Minneapolis, MN 55455-0344</td>
</tr>
<tr>
<td>Basic Research Office</td>
<td>Dr. Carl Ross</td>
</tr>
<tr>
<td>Army Research Institute</td>
<td>CNET-PDCD</td>
</tr>
<tr>
<td>5001 Eisenhower Avenue</td>
<td>Building 90</td>
</tr>
<tr>
<td>Alexandria, VA 22333</td>
<td>Great Lakes NTC, IL 60088</td>
</tr>
<tr>
<td>Dr. Jesse Orlansky</td>
<td>Dr. J. Ryan</td>
</tr>
<tr>
<td>Institute for Defense Analyses</td>
<td>Department of Education</td>
</tr>
<tr>
<td>1801 N. Beauregard St.</td>
<td>University of South Carolina</td>
</tr>
<tr>
<td>Alexandria, VA 22311</td>
<td>Columbia, SC 29208</td>
</tr>
<tr>
<td>Dr. Peter J. Pashley</td>
<td>Dr. Fumiko Samejima</td>
</tr>
<tr>
<td>Educational Testing Service</td>
<td>Department of Psychology</td>
</tr>
<tr>
<td>Rosedale Road</td>
<td>University of Tennessee</td>
</tr>
<tr>
<td>Princeton, NJ 08541</td>
<td>310B Austin Peay Bldg.</td>
</tr>
<tr>
<td></td>
<td>Knoxville, TN 37916-0900</td>
</tr>
<tr>
<td>Dr. Wayne M. Patience</td>
<td>Dr. Drew Sands</td>
</tr>
<tr>
<td>American Council on Education</td>
<td>NPRDC Code 62</td>
</tr>
<tr>
<td>GED Testing Service, Suite 20</td>
<td>San Diego, CA 92152-6800</td>
</tr>
<tr>
<td>One Dupont Circle, NW</td>
<td>Lowell Schoer</td>
</tr>
<tr>
<td>Washington, DC 2003E</td>
<td>Psychological &amp; Quantitative Foundations</td>
</tr>
<tr>
<td>Dr. James Paulson</td>
<td>College of Education</td>
</tr>
<tr>
<td>Department of Psychology</td>
<td>University of Iowa</td>
</tr>
<tr>
<td>Portland State University</td>
<td>Iowa City, IA 52242</td>
</tr>
<tr>
<td>P.O. Box 751</td>
<td>Dr. Mary Schratz</td>
</tr>
<tr>
<td>Portland, OR 97207</td>
<td>905 Orchid Way</td>
</tr>
<tr>
<td>Dr. John Paulson</td>
<td>Carlsbad, CA 92009</td>
</tr>
<tr>
<td>Dept. of Administrative Sciences</td>
<td>Dr. Dan Segall</td>
</tr>
<tr>
<td>Code 54</td>
<td>Navy Personnel R&amp;D Center</td>
</tr>
<tr>
<td>Naval Postgraduate School</td>
<td>San Diego, CA 92152</td>
</tr>
<tr>
<td>Monterey, CA 93943-5026</td>
<td>Dr. W. Steve Sellman</td>
</tr>
<tr>
<td>Dr. Mark D. Reckase</td>
<td>OASD (MRA&amp;L)</td>
</tr>
<tr>
<td>ACT</td>
<td>28269 The Pentagon</td>
</tr>
<tr>
<td>P.O. Box 168</td>
<td>Washington, DC 20301</td>
</tr>
<tr>
<td>Iowa City, IA 52243</td>
<td>Dr. Kazuo Shigemasu</td>
</tr>
<tr>
<td>Dr. Malcolm Ree</td>
<td>7-9-24 Kugenuma-Kaigan</td>
</tr>
<tr>
<td>AFHRL/MDA</td>
<td>Fujisawa 251</td>
</tr>
<tr>
<td>Brooks AFB, TX 78235</td>
<td>JAPAN</td>
</tr>
</tbody>
</table>
Dr. Howard Wainer  
Educational Testing Service  
Princeton, NJ 08541

Dr. Ming-Mei Wang  
Lindquist Center for Measurement  
University of Iowa  
Iowa City, IA 52242

Dr. Thomas A. Warm  
FAA Academy AAC934D  
P.O. Box 25082  
Oklahoma City, OK 73125

Dr. Brian Waters  
HumRRO  
12908 Argyle Circle  
Alexandria, VA 22314

Dr. David J. Weiss  
N660 Elliott Hall  
University of Minnesota  
75 E. River Road  
Minneapolis, MN 55455-0344

Dr. Ronald A. Weitzman  
Box 146  
Carmel, CA 93921

Dr. David J. Weiss  
N660 Elliott Hall  
University of Minnesota  
75 E. River Road  
Minneapolis, MN 55455-0344

Dr. Ronald A. Weitzman  
Box 146  
Carmel, CA 93921

Major John Welsh  
AFHRL/MAAN  
Brooks AFB, TX 78223

Dr. Douglas Wetzel  
Code 51  
Navy Personnel R&D Center  
San Diego, CA 92152-6800

Dr. Rand R. Wilcox  
University of Southern California  
Department of Psychology  
Los Angeles, CA 90089-1061

German Military Representative  
ATTN: Wolfgang Wildgrube  
Streitkraefteamt  
D-5300 Bonn 2  
4000 Brandywine Street, NW  
Washington, DC 20016

Dr. Bruce Williams  
Department of Educational Psychology  
University of Illinois  
Urbana, IL 61801

Dr. Hilda Wing  
NRC MH-176  
2101 Constitution Ave.  
Washington, DC 20418

Mr. John H. Wolfe  
Navy Personnel R&D Center  
San Diego, CA 92152-6800

Dr. George Wong  
Biostatistics Laboratory  
Memorial Sloan-Kettering Cancer Center  
1275 York Avenue  
New York, NY 10021

Dr. Wallace Wulfeck, III  
Navy Personnel R&D Center  
Code 51  
San Diego, CA 92152-6800

Dr. Kentaro Yamamoto  
03-T  
Educational Testing Service  
Rosedale Road  
Princeton, NJ 08541

Dr. Wendy Yen  
CTB/McGraw Hill  
Del Monte Research Park  
Monterey, CA 93940

Dr. Joseph L. Young  
National Science Foundation  
Room 320  
1800 G Street, N.W.  
Washington, DC 20550

Mr. Anthony R. Zara  
National Council of State Boards of Nursing, Inc.  
625 North Michigan Avenue  
Suite 1544  
Chicago, IL 60611
Dr. Ratna Nandakumar  
Dept. of Educational Studies  
Willard Hall, Room 213  
University of Delaware  
Newark, DE 19716