ED 387 527 TM 023 987

AUTHOR Wainer, Howard
TITLE Measurement Problems. Program Statistics Research. Technical Report No. 92-20.
INSTITUTION Educational Testing Service, Princeton, N.J.
REPORT NO ETS-RR-92-12
PUB DATE Jan 92
NOTE 28p.
PUB TYPE Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Data Collection; Educational Assessment; Educational Research; *Measurement Techniques; Predictor Variables; *Psychometrics; Research Needs; *Research Problems; Scoring; *Statistical Analysis; Test Bias; *Test Validity
IDENTIFIERS *Standard Setting

ABSTRACT

This paper reports 20 unsolved problems in educational measurement and points toward what seem to be promising avenues of solution. The first group of concerns involves validity, posing the problems of obtaining suitable validity criteria, determining and measuring the predictor variables that best characterize the traits and proficiencies of interest, and determining the validity of a test being used for selection. Other unresolved issues are those of data insufficiencies, statistical adjustment, test fairness, domain coverage, reliable scoring, and standard setting. A number of technical issues remain to be resolved, including considerations of difficulty, problems of estimation, response time, and the complexity of psychometric models. The final question is that of attracting, training, and retaining technically-talented individuals in the field of educational measurement. (Contains 44 references.) (SLD)

RR-92-12

Measurement Problems

Howard Wainer
Educational Testing Service

PROGRAM STATISTICS RESEARCH

TECHNICAL REPORT NO. 92-20

Educational Testing Service
Princeton, New Jersey 08541

Measurement Problems

Howard Wainer
Educational Testing Service

Program Statistics Research
Technical Report No. 92-20

Research Report No. 92-12

Educational Testing Service
Princeton, New Jersey 08541

January 1992

The Program Statistics Research Technical Report Series is designed to make the working papers of the Research Statistics Group at Educational Testing Service generally available. The series consists of reports by the members of the Research Statistics Group as well as their external and visiting statistical consultants.

Reproduction of any portion of a Program Statistics Research Technical Report requires the written consent of the author(s).

Measurement Problems

Howard Wainer[1]
Educational Testing Service

Abstract

History teaches the continuity of science; the developments of tomorrow have their genesis in the problems of today. Thus any attempt to look forward is well begun with an examination of unsettled questions. Since a clearer idea of where we are going smooths the path into the unknown future, a periodic review of such questions is prudent. The present day, lying near the juncture of the centuries, is well suited for such a review. This paper reports twenty unsolved problems in educational measurement and points toward what seem to be promising avenues of solution.

[1] This research was supported by the Educational Testing Service through the Research Statistics Group and the Senior Scientist Award. The impetus to prepare this essay came from an invitation from Bruno Zumbo to address a conference on "Modern Theories in Measurement" that was held on November 10-13, 1991, at the Chateau Montebello in Quebec. Many of the ideas expressed here come from conversations with my friends and colleagues. Prominent among these are: Bill Angoff, Randy Bennett, Darrell Bock, Charles Davis, Neil Dorans, Susan Embretson, Paul Holland, Skip Livingston, Bengt Muthén, Jim Ramsay, Malcolm Ree, Lawrence Stricker, David Thissen, Neal Thomas, and Xiang-Bo Wang. In addition I am especially indebted to Gunnar Carlsson for his insightful discussion of the current state of Hilbert's problems, and to Linda Steinberg for her encouragement, careful reading, and thought provoking criticisms.

Measurement Problems

> *"Who of us would not be glad to lift the veil behind which the future lies hidden; to cast a glance at the next advances of our science and at the secrets of its development during future centuries? What particular goals will there be toward which the leading ... spirits of coming generations will strive? What new methods and new facts in the wide and rich field of (scientific) thought will the new centuries disclose?*
>
> *History teaches the continuity of the development of science. We know that every age has its own problems, which the following age either solves or casts aside as profitless and replaces by new ones. If we would obtain an idea of the probable development of ... knowledge in the immediate future, we must let the unsettled questions pass before our minds and look over the problems which the science of today sets and whose solution we expect from the future. To such a review of problems the present day, lying at the meeting of the centuries, seems to me well adapted. For the close of a great epoch not only invites us to look back into the past but also directs our thoughts to the unknown future."*

So begins what was perhaps the most influential speech ever made in mathematics. It was David Hilbert's presentation to the International Congress of Mathematicians in Paris in July of 1900. In it he described what he felt were the 23 most important unsolved problems in mathematics. His choice was wise and the formidable task of solving these problems occupies mathematicians to this day. Although most of these problems have now been satisfactorily resolved[2] there remains much to be done.

We have not yet reached the millennium. Given Hilbert's example, it would profit us to begin to formulate the problems of our field. Perhaps with an eight year running start and with the synergistic wisdom of those gathered here, we can begin to match, even in a small way, the wisdom of choice that Hilbert accomplished on his own. If so, I am assured that our field will move forward more surely and smoothly than would otherwise have been the case.

In this account I limit myself to educational measurement problems within a particular context. Specifically, measurements that lead to a decision and a consequence. Thus this discussion is not directly aimed at a process analogous to measuring someone's height: in which a measurement is made, the result is reported, and that's it. The situation I concern myself with here is more like that in the measurement of weight. There are standards asso-

[2] Hilbert's problems 1, 3, 5, 7, 10, 11, 15, 17, 19, 21 and 22 have been solved. Problems 14 and 20 proved to be false; there are partial solutions to problems 13 and 18; and substantial progress on 8, 9 and 23. To the best of my knowledge only two of Hilbert's problems truly remain open (12 and 16); the former is in the field of algebraic number theory and the latter in algebraic topology. It is usually not clear, among the other incompletely solved problems, what would constitute a complete solution (i.e. Problem 6 asks for a rigorous treatment of the axioms of physics — considering the great progress physics has made since 1900 it is clear what would have satisfied Hilbert then would not be satisfying now). For those interested in more details see Browder (1974).

ciated with good health that connect to weight, actions that can take place that can affect weight, and the success of those actions can be assessed.

I believe that since the lion's share of the educational situations requiring measurement are of the sort that suggest an action and have an outcome, this paper focuses on that arena. Other uses of educational measurement, such as they may exist, surely share some of the problems, but I mean to limit this account explicitly.

With this boundary stated, let us begin. What follows are a preliminary list of 20 problems; some are stated sharply (problems 10, 16, 17, 19), some are more vague (problems 1, 4, 7). They are grouped into classes with a rough ordering of importance. By far the biggest single problem is that of

Validity

Validity is too low and must be increased. On the SAT, perhaps the most studied test there is, the correlation with first year college grades is generally in the .3 to .4 range. Where is the remaining .6 to .7? The answer to this problem must surely require answering three component questions. These are:

1. How can we obtain suitable validity criteria?

Any thoughtful person when asked about developing a prediction model would suggest that the first step is to get a good measure of what you are trying to predict. It might appear that without a good criterion success would be hard to determine. Although this has always been known, the amount of resources expended in the development and measurement of criteria always seems to be dwarfed by those allocated for developing predictors.

We must first ask, "Did we choose a reasonable criterion?" We can certainly boost the validity coefficient through a judicious choice of criterion. For example if we used the Graduate Record Exam (GRE) as the validity criterion for the SAT, we would see a substantial increase in the correlation of the test with the criterion. Nothing predicts a test like another test. But this is not what we usually mean by validity. College grades may be better, but they contain a hodgepodge of self-selection reflecting different criteria and different proficiencies. How to adjust for this is a major problem. But are college grades really what we want to predict? It seems to me that we ought to be less interested in first year grades than in predicting who will be a good doctor, engineer, lawyer, teacher, etc.. But how can we concretize this when we do not have a clear indicator of the variable that we wish to use as a criterion?

Consider, for example, the National Teacher Exam (the NTE). We would like this to predict who will be a good teacher. Because we do not have any single measure of the construct "good teacher" does not mean that we cannot build an index of teacher quality. To build such an index we need to collect a set of indicators that generally vary in the same direction as the unobserved construct. They need not be perfectly correlated (indeed if they were we would have found our index), but only related in the same direction in general. Some obvious candidate variables might be: teacher training (more is better), student test scores (higher is better), student gain scores (greater is better), supervisor ratings (higher is better), etc.. This does not mean that some teachers with less training are not better than others with more. Nor does it mean that there are not some extraordinary teachers whose students (for a variety of obvious reasons) do poorly. It only means that, all things being equal, this is the way that things ought to go. If this collection of indicators are all pointing in the same direction (more formally, if they form a positive manifold in the criterion space), all linear combinations of these variables will, asymptotically, become the same. This was proved by Wilks (1938). It was later shown (Wainer, 1976, 1978) that one consequence of the theorem was that under broad conditions with very few variables, an equally weighted linear composite provided prediction that was nearly as accurate as optimally chosen weights.

Wilks' theorem can be invoked easily to construct a criterion variable when none exist from correlated components. We won't have **the** index of teacher competence, but we will have another index that is highly correlated with it.[3] To build such an index we need only choose a set of suitable variables, place them onto a common scale, and add them up. Such schemes have a long history of use in the development of social indicators (e.g., Angoff & Mencken, 1931), but only recently has gotten wide attention within the area of criterion development[4].

It is well established that the criteria we use can affect in a nontrivial way the instruction that precedes it. The unity of enterprise is clear. Unless we are extraordinarily vigilant, if we assess in a particular fashion, instruction will be shifted, often to a major degree, in the direction of that assessment. If assessment in a course in French is made with a written exam, how much class emphasis will be given to verbal skills?[5] However by

[3] This may be enough for many uses. If the *Iliad* wasn't written by Homer it was produced by some other Greek of the same name.

[4] This invasion has principally been along the route of validity generalization; see Raju, Fralicx & Steinhaus (1986) for one promising result.

[5] Criticizing standardized testing for distorting the instructional process has been popular lately. Gitomer (1991, p. 2) characterizes this criticism as "If, in fact, teachers and schools are 'teaching to the test,' then we ought to develop assessments in which teaching to the test is a valid use of instructional time." The conclusion that the test must be changed is not the only one that can be reached. We might, instead, invoke sanctions to inhibit teachers from doing such a thing, if it is silly.

Even though driving tests are given within very narrowly proscribed circumstances, driving schools do not spend all their time just going over the test course. They don't, because people taking driving courses want to learn to drive and are not solely interested in passing the test. If a course in English composition focused on solving verbal analogies the teacher should be reprimanded by both students and school administrators. Just as it is impractical (too unsafe) to test driving skills on high-speed freeways, so too may be the large-scale utilization of subjectively graded essays (too expensive of time and money). A language test containing a considerable number of multiple-choice items that are abstracted from general verbal skill has not been a bad compromise, since it does provide an practical method for fair and objective assessment. If the instructional process is distorted by the measurement process, there is more than a single avenue of remediation to consider.

adopting a criterion index that is a mixture of many things such distortions of the instructional process ought to be reduced.

2. How do we best determine and measure the predictor variables that most adequately characterize the traits and proficiencies of interest?

Are the proficiencies tested by the SAT all that are required for success in college? No one I know would agree to this. What other things are needed? My colleague, Bill Angoff, has long called for a formal job analysis of what is required to be a student, but so far I do not know of anyone who has done this seriously. The literature surrounding sex differences in college performance suggests that being responsible (showing up for class regularly, handing legible home work in on time) is important. How can we include a test for such skills/propensities on an entrance test? Grit and determination are surely important. Perhaps we could get a measure of these if we follow the Chinese example and lengthen the SAT considerably. What about making a serious effort to measure some of the other personality variables that we feel are important for success? Should we include the cognitive variables that psychologists like Bob Sternberg (1985) have identified as being important?

I suspect that much of the current movement toward authentic assessment has, as its genesis, attempts to answer this question. But, so far at least, I have been disappointed in the lack of scientific rigor that has characterized many of the initial forays in this area. See problems 4 and 9.

3. How do we determine the validity of a test when it is being used for selection?

Ironically, many of the problems assessing the efficacy of educational measurement exist because the measurement is used; if a test was administered and its results ignored, we could measure its validity more directly. However tests that serve a gate-keeping role yield us a sample of only those who were admitted; tests that direct.instruction only provide us with information on those who have taken a particular instructional path.

Selection takes place at many points in the examination process. Some people may choose not to take the test. Others, after taking the test, may choose not to apply. Still others might apply but not be accepted. All we can observe are the scores of those who took it and the criterion for those who were admitted and who persevere. The nonresponse almost surely cannot be ignored; we cannot estimate the correlation between test and criterion and call it validity. How to model the nonresponse is unknown. The best we can do at the moment is to use methods of multiple imputation to estimate bounds on validity; these are likely to be too wide to be practical. Perhaps experiments like the legendary one that John Flanagan (1948) accomplished, in which the test results were ignored in all decisions, could be done. The results of such experiments would help to guide the modelling of nonresponse in the future. But experimentation is much more difficult for sequential decisions in which one test determines admission to a course of training from which further selection takes place. What would be the validity of the Graduate Record Exam if the pool of examinees was not limited to college seniors anticipating graduate study, but was instead a random sample of all seniorss? Note that the 500 subjects in the pilot sample used to assess the validity of first form of Army Alpha were broadly sampled, coming from such diverse sources as a school for the retarded, a psychopathic hospital, a reformatory, some aviation recruits, some men in an officers' training camp, 60 high school students and 114 Marines at a Navy yard.

The separation of issues into separate problems is, in a real sense, misleading. They are all of a piece, interlocked, joined, and bound. The next two major headings for problems are **data adjustment** and **data insufficiencies.** I view them as almost equally important, but I give a slight edge to the latter.

Data Insufficiencies

Some aspects of educational data collection are carefully thought out: we have SAT scores on well-equated tests for millions of individuals. Yet the individuals on whom these scores have been collected chose to participate in the testing process for reasons known only to themselves. Our knowledge about what the scores of individuals who chose not to take the test would have been is vague and incomplete. Inferences about the entire age cohort are severely limited by this self-selection. Other educational data are very much catch-as-catch-can. Validity criteria are often found lying in the street (First Year Grade Point Average in College) with little or no standardized meaning. Efforts, often valiant ones, are sometimes made to adjust grades to make them more meaningful; oftentimes these adjustments are well-meaning but circular. For example, Crouse & Trusheim (1988) use the mean SAT score of students attending a college to adjust the GPA at that particular college; higher mean SAT score implies that a B is a greater accomplishment than a B at a lower scoring school. They then use the SAT adjusted GPA as the criterion against which the judge the efficacy of the SAT. We need to do better.

There is a tradition in the social sciences of small scale studies with few replicants and many variables: the opposite of what is needed. But large scale studies with carefully gathered data and well-expressed hypotheses can have serious implications. Therefore they are rarely if ever done.

4. In the imperfect world which we inhabit, how can we collect data that are sufficient for our needs?

We can never do any better than the data we are working with. Educational data are too unstructured, too unreliable and too often lacking in validity. Moreover, there is a general misunderstanding that one can compensate for too few replicants by using more variables. More variables makes the problems of a limited number of replications worse, not better. Collecting data is expensive and too much research is done by too many people operating with small budgets. A smaller number of larger and better planned studies with careful attention to issues like power, prior hypothesis definition and control or matching is what we need. One attractive compromise would be greater cooperation among researchers that resulted in smaller studies that built on one another through formal meta analysis. This would require agreement on definitions of variables and their manifestations.

I recently had a conversation with a colleague who described, with great enthusiasm, how the state of Kentucky had recently decided (due to litigation) to completely revamp its educational system. I asked if they planned to shift over gradually so that the unshifted part might serve as a control condition. She replied that, "they didn't have a control because it was **only** an experiment." How are we to know how well anything works unless we can compare it to something else. What does "works" mean anyway?

Sometimes the "problem of insufficient data" is due to too little work, thought and/or money being devoted to the gathering of the data. These can be solved if we can convince those who control resources that the problem is worth solving and that our approach will do it. Money was allocated to run the Salk Polio Experiment, with its hundreds of thousands of subjects, because it was important to have the right answer. Billions are spent on super colliders to study elementary particles, on plasma research to study controlled fusion reactions, on space stations to study the heavens. We need to stop making due with insufficient data and insist that proper data gathering efforts can shed important light on the problems of education. To be convincing however we must first think hard about what questions must be answered and what data need to be gathered to answer them.

Let me illustrate this problem with two important subproblems.

- *How can we separate demonstrated proficiency from motivation? Are examinees poorly taught or just not trying?*

Often what we are trying to measure is "examinees' proficiency when they are really trying." If a test doesn't 'count' for a specific individual, how can we be sure that they are trying as hard as they might if it mattered? What inferences can we draw from survey data that have no consequences for the respondents? I am reminded of the experiences of the school administrators in Ojai, California a few years back. As part of the "Cash for CAP" program, the state rewarded extraordinary performance by graduating seniors on the California Assessment Program with budget augmentations that could be used for whatever school officials saw fit. Representatives from the senior class demanded that the principal allocate a certain portion of this anticipated wind-fall to subsidize some graduation festivities. The principal refused, telling them that the funds were to be used for additions to the new computer laboratory. Predictably, student performance on the test was abysmal.

Trying to get students to give their best efforts when the exam doesn't have important direct consequences to them becomes an especially thorny problem if we wish to use test items that require extended answers. Evidence from the California Assessment (Bock, 1991) indicates clearly that students are more willing to respond thoughtfully to multiple choice items than they are to items that require writing an essay. This has the ironic consequence that tests having a substantial impact on an individual student (like competitive exams for scholarships) can contain essay questions, whereas an assessment instrument that may have huge policy implications cannot. The irony is that for highly competitive exams we typically want the psychometric characteristics that are the hallmark of multiple choice exams, whereas large-scale assessments can tolerate the unreliability that is associated with the human rating of essay questions.

Perhaps there is something to be learned from the results of the recent math and science portions of the International Assessment of Educational Progress (IAEP). In it the performance of Korean students far outstripped that of all other countries. Are these students brighter than all the rest? Are they taught more competently? Do they work harder in school? Was the test, once translated into Korean, much easier than in English or French? Were they trying harder on the test than other students? Some of these hypotheses are subject to empirical verification. Student-kept diaries can shed light on study habits. Mathematical items are less subject to the effects of language, and nonlinguistic forms can be constructed. Differential racial intelligence is a silly idea. Teachers can be tested as to their subject-matter and methodological competence. But I digress. My colleague, Sung-Ho Kim observed the administration of the IAEP tests in Korea. He noted that although the students chosen to take the test were selected at random, just as in all the other countries, they were not anonymous. No individual scores were obtained, but it was made quite clear that these chosen students were representing the honor of their school and their country in this competition. To be so chosen was perceived as an individual honor and hence to give less than one's best effort was unthinkable. Contrast this with what performance you would expect from an American student hauled out of gym class to take a tough test that "didn't count."

This same problem manifests itself in spades when tests face the joint requirements of releasing all operational items after use and explicitly identifying experimental (nonoperational) sections of the test. Equating items cannot be released if they will be used to link with a future form. Yet if examinees know that they "don't count" how hard will they work on them? Small studies at ETS confirm that such equating items look more difficult than they ought to (based on independent calibration) under such circumstances. This has the effect of lowering the equated scores of all examinees, not just those who don't give their best efforts on these items.

- *How can we make longitudinal inferences from cross-sectional data?*

Many of the most important questions facing education are longitudinal. We need to accurately measure growth. Moreover growth is often over long time periods. Collecting longitudinal data over long time periods takes a long time. Drop-outs are a big problem. Dropping out is not likely to be ignorable with respect to the 'treatment' of interest. Keeping track of drop-outs is very expensive. The most common method to circumvent this is to attempt to model longitudinal change with cross-sectional data. This can lead one astray in a serious way. For example, my observations during a recent visit to North Miami Beach led me to formulate a theory of language development. When people are young they speak Spanish, when they are old they speak Yiddish. I found confirmation for this theory when I observed adolescents working in shops who spoke mostly Spanish, but also a little Yiddish. See also problem 8b.

Adjustment

Issues of statistical adjustment are rife within educational measurement. They span many arenas. I choose a few of them here.

5. How do we equate essays and other 'large' items?

If we try to build tests with large (time consuming) items while holding testing time more or less constant we are faced with a series of critical decisions. Do we use general kinds of questions ("Describe the most unforgettable character you have ever met.") or specific ones ("Characterize Kant's epistemology.")? If we do the former, it is not far-fetched to suppose that we will soon see the "Princeton Review Essay;" coaching is easy and so fairness is compromised. If we use more specific questions we may not span the domain of knowledge. In this case, an examinee may know a great deal, for example, about both Kant and about epistemology, but just lack knowledge about their intersection. Almost any other question would have elicited a full response. Is it proper to judge a student's knowledge on a small selection? If we agree that it isn't, we can have a large number of specific topics and either require answers to all of them or allow the examinee to choose. If the former we will probably have to (massively) expand testing time and cost. If the latter, how do we equate the various choices for differences in their inherent difficulty? Equating can sometimes be done by having enough common essays to provide a reliable anchor, but this runs into time and resource constraints. Equating can also be done with an anchor consisting of a number of small items. If we equate with small items we must consider whether they are testing the same thing that the large items test. If they do not, the errors of equating loom large. If they do, why are we bothering with the large items in the first place? How much does the fact that examinees select which items they will answer affect the equating? See problems 8c and 10.

6. How do we adjust for nonstandard conditions?

Sometimes one form of a test is given under nonstandard conditions. Some obvious cases of this are test forms presented in braille or orally for visually impaired examinees, or test forms given without time constraints. In these situations tests often face the dual requirements of forbidding the identification of handicapped individuals while simultaneously insisting that all tests administered under nonstandard conditions be flagged. One way to comply with such seemingly contradictory requirements might be to adjust test scores given under nonstandard conditions. This might, if done well enough, allow the comparison of individuals who have taken a test under rather different conditions. Don Rubin (1988) has suggested that under some conditions a variant of low-dose extrapolation, a statistical procedure often used in drug research, might be useful.

To illustrate this methodology, consider the research problem associated with ascertaining whether or not a particular additive is carcinogenic (so-called 'Delaney Clause Research'). For most additives the usual dosages increase cancer risk only slightly, and so to reliably estimate the size of the effect would require unrealistically large numbers of experimental animals. So instead the animals are given gigantic dosages; one group of rats might get the equivalent of 20 cases of Diet Pepsi a day, a second group 10 cases, and a third perhaps 5 cases. The incidence of cancer under each of these conditions is noted and a response surface is fitted that relates cancer incidence with dosage. This function is then extrapolated down to dosage levels more likely to be encountered under real-world conditions. This method is closely akin to 'accelerated life testing' common in engineering wherein car doors are slammed 20 times a minute, or light bulbs are burned at higher than usual voltages.

One extension of low-dose extrapolation to equating tests, where the nonstandard condition is 'unlimited' time for some subgroup, can be accomplished by experimentally varying the allotted time (one group might be given 3 hours, a second 4 hours, a third 5, a fourth 10, a fifth unlimited) and measuring to what extent additional time affects scores. Then use the fitted function to adjust scores for those individuals who had unlimited time. Such a procedure seems statistically feasible, but because the most likely effect of such an adjustment would be the adjusting downward of scores from unlimited time administrations, it would probably run into some political resistance.

While this sort of approach can help out for some problems, others, like how to compare a test in braille with a printed form remains mysterious to me.

7. How do we equate across nonoverlapping populations?

Comparing tests given under very different conditions, in which there is no possibility of someone from one group taking it under the same condition as one from the other group (i.e. a blind person taking a printed version); or if possible, (i.e. a sighted person taking a braille form) it is not credible to believe that performance is at the same level in both, brings us to a problem that is formally identical to tests in which individuals of different ethnic and language groups are compared.

Suppose we want to make comparisons among individuals and groups who have taken translated forms of the same test. The international comparisons made by the *International Assessment of Educational Progress* is but one instance. There are Spanish language versions of the SAT (the *Prueba de Aptitud Academica*). The Canadian military has been making decisions about its potential recruits on the basis of a French and English version of its screening test (*Canadian Forces Classification Battery*) for years. Entrance to Israeli universities is, in some part, determined by an applicant's performance on their "*Psychometric Entrance Tests*" which are given in Hebrew, English, Arabic, French, Spanish and Russian. Are such efforts fruitless?

Formal equating is impossible through traditional procedures (common items or overlapping samples). There is some hope that a regression procedure accomplished through a common criterion might work (i.e. score the test with the predicted score on the validity criterion based upon separate prediction equations by group), but many problems with this remain. This is not equating and thus two obvious problems are:

- *differential validity of forms* — the Israeli PET is very valid at the Hebrew University for the Hebrew language form, but much less so for the Russian form. Thus regression will make a very high score on the Russian form yield a much lower predicted grade at the university than a considerably lower score on the Hebrew form. Is this fair?

- *differential selection* — how accurate are our estimates of predicted performance when they are based on only students who actually matriculate.

I believe that some sort of 'quasi-equating' through a common criterion is the only way available at the moment, yet I am pessimistic about it working well enough on its own to be useful. And, if we cannot do it well when there is a common criterion, what can we do when even this link is missing? How are we to compare the performance of American students with that of their Japanese counterparts? Are newspaper headlines reporting that "U.S. students rank last" in some international comparison specious?

The problem arises again, although it may be a bit more subtle, when we try to equate for sex differences. There is ample evidence that tests predict differently by sex, and yet typically the scores are never equated between the sexes. Ought they be? If so, how? If we cannot deal with differential validity due to sex with American men and women who are attending the same classes in the same universities, how are we to fairly use scores obtained over the broader gulf of language and culture that I described earlier?

8. How do we correct for self-selection?

The bias introduced by self-selection is a terrible problem. It is terrible for at least three reasons. First, because it often has large effects. Second, because it is subtle and may be missed. And third, because there is no good way to correct for it statistically. The best we can do, through the method of multiple imputations, is measure the size of its effect. To illustrate the first two of these reasons consider the following four examples of wildly incorrect conclusions that were made on the basis of a self-selected sample:

(i) The *NY Times* reported (3/91) the results of data gathered by the American Society of Podiatry that 88% of all women wear shoes at least one size too small. Who would be most likely to participate in such a poll?

(ii) In 90% of all deaths resulting from bar-room brawls the victim was the one who instigated the fight. One wonders about the wit of the remaining 10% who didn't point to the floor when the police asked, 'Who started this?'

(iii) In 100 autopsies a significant relationship was discovered between age at death and length of the 'lifeline' on the palm (Newrick, Affie, & Corrall, 1990). Actually they discovered that wrinkles and age are correlated; not a breathtaking conclusion.

(iv) A Swiss physician during the time of Newton tabulated the average age of death by profession as listed on death certificate. He found that the most dangerous profession is 'student.'

I hope that this cements the size of the possible effects of selection as well as its sometimes subtle nature. Unfortunately selection looms large in many of the areas where we would like to make inferences from test scores. Three of these are:

(a) comparing political units — In the absence of a universally subscribed-to national testing program, policy makers have used various in-place tests as educational indicators. Perhaps because it is the most widely taken, or perhaps because it samples a very valuable segment of the population, SAT scores have been broadly used for this purpose. Many investigators (e.g., Dynarski, 1987; Page & Feifs, 1985; Powell & Steelman, 1984; Steelman & Powell, 1985) have tried to make inferences about the performance of state educational systems on the basis of mean SAT scores in that state. Most agree that the raw means cannot be used 'as is,' and that some adjustments must be made. There are two kinds of adjustments that are usually considered.

The first is to correct for differential participation rate. In some states a substantial proportion of the population of high school seniors take the SAT (e.g., Connecticut has 78%, New Jersey has 66%). In other states very few take the test (e.g., Iowa has a 4% participation rate, South Dakota only 3%). It is usually felt that the decision to take or not to take the SAT is, to some extent, based upon the student's anticipated success on the test. Hence nonparticipation is viewed as "nonignorable" in the sense of Rubin (1987), and correction for participation is hoped to adjust for nonignorable nonparticipation.

After correcting for differential participation, investigators will often try to adjust for differences in the demographic structure across the states. This latter adjustment is made in order to use the state SAT data to be able to make statements like, "increasing per pupil expenditures by x increases SAT scores by y."

These two kinds of adjustments are fundamentally different.

The initial attempts (cited above) to accomplish either of these adjustments were complete failures. Using sensitivity methods (multiple imputations for missing data) it became clear that the variability imposed by the self-selected aspect of these data made the results so unstable that they were worthless for any practical purpose. One more recent attempt to model just the self-selection (Taube & Linden, 1989) used a truncated Gaussian selection function. This too was a complete failure, but subsequently Edwards & Beckworth (1990) used this model to provide, for the first time, a reasoned approach to modeling self-selection on the SAT. Unfortunately this model was shown to be insufficient (Holland & Wainer, 1990a,b). I have spent some effort in the past describing the problems associated with selection and trying to point toward the structure of an acceptable solution (Wainer, 1986a,b; 1989a,b; 1990).

Much more work is required to educate the users of self-selected data about their limitations. Good models of imputation need to be established so that the errors associated with selection can be accurately determined.

(b) comparing across time — Is the performance of our public educational system declining? Do students know as much now as they used to? The average scores on college admission tests in the United States declined for more than 20 years. Why? Were students more poorly prepared? Or had education become more democratized and consequently we found a broader cross-section of American youth taking the tests? These questions are poorly phrased, for we do not want a "yes" or "no," but rather we need to be able to measure the causal effect of self-selection on average performance. How much of the decline was due to demographic changes? There have been many attempts to do this, but they have never been completely satisfying. For example, it was clearly shown that demographic shifts in the test-taking population accounted for the SAT score decline observed from 1963 until 1972 (Beaton, Hilton & Schrader, 1977). However, although these demographic shifts were essentially complete by 1972, the decline continued into the eighties.

This one example illustrates the difficulties one encounters when trying to make comparisons across time when the individuals who are measured self-select into your sample. Statistical adjustment can be done, but one can never be sure of the answer. Major trends might be visible, if those trends are large enough to dominate the variability associated with selection (and measured through multiple imputations). But suppose we wish to measure the efficacy of some sort of educational intervention. Most funders of intervention programs require some indication of positive outcome; we cannot wait several years for a strong trend to make itself visible. We must be able to detect subtle changes. But what do we do with drop-outs? With those who refuse to participate? Usually the amount of uncertainty introduced by these and other self-selection factors dwarf the size of the treatment effect that we are looking for.

(c) comparing within a test across different items — There is increasing pressure to build tests out of units that are larger than a single multiple choice item. Sometimes these units can be thought of as aggregations of small items, e.g., testlets (Wainer & Kiely, 1987; Wainer & Lewis, 1990), sometimes they are just large items (e.g., essays, mathematical proofs, etc.). Larger items, by definition, take the examinee longer to complete than do short items. As such, fewer large items can be completed within a given period of testing time. The fact that an examinee cannot complete very many large items within the usually allotted amount of testing time places the test builder in something of a quandary. One must either be satisfied with fewer items, and so open the possibility of not spanning the content domain as fully as might have been the case with a much larger number of smaller items, or expand the testing time sufficiently to allow the content domain to be sufficiently well represented. Often practicality limits testing time, and so compromises on domain coverage must be made. A common compromise is to provide several large items and allow the examinee to choose among them. The notion is that in this way the examinee is not disadvantaged by an unfortunate choice of domain coverage by the test builder.

Allowing examinees to choose which items they will answer opens up a new set of problems. Some items might be more difficult than others and so if examinees who choose different items are to be fairly compared with one another, the scores obtained on those items must be equated. How?

All methods of equating are aimed at producing the subjunctive score that an examinee would have obtained had that examinee answered a different set of items. To accomplish this feat requires that the item responses that are not observed are "missing-at-random." Why this is true becomes obvious with a little thought. The act of equating means that we believe that the performance that we observe on one item tells us something about what performance would have been on another item. If we know that the procedure by which an item was chosen has nothing to do with any specialized knowledge that the student possesses we can believe that the missing responses are missing-at-random. However if the examinee has a hand in choosing the items this assumption becomes considerably less plausible.

To understand this more concretely consider two different construction rules for a spelling test. Suppose we have a corpus of 100,000 words of varying difficulty out of which we wish to manufacture a 100 item spelling test. From the proportion of the test's items that the examinee can correctly spell we will infer that the examinee can spell a similar proportion of the total corpus. Two rules for constructing such a test might be:

- *Missing-at random* — We select 100 words at random from the corpus and present them to the examinee. In this instance we can believe that what we observe is a reasonable representation of what we did not observe.

- *Examinee selected* — A word is presented at random to the examinee, who then decides whether or not to attempt to spell it. After 100 attempts the proportion spelled correctly is the examinee's raw score. The usefulness of this score depends crucially on the extent to which we believe that an examinee's judgement of whether or not she can spell a particular word is related to her actual ability. If there is no relation between spelling ability and *a priori* expectation, then this method is as good as method (i). At the other extreme, if an examinee spells 90 words correctly all we can be sure of is that that examinee can spell no fewer than 90 words and no more than 99,990. A clue that helps us understand how to position our estimate between these two extremes is the number of words passed over during the course of obtaining the 100 sample. If the examinee has the option of omitting a word, but in fact does attempt the first 100 words that are presented, our estimate of that examinee's proficiency will not be very different than that obtained under 'missing-at-random.' If it takes 50,000 words for the examinee to find 100 to attempt we will reach quite a different conclusion. If we have the option of forcing the examinee to spell some of the words that she had previously rejected (sampling from the unselected population) we can further reduce uncertainty due to selection.

This example should make clear that the mechanism by which items are chosen to be presented to an examinee is just as crucial for correct interpretation as the examinee's performance on those items. Is there any way around this problem? How can we equate tests in which all, or some, of the items are selected by the examinee?

Other Issues

9. How can we provide the fairness of a standardized exam within the confines of 'authentic assessment' or 'portfolio assessment' or some of the other recently suggested possibilities for the improvement of traditional testing practice?

19th Century educators argued successfully for written standardized exams rather than oral exams whose contents varied with the whim of the examiner. Horace Mann (1845), to pick one outstanding example, provided eight reasons for standardized exams. Quoting him (p. 37):

> *1. They are impartial.*
> *2. They are just to the pupils.*
> *3. They are more thorough than older forms of examination.*
> *4. They prevent the officious interference of the teacher.*
> *5. They determine, beyond appeal or gainsaying, whether the pupils have been faithfully and competently taught.*
> *6. They take away all possibility of favoritism.*
> *7. They make the information obtained available to all.*
> *8. They enable all to appraise the ease or difficulty of the questions.*

Mann was obviously trying to control for inept teaching and unfair assessment by making public standardized criteria for success. His arguments sound remarkably contemporary. In his support for standardization he compared a test to a footrace (p. 39),

> "*Suppose a race were to be run by twenty men in order to determine their comparative fleetness; but instead of bringing them upon the same course, where they all could stand abreast and start abreast, one of them should be selected to run one mile, and so on, until the whole had entered the lists; might it not, and would it not so happen that one would have the luck of running up hill, and another down...? Pupils required to answer dissimilar questions are like runners obliged to test their skill by running on dissimilar courses.*"

Compare this ideal of a controlled and recorded assessment with the description of the Kentucky "experiment" described in Problem 4. Since one of Mann's purposes in insisting on standardized questions and written answers was to assess the competence of teachers (points 4 and 5), we must ask to what extent can this be done credibly if the individual teacher is responsible for making up and scoring the assessment instruments?

10. How can we provide the sufficient breadth of domain coverage required in any standardized exam that consists primarily of items that demand extensive time to answer?

Mann further argued (p. 40) for importance of as full a sampling from the content domain as possible.

> *"Again, it is clear that the larger the number of questions put to a scholar, the better is the opportunity to test his merits. If but a single question is put, the best scholar in the school may miss it, though he would succeed in answering the next twenty without a blunder; or the poorest scholar may succeed in answering one question, though certain to fail in twenty others. Each question is a partial test, and the greater the number of questions, therefore, the nearer does the test approach to completeness. It is very uncertain which face of a die will turn up at the first throw; but if the dice are thrown all day, there will be a great equality in the number of faces turned up."*

11. How can we reliably score constructed responses?

In studies of scoring subjectivity, Ruggles (1911) found that "there is as much variation among judges as to the value of each paper as there is variation among papers in the estimation of each judge." More recent experiences in the California Assessment Program (CAP) and the National Assessment of Educational Progress (NAEP) provide similar results. Bock (1991) reported that 60% of the variance of the proficiency distribution of the examinees in the CAP was due to variation among raters. During the course of the 1988 NAEP writing assessment, some 1984 essays were included in the mix for the purpose of assessing change. The differences in the scores of these essays from one assessment to the next was so large that it was deemed wise to determine change through the very expensive rescoring of a large sample of 1984 essays by the 1988 judges (Johnson & Zwick, 1988). No mere statistical adjustment was apparently sufficient.

One conclusion that we can draw from these and other, similar findings is that the use of expert judges to score constructed responses will only yield acceptable levels of accuracy when the scoring schema are so rigidly defined that one might as well be using a multiple choice test. The latter format can, of course, test the same constructs far more quickly and easily. Such findings provide ample evidence for extreme caution before considering constructed response items in any assessment of consequence. These are in addition to those concerns discussed previously, e.g. problem 4.

12. How can we best include prospective losses in the setting of passing standards?

When a test is used for gate-keeping purposes, a difficult problem is the setting of passing standards. A number of techniques have been proposed (see the Summer, 1991 issue of *Educational Measurement: Issues and Practice* for extended discussions of both methods and goals). Previous research has shown (Peterson & Novick, 1976) that unless the goals of the testing program are explicitly included in the objective function, circumstances are likely to arise that will seriously compromise the test's effectiveness. I believe that their advice is wise indeed: that in the course of determining the passing standards for any test we must explicitly include the loss function for both incorrect failure as well as improper passage. Bayesian methods pioneered by Mel Novick within the arena of educational measurement seem well suited for both the measurement of subjective loss functions and their efficient inclusion within a comprehensive testing/licensing program.

13. How can we maintain the highest possible quality of examinations within the confines of existing political and economic constraints?

Tests have consequences in addition to the enforcement of the sort of democratic meritocracy that we all envision. Test-givers operate within a political context and sometimes the political constraints introduce a structure that is at odds with what would have been the unfettered goals of a testing program. These constraints are real and important and cannot be dismissed. Similarly there are very real economic constraints that all testing programs operate under; one cannot have every essay read by 100 readers no matter how much we need the extra reliability that such extremism would yield. The challenge is to recognize these constraints and maximize the quality of the test within them.

Technical Issues

I may be betraying my own biases, but I believe that the most important contribution to testing in the 20th century has been the development of formal statistical models for both the scoring of tests and the characterization of their quality. This development parallels the growth of statistical knowledge. It is no coincidence that many of the major innovators in 20th century statistics are also major contributors to psychometrics. The names Pearson, Spearman, Hotelling, Thurstone, and Wilks dominate statistical and psychometric thought in the first half of the century; Tukey, Bock, Lord, Novick, Holland and Rubin are found with frequency on the pages of both *Psychometrika* and the *Journal of the American Statistical Association* in the second half. Because I have so far avoided a detailed discussion of technical issues is not meant to suggest that all such problems have been resolved. Rather I did not want to distract attention from profound conceptual problems. It is too easy to bring the formidable powers of modern statistics and computation to bear on some smaller problem and lose track of the soft underbelly.

As a brief example of how easy it is to do this, consider the considerable body of often ingenious work on computerized test construction. Recent work has overcome horrible problems of integer programming with multiple inequality constraints to build a test form in seconds that would take a human test developer several hours. But test developers are not impressed. Why? Michael Zieky, a veteran test developer at ETS, has pointed out to me that even here, a hot bed of testing, no one makes up a test more often than once every six to eight weeks. And so the jillion or so dollars spent developing a computerized test builder saves such a test developer a few hours a month. Most of the rest of the time is spent writing items. If a computer algorithm could do **that** it would be a big help. But instead computers are used to do what they have done in the past. Real progress will only be made when they can begin to do what only humans seem to be able to do now.

With this overlong preamble explaining why I have relegated technical issues to the end let me get them out of the way in a hurry.

14. How can we automate (computerize) the writing of high quality test items?

I believe that before we can do this we need to solve a variety of component problems. Principal among these is,

15. What makes an item hard?

This is perhaps an unfairly overshort statement of the problems associated with connecting psychometric practice with contemporary knowledge of cognitive skills and learning. Yet, I believe that if we understood what makes any particular item as hard as it turns out to be empirically, we will have nmade a real advance. In some areas limited advances have been made. Susan Embretson and Gerhard Fisher have provided us with some psychometric models and psychological insight that help, but we are still a long way from a full enough understanding to be able to apply this knowledge broadly. Kikumi Tatsuoka has shown us how we can make pretty good estimates of the relative difficulty of certain kinds of arithmetic problems.Readability formulas for descriptive prose seem to be highly correlated with item difficulty for certain kinds of tasks (Koslin, Zeno & Koslin, 1987). But what about in the testing of physics, or chemistry, or plumbing? My aging cynicism may be showing, but I am not sanguine about major advances in this area in the near future.

16. How can we marginalize high dimensional distributions for Bayes estimation when there are no closed forms?

The estimation of parameters of IRT models is now done through marginal maximum likelihood (Bock & Aiken, 1981). This approach is clearly maximizing the 'right' likelihood, but exactly how to accomplish this in higher dimensions is not clear. Bock (personal communication, June 17, 1991) reports progress, but a general solution is still needed.

17. What is the best way to estimate latent distributions (parametric, semi-parametric, non-parametric)?

Latent trait theory has as its most fundamental component one or more hypothesized underlying traits. An examinee's position on these traits is never observed, but only inferred from observed performance on a set of items. In assessments we are usually interested in estimating the latent distribution of these traits. The under-identification of the model (a trade-off involving item characteristics and examinee characteristics) makes this problem very difficult indeed.

18. How can we combine response time with other measures of quality of response?

Discussions of the relative virtues of speeded versus power tests extend back for more than 50 years (Paterson & Tinker, 193[illegible]). Advice about how to measure the psychometric characteristics of speeded tests was prominent in the literature of the early 1950s (e.g. Gulliksen, 1950; Cronbach & Warrington, 1951). The general consensus was, and remains, that there is no such thing as a pure power test; all tests were, to some extent, speeded for some examinees. With the widespread availability of inexpensive computing arose the possibility of computer administered exams. One concomitant of this is that "amount of time required to answer" became an additional piece of data that was easily obtained. How can we use this information? Many believe that if two examinees answer the same question correctly the one who could do so in a shorter period of time is entitled to a higher proficiency estimate. But this is less credible if they both got it wrong. Moreover, the trade-off between speed and accuracy is well documented in vast numbers of situations. Response time tells us something, but what? Bruce Bloxom (1985) provides an up-to-date description of the technical issues involved in the modeling of response time, although how to combine response time with other measures of quality of response (like "did you get it right?") remains unclear. Thissen (1983) summarizes work in this area as well as proposing a generalization to Furneaux's (1961) model. It seems to me that a solution to this problem requires experimentally limiting examinee control of response time. If examinees can take as long as they like many variables can enter into the mix. If, instead, we experimentally assign examinees to several time limit categories we can estimate the relationship between time and performance in a less contaminated atmosphere (see the discussion of Rubin's plan to use such a scheme described in Problem 6).

19. What is the right level of complexity of psychometric models?

I often hear, during the course of methodological presentations, the statement, "It is hard to find good data for this model." This comes up often with structural equation models, but it happens a lot with other sorts as well. One might naively then think, "we must learn to gather better data." Indeed, if the measurement model is providing us with vital information perhaps we ought to work harder at data gathering (problem 4). But if we need the population of the entire Indian subcontinent to obtain sufficiently stable parameters to draw useful conclusions, we must rethink our model.

Conclusion

"The problems mentioned are merely samples of problems, yet they will suffice to show how rich, how manifold and how extensive the ... science of (measurement) today is, and the question is urged upon us whether (it) is doomed to the fate of those other sciences that have split into separate branches, whose representatives scarcely understand one another, and whose connection becomes ever more loose. I do not believe this nor wish it. (Measurement) is in my opinion an indivisible whole, an organism whose vitality is conditioned upon the connection of its parts. For with all the variety of ... knowledge, we are still clearly conscious of the similarity of the logical devices, the relationship of the ideas in (measurement) as a whole and the numerous analogies in its different departments. We also notice that, the further a ...theory is developed, the more harmoniously and uniformly does its construction proceed, and unsuspected relations are disclosed between

hitherto separate branches of the science. So it happens that, with the extension of (our science), its organic character is not lost but only manifests itself the more clearly."

I end on a pessimistic note. Many of the problems discussed here require, indeed demand, substantial technical expertise on the part of the investigator for their solution. Unfortunately, it appears that instruction in these technical areas is declining within the social sciences (Aiken et al, 1990). This contributes to, as well as reflects, the interests and talent of the students. Consequently I will number, as the last of our principal problems

20. How do we attract, train, and keep the technically talented individuals who are needed?

David Hilbert suggested, almost a century ago, that although measurement is the foundation of all exact knowledge, in order for it to completely fulfill its high mission, the new millennium must

"*bring it gifted masters and many zealous and enthusiastic disciples*."

We must work harder to find and keep both.

References

Aiken, L. S., West, S. G., Sechrest, L., & Reno, R. R. (1990). Graduate training in statistics, methodology, and measurement in psychology. *The American Psychologist, 45,* 721-734.

Angoff, C., & Mencken, H.L. (1931). The worst American state. *American Mercury, 31,* 1- 16, 175-88, 355-71.

Beaton, A. E., Hilton, T. L., & Schrader, W. B. (1977). *Changes in the verbal abilities of high school seniors, college entrants, and SAT candidates between 1960 and 1972.* Princeton, NJ: Educational Testing Service.

Bloxom, B. (1985). Considerations in psychometric modeling of response time. *Psychometrika, 50,* 383-397.

Bock, R. D. (1991). "The California Assessment." A talk given at the Educational Testing Service, Princeton, N.J. on June 17, 1991.

Bock, R.D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: An application of an EM algorithm. *Psychometrika, 46,* 443-459.

Browder, F. (ed.) (1974). *Mathematical Developments Arising from the Hilbert Problems.* Symposia in Pure Mathematics, Vol. 28.

Cronbach, L. J., & Warrington, W. G. (1951). Time-limit tests: Estimating their reliability and degree of speeding. *Psychometrika, 16,* 167-188.

Crouse, J., & Trusheim, D. (1988). *The Case against the SAT.* Chicago: University of Chicago Press.

Dynarski, M. (1987). The Scholastic Aptitude Test: Participation and performance. *Economics of Education Review, 6,* 263-273.

Edwards, D., & Beckworth, C. M. (1990). Comment on Holland & Wainer's "Sources of uncertainty often ignored in adjusting state mean SAT scores for differential participation rates: The rules of the game." *Applied Measurement in Education, 3,* 369-376.

Flanagan, J. C. (1948). *The Aviation Psychology Program in the Army Air Forces.* Report 1, AAF Aviation Psychology Program Research Reports, U.S. Government Printing Office. Pp. xii+316.

Furneaux, W.D. (1961). Intellectual abilities and problem solving behavior. In H.J. Eysenck (Ed.), *The handbook of abnormal psychology* (pp. 167-192). London: Pittman.

Gitomer, D. H. (1991). Performance assessment and educational measurement. Unpublished manuscript, Princeton, NJ: Educational Testing Service.

Gulliksen, H. O. (1950). The reliability of speeded tests. *Psychometrika, 15,* 259-269.

Hilbert, D. (1902). Mathematical problems. *Bulletin of the American Mathematical Society, 8,* 437-479.

Holland, P. W., & Wainer, H. (1990a). Sources of uncertainty often ignored in adjusting state mean SAT scores for differential participation rates: The rules of the game. *Applied Measurement in Education, 3 (2),* 167-184.

Holland, P. W., & Wainer, H. (1990b). Edwards & Cummings' "Fuzzy Truncation Model" is a step in the right direction. *Applied Measurement in Education, 3(4),* 377-380, 1990.

Johnson, E. G., & Zwick, R. J. (1988). *The NAEP Technical Report.* Princeton, N.J.: Educational Testing Service.

Koslin, B. L., Zeno, S., & Koslin, S. (1987). *The DRP: An effectiveness measure in reading.* New York: College Entrance Examination Board.

Mann, H. (1845). *A description of a survey of the Grammar and Writing Schools of Boston in 1845.* Quoted in O. W. Caldwell & S. A. Courtis (1923). *Then and now in education,* Yonkers-on-Hudson, New York: World Book Company, p. 37-40.

Newrick, P. G., Affie, E., & Corrall, R. J. M. (1990). Relationship between longevity and lifeline: A manual study of 100 patients. *Journal of the Royal Society of Medicine, 83,* 499-501.

Page, E. B., & Feifs, H. (1985). SAT scores and American states: Seeking for useful meaning. *Journal of Educational Measurement,* 22, 305-312.

Paterson, D. G., & Tinker, M. A. (1930). Time-limit vs. work-limit methods. *American Journal of Psychology, 42,* 101-104.

Petersen, N. S. & Novick, M. R. (1976). An evaluation of some models for culture-fair selection. *Journal of Educational Measurement, 13,* 3-29.

Powell, B., & Steelman, L. C. (1984). Variations in state SAT performance: Meaningful or misleading? *Harvard Educational Review, 54,* 389-412.

Raju, N. S., Fralicx, R., & Steinhaus, S. D. (1986). Covariance and regression slope models for studying validity generalization. *Applied Psychological Measurement, 10.* 195-211.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. New York: John Wiley & Sons.

Rubin, D. B. (1988). Discussion. In H. Wainer and H. Braun (Eds.), *Test Validity* (pps. 241-256) Hillsdale, N.J.: Lawrence Erlbaum Associates.

Ruggles, A. M. (1911). *Grades and grading*. New York: Teacher's College.

Steelman, L. C., & Powell, B. (1985). Appraising the implications of the SAT for educational policy. *Phi Delta Kappan*, 603-606.

Sternberg, R. J. (1985) *Beyond IQ: A triarchic theory of human intelligence*. New York: Cambridge University Press.

Taube, K. T., & Linden, K. W. (1989). State mean SAT score as a function of participation rate and other educational and demographic variables. *Applied Measurement in Education, 2,* 143-159.

Thissen, D. (1983). Timed testing: An approach using item response theory. In D. Weiss (Ed.), *New horizons in testing: Latent trait test theory and computerized adaptive testing* (pp. 179-203). New York: Academic Press.

Wainer, H. (1976). Estimating coefficients in linear models: It don't make no nevermind. *Psychological Bulletin, 83*, 213-217.

Wainer, H. (1978). On the sensitivity of regression and regressors. *Psychological Bulletin, 85*, 267-273.

Wainer, H. (1986a). The SAT as a social indicator: A pretty bad idea. In H. Wainer (Ed.) *Drawing inferences from self-selected samples* (pp. 7-21). New York: Springer-Verlag.

Wainer, H. (1986b). Five pitfalls encountered while trying to compare states on their SAT scores. *Journal of Educational Statistics, 11,* 239-244.

Wainer, H. (1989a). Eelworms, bulletholes and Geraldine Ferraro: Some problems with statistical adjustment and some solutions. *Journal of Educational Statistics, 14,* 121-140.

Wainer, H. (1989b). Responsum. *Journal of Educational Statistics, 14,* 187-200.

Wainer, H. (1990). Adjusting NAEP for self-selection: A useful place for "Wall Chart" technology? *Journal of Educational Statistics, 15*, 1-7.

Wainer, H., & Kiely, G. L. (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational Measurement, 24,* 185-201.

Wainer, H., & Lewis, C. (1990). Toward a psychometrics for testlets. *Journal of Educational Measurement, 27*, 1-14.

Wilks, S. S. (1938). Weighting systems for linear functions of correlated variables when there are no dependent variables. *Psychometrika, 3*, 23-40.