Journal of Official Statistics
Vol. 9. No. 1, 1993, pp. 189–201
© Statistics Sweden

Some Research Problems Encountered at the Educational Testing Service

Howard Wainer, Eugene G. Johnson, Charles Lewis, and Robert J. Mislevy[1]

Abstract: The Educational Testing Service is the largest private testing organization in the world. In the course of fulfilling its mission many perplexing statistical problems are encountered. In this article each of four ETS statisticians describe a problem that currently engages their attention.

Key words: Nonignorable nonresponse; Bayesian estimation; choice; missing data; Hilbert.

1. Introduction

The Educational Testing Service (ETS) was founded in 1947 by the American Council on Education, the Carnegie Foundation for the Advancement of Teaching, and the College Board. The primary purpose of ETS is to serve and improve education through development and use of high quality measurement procedures and carefully performed research and related services.

ETS conducts research on measurement theory and practice, teaching and learning, and educational policy. Research at ETS has four essential missions:

1. Basic research, embracing both the technical and the substantive foundations of educational measurement, conducted in support of the goals of ETS and its clients.
2. New product research which currently emphasizes innovative uses of technology in support of education and measurement, and the development of assessment techniques that contribute to more effective teaching and learning.
3. Research to enhance and maintain the technical quality of tests including methodological, psychometric, and statistical studies.
4. Public service research provides program evaluation for a variety of clients and is also involved with policy research. Policy research deals with the implications of judicial and legislative actions and with issues of access and equity for women and minorities.

In this paper each author describes what he considers some of the most vexing problems. While these problems are not all completely statistical, they all have major statistical components.

In the first section Charles Lewis describes a technical problem that occurs in taking a Bayesian approach to traditional

[1] Educational Testing Service, Rosedale Road, Princeton, NJ, 08541, U.S.A.

Acknowledgment: The authors gratefully acknowledge support from the Graduate Record Board and the National Center for Education Statistics, Office of Educational Research and Improvement. This article has benefited from conversations with Henry Braun and Bill Ward.

test theory. This problem seems to cut to the core of the sorts of models he describes. These linear models (so-called 'true score models') are the basis of most test scoring schemes and so the problem he describes has analogs in many other fields.

In the second section Robert J. Mislevy explains the growing dissatisfaction with models of the sort described previously, and points toward the need for a broader outlook. This broader viewpoint has yet to be rigorously characterized; such rigor is sorely needed to avoid serious errors of interpretation that seem too often to sneak into current educational measurement.

The third section, by Howard Wainer, discusses the general problem of nonignorable nonresponse in one circumstance, specifically an increasingly popular innovation in testing practice: allowing examinees to choose to answer only a small number of test items from a larger selection. He points out the pitfalls of such a practice and laments its consequences.

The fourth section, by Eugene G. Johnson, describes the inferences that are occurring within the ongoing American educational survey called the "National Assessment of Educational Progress" (NAEP) due to nonignorable nonresponse. By law, students and schools may opt to not participate in the assessment. The problems caused by nonresponse and the current methods of adjustment are discussed. In response to the educational reforms described by Mislevy and Wainer, NAEP uses new testing methodologies. Johnson describes some of these and indicates some statistical issues they engender.

2. Problems with the Simultaneous Estimation of Many True Scores

A basic goal of psychometric theory is to make inferences about assumed underlying states of knowledge that individuals possess on the basis of their observed behavior. In the so-called classical theory of mental tests (see Lord and Novick (1968) for a complete treatment of the subject), observed test scores are described as the sum of a true score and an error score

$$X = T + E \tag{1}$$

with, by the definition of T,

$$E(X \mid T) = T, \tag{2}$$

so that the variance of the observed scores may be expressed as the sum of the variances of the components

$$\sigma_X^2 = \sigma_T^2 + \sigma_E^2 . \tag{3}$$

Moreover, it is common to assume that T and E have independent normal distributions (with means μ and 0, respectively). In other words, a one-way, random effects analysis of variance model is adopted for observed test scores.

In addition to the usual interest in making inferences about the variance components σ_T^2 and σ_E^2, mental test theory focuses attention on individual true scores. In the model just described, the conditional distribution of T, given X, is normal with

$$\mathrm{E}(T|X) = \frac{\sigma_T^2 X + \sigma_E^2 \mu}{\sigma_T^2 + \sigma_E^2} \tag{4}$$

and

$$\sigma^2(T|X) = \frac{\sigma_T^2 \sigma_E^2}{\sigma_T^2 + \sigma_E^2} . \tag{5}$$

Equations 4 and 5 assume that the mean and variance components are known. A standard practice, sometimes referred to as Empirical Bayes estimation (Braun 1989), has been to estimate μ, σ_T^2 and σ_E^2, and insert these estimates into Equations 4 and 5 as the basis for inferences about true scores, given observed test scores.

To address this problem more formally,

we may adopt a Bayesian framework (following, for instance, Novick, Jackson and Thayer 1971; see Lewis 1989, for additional references and discussion). The first step is to obtain a posterior distribution for a set of true scores and the unknown population parameters, given a set of observed scores. For purposes of this discussion, let us restrict our attention to the case of one test score per individual, and assume that σ_E^2 is known. With a vague prior density for μ and σ_T^2, and a total of m individuals, the posterior density for all parameters has the form

$$p(\mathbf{T},\sigma_T^2,\mu|\mathbf{X}) \propto (\sigma_T^2)^{-m/2} \times \exp\left[-\sum_{i=1}^{m}(X_i - T_i)^2/(2\sigma_E^2) - \sum_{i=1}^{m}(T_i - \mu)^2/(2\sigma_T^2)\right]. \qquad (6)$$

As m increases, our knowledge about μ and σ_T^2 becomes more precise, and the posterior mean and variance of T_i approach the expressions given in Equations 4 and 5 (Box and Tiao 1973). To further study the posterior density given in Equation 6, it will be useful to consider its joint mode. First, it may be shown that the modal values of the T_i will have the form given in Equation 4, with the modal estimates of μ and σ_T^2 substituted for the true values

$$\bar{T}_i = \frac{\tilde{\sigma}_T^2 X_i + \sigma_E^2\tilde{\mu}}{\tilde{\sigma}_T^2 + \sigma_E^2}. \qquad (7)$$

From Equation 6, it also follows that the modal estimate of μ equals the mean of the modal values of the T_i, which in turn equals the mean of the observed test scores

$$\bar{\mu} = \bar{X}. \qquad (8)$$

Substituting the values from Equations 7 and 8 into Equation 6 and simplifying yields the following function of σ_T^2, which may be maximized to obtain the joint mode of the posterior density:

$$g(\sigma_T^2) = (\sigma_T^2)^{-m/2} \times \exp\left[-\frac{1}{2}\frac{\sum_{i=1}^{m}(X_i - \bar{X})^2}{\sigma_T^2 + \sigma_E^2}\right]. \qquad (9)$$

There is, however, a problem with Equation 9, namely that

$$\lim_{\sigma_T^2 \to 0+} g(\sigma_T^2) = \infty. \qquad (10)$$

In other words, the joint posterior density may be made arbitrarily large with sufficiently small values of σ_T^2. The corresponding limiting modal estimates for the T_i may be derived from Equation 7 as

$$\bar{T}_i = \bar{X} \qquad (11)$$

for all i, regardless of the value of the observed score X_i. The estimates given in Equation 11, sometimes referred to as completely regressed or pooled estimates, obviously have no practical value for the reporting of test results.

It may be useful to take a closer look at the behavior of the joint posterior density as a function of σ_T^2. Consider, as an example, a case with 100 observed test scores, whose sample variance is 5.0, and for whom the error variance is known to be 1.0. Based on Equation 3, a consistent estimate of true score variance would be 4.0. Figure 1 shows the form of $\log_{10} g(\sigma_T^2)$ for σ_T^2 between 0.0 and 10.0. Besides illustrating the result in Equation 10, namely that the density increases without limit as σ_T^2 approaches zero, Figure 1 shows a secondary mode at 2.6, with a density almost ten times that in the neighborhood of 4.0. Thus, ignoring the limiting behavior of the joint posterior and restricting attention to interior modes still produces values for the true scores in this example which show substantially more regression than would be expected on the basis of the posterior means of the T_i.

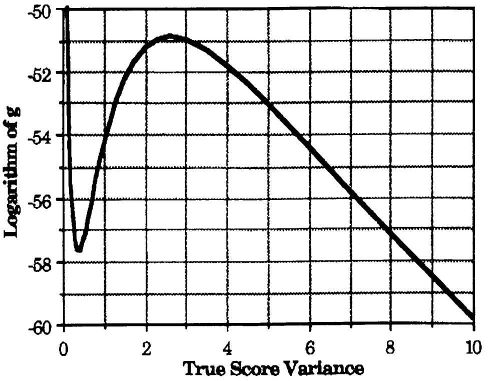

Fig. 1. Joint posterior density with 100 examinees, sample variance 5, and error variance 1.

To study interior modes more generally, one may set the derivative of the logarithm of g with respect to σ_T^2 equal to zero and solve for σ_T^2. The result may be written as

$$\bar{\tilde{\sigma}}_T^2 = s_X^2 \left(\frac{1 + \sqrt{1 - 4\sigma_E^2 / s_X^2}}{2} \right) - \sigma_E^2 \tag{12}$$

where s_X^2 has been used to denote the sample variance of the observed test scores. Of course, this expression can only be evaluated when

$$s_X^2 \geqslant 4\sigma_E^2 . \tag{13}$$

For smaller values of the sample variance, there is no interior mode for the joint posterior density and, consequently, no alternative to the complete regression mode.

Assuming Inequality 13 holds, we may compare the result given in Equation 12 with the alternative

$$\hat{\sigma}_T^2 = s_X^2 - \sigma_E^2 . \tag{14}$$

It is clear that Equation 14 will always produce the larger of the two values, resulting in less regression for the corresponding true score values. In fact, as s_X^2 approaches the lower limit given in Inequality 13, $\hat{\sigma}_T^2$ will approach three times $\bar{\tilde{\sigma}}_T^2$. For larger values of s_X^2, the difference between the two estimates approaches σ_E^2. In other words, for $s_X^2 \gg 4\sigma_E^2$,

$$\bar{\tilde{\sigma}}_T^2 \doteq s_X^2 - 2\sigma_E^2 . \tag{15}$$

What is perhaps most disturbing about these results is that they are unaffected by the number of observed test scores being analyzed. That is, although $\hat{\sigma}_T^2$ converges in probability to the true value of σ_T^2 as the number of test scores increases without limit, the same cannot be said of $\bar{\tilde{\sigma}}_T^2$. Consequently, the joint posterior modal estimates of the true scores – even if an interior mode exists and we restrict our attention to it – do not approach the 'true' regression estimates given in Equation 4.

In conclusion we may say that Empirical Bayes estimates of true scores are clearly appropriate and receive asymptotic support from the behavior of the posterior true score means. Nonetheless, the asymptotic properties of the posterior mode and, hence, of the joint posterior density, demonstrate that there are still gaps in our understanding of the inferential foundations for this important problem in mental test theory.

3. Test Theory Reconceived

3.1. *Introduction*

Educational test theory is a corpus of concepts, models, and methods for making inferences about students' proficiencies. The principles of statistical inference are thus brought to bear on practical problems in selection, instruction, and evaluation. Recent decades have witnessed considerable progress in the development of models and methods, but the conceptual foundations have advanced precious little in the past century. The problem: The standard models of test theory evolved to address problems cast in the psychology of the first half of the

Twentieth Century. They fall short for solving the range of problems cast in the emerging view of how people think, learn, and solve problems. The challenge: To extend test theory to a broader family of student models, in directions indicated by recent developments in cognitive and educational psychology.

3.2. *The evolution of standard test theory*

The conceptual foundations of standard test theory are found in the psychology of Charles Spearman (e.g., 1904) and L.L. Thurstone (e.g., 1947). A person is characterized by a small number of real-valued variables, "traits," that drive the probabilities of his or her observable responses in specified settings. In educational testing, for example, the student model is typically a single variable, say, *ability*, and the observations are responses to test items, assumed conditionally independent given ability. Gulliksen (1961) described the central problem of test theory as "the relation between the *ability* of the individual and his *observed score* on the test" (p. 101; emphasis original).

The paradigm of trait psychology suits the mass educational systems that arose in the United States at the turn of the century, and dominate practice yet today. 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. This problem context encourages one to build student models around abilities that are few in number, broadly construed, stable over time, applicable over wide ranges of students, and discernible by data that are easy to gather and analyze.

Pointing to Lord and Novick's (1968) *Statistical theories of mental test scores* as a watershed event, Lewis (1986) stated that "much of the recent progress in test theory has been made by treating the study of the relationship between responses to a set of test items and a hypothesized trait (or traits) of an individual as a problem of statistical inference" (p. 11). Indeed, we note the appearance of sophisticated estimation procedures (e.g., Bock and Aitkin 1981), hierarchical modeling techniques (e.g., Muthén and Satorra 1989), approaches for test theory based on missing-data theory (e.g., Mislevy 1991), and theoretical advances into latent-variable modeling (e.g., Holland and Rosenbaum 1986). All of these achievements, however, remain largely within the paradigm of trait psychology.

3.3. *The cognitive revolution*

Recent decades have also witnessed a paradigmatic revolution in the psychology of learning and cognition. The emphasis has shifted *away* from the characterization of stable, universally applicable traits, revealed in the same way for all subjects by constrained responses in standardized observational settings, and *toward* an understanding of how individuals organize and update knowledge, how they bring that knowledge to bear in meaningful problems. Learners increase their competence not by simply accumulating new facts and skills, but by reconfiguring their knowledge structures ("schemas"), 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 cognitive paradigm shapes conceptions of how to increase students' competencies, to help them develop knowledge bases that are increasingly *coherent, principled,* *useful*, and *goal-oriented* (Glaser 1991, p. 26). The implications for test theory that follow might be best introduced by an analogy from physics.

3.4. On the theory of bridge design

A hundred years ago, civil engineers designed bridges in accordance with Newton's laws and Euclid's geometry, in the prevailing belief that these models were an accurate description of the true nature of the universe. The quantum and relativistic revolutions shattered this paradigm. Nevertheless, today's civil engineers still design bridges according with the same approach. What's different?

First, even though the same formulas are employed, they are comprehended from the perspective of the new physical paradigm. The formulas through which the bridge is designed and constructed are no longer thought of as approximations departing from truth only by measurement error, but as engineering tools useful for addressing the problem at hand. The bridge is neither so small as to require modeling quantum effects, nor so massive or fast moving as to require relativistic effects.

Secondly, today's civil engineers work with materials that did not exist a hundred years ago, with strengths, flexibilities, and durabilities tailored through modern metallurgy using, in part, concepts from quantum physics. Even though the same bridge-building theory is employed, the materials and the products are improved in ways unanticipated in the previous paradigm.

Finally, while civil engineers continue to solve problems that arose under the previous paradigms using the still-useful formulas of Newtonian physics, albeit more effectively, other scientists and engineers in fields that did not exist last century are attacking problems that could not even be conceived of then – problems in superconductivity, microchip design, and fusion research, as examples.

3.5. On the theory of educational tests

I see the same multiple paths of progress for educational test theory, to support educational inference and decision-making from the perspective of contemporary psychology. The role of the statistician is working with the educational and cognitive psychologist to develop useful student models that express the key aspects of knowledge and proficiency, and support defensible and cost-effective statistical inference in practical settings. My comments fall in the realms suggested by the preceding analogy.

First, educational testing for large-scale selection and placement decisions will continue to be useful when resources dictate constraints similar to those that originally spawned standard test theory. These applications must be re-examined in the conceptual framework of the new paradigm – to be re-justified from new premises, revised so that they can be, perhaps abandoned if they cannot. That an application falls into the last of these categories signals not a failure of test theory to accomplish what it was designed to do, but an inadequacy of the conceptual framework from which the decision-making alternatives were derived.

An example: The Scholastic Aptitude Test (SAT), a multiple-choice test comprised of verbal and quantitative reasoning items, was introduced in 1926 to help colleges select among applicants. The goal was "measuring verbal aptitude and . . . mathematical aptitude" (Angoff and Dyer 1971, p. 2), purportedly to identify those with high enough trait values to succeed. A strong predictive relationship was sufficient to justify the test. A wide diversity of students, varying in cultural backgrounds, educational experiences, and personal qualities, can obtain the same SAT score, however. Does the same score convey the same information about each? On the basis of these scores, should the same inference or the same decision be made about all? Maybe, but to justify the program today, one must demonstrate a direct relation between performance and relevant skills; for example, showing that difficulty of verbal test items depends on features of items linked to elements in theories of comprehension (Scheuneman, Gerritz, and Embretson 1991).

Secondly, testing may still employ overall proficiency measurement model, but with different raw materials – i.e., types of observations. Rather than obtaining information with more easily processed multiple-choice items, for example, an assessment may require examinees to solve more complex multi-step tasks, to formulate problems rather than solve problems presented by the examiner, or to carry out an extended project at least partly of the student's own choosing. The rationale, again, is that correlation alone is not enough. Students do not increase proficiency by "increasing their trait values," but by studying, learning, and practicing particular content and skills; the content and skills tests assess are the content and skills teachers teach. Methods of data collection deemed inefficient under the trait paradigm gain currency when viewed as more direct indicators of the desired outcomes of learning (see, e.g., Frederiksen and Collins 1989). In this arena, test theory must develop observational models and inferential procedures to connect trait-based student models with a broader range of observations.

An example: The College Board's Advanced Placement (AP) Studio Art test differs from standard tests in that students present for evaluation a portfolio of works they develop during the course of instruction. An outline of requirements is specified, but, in order to elicit evidence about the process of developing proficiency as an artist, students are necessarily provided almost unbridled choice in the specific projects they undertake. Together, experts in art and statisticians have developed a framework for evaluating portfolios along performance scales; they must continue, using statistical methodology, to refine systems by which judges can monitor, control, communicate, and improve their procedures for ratings of complex performances.

Finally, statisticians interested in educational applications must work with psychologists and educators to develop workable models for applications that are not addressed by standard test theory. In this vein are more detailed models of aspects of student knowledge, for the purposes of immediate, short-term educational decisions. The important questions become not "How many items did this student answer correctly?" but, in Thompson's (1982) words, "What can this person be thinking so that his actions make sense from his perspective?"

An example of this type is a computerized intelligent tutoring system (ITS; see, e.g., Lesgold, Lajoie, Bunzo, and Eggan 1988). In an ITS, students learn concepts and practice problems while the tutor continuously updates a student model. The status of the student model drives short term instructional decisions, for hints, direct instruction, or problem selection. The psychology of learning in the domain determines the nature of the student model. Test theory, broadly construed, connects observations to the student model. Statistical principles serve as the foundation for inference and decision-making.

3.6. *Conclusion*

The cognitive revolution in psychology challenges the very premises upon which educational testing was founded. Reconceiving test theory, to tackle both old problems as viewed from the new paradigm and new problems that did not previously exist, is a task that demands the creative efforts, in concert, of theoreticians, educators, and, I would submit, statisticians.

4. Allowing Examinee Choice in Exams

There is a growing movement in education to radically change the structure of standardized exams. This movement has grown from a dissatisfaction with the results obtained from the kind of standardized exams currently in use. These exams are typically composed of a substantial number (commonly between 50 and 100) of multiple choice items[2]. This dissatisfaction spans many areas, but is principally focused on the perceived molecular nature of multiple choice items. Many of the complainants express a preference for larger, more "authentic" items[3].

It has long been understood that a good test must contain enough questions to cover fairly the content domain. In his description of an 1845 survey of the Grammar and Writing Schools of Boston, Horace Mann argued that

". . . 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."

Despite the force of Mann's argument, pressure continues to build to make tests from 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 and Kiely 1987; Wainer and Lewis 1990); sometimes they are just large items (e.g., essays, mathematical proofs, etc.). Large items, by definition, take the examinee longer to complete than do short items. Therefore, fewer large items can be completed within the given testing time.

The fact that an examinee cannot complete very many large items within the allotted testing time places the test builder in something of a quandary. One must either be satisfied with fewer items, and possibly not span the domain of material that is to be examined 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 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 placed at a disadvantage by an unfortunate choice of domain coverage by the test builder.

Allowing examinees to choose the items they will answer presents a difficult set of problems. Despite the most strenuous efforts to write items of equivalent difficulty, some are inevitably more difficult than others. If examinees who choose different items are to be fairly compared with one another, a basis for that comparison must be established. How might that be done?

[2] A "multiple choice item" is a question paired with several possible answers. The most usual task for an examinee is to choose the best option from among those offered.

[3] "Authentic" here is used to mean items that more closely resemble the real world tasks that the test is supposed to be predicting.

This question is akin to the question answered by equating in traditional methods of test construction in which different forms of a test are prepared[4] and administered at random to different segments of the examinee population.

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 unobserved item responses are "missing-at-random."[5] The act of equating means that we believe that the performance that we observe on one test form (or item) tells us something about what performance would have been on another test form (or 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. There is an important difference between examinee-chosen data and the data usually used to equate alternate forms – the latter have data missing by the choice of the examiner, not the examinee.

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, and we wish to create a 100-item spelling test. From the proportion of test's items that the examinee correctly spells 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 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 examinees' judgments of whether or not they can spell particular words are related to actual ability. If there is no relation between spelling ability and *a priori* expectation, then this method is as good as missing-at-random. At the other extreme, we might believe that examinees know perfectly well whether or not they can spell a particular word correctly. In this instance a raw score of 100% has quite a different meaning. Thus, if an examinee spells 90 words correctly all we know with certainty is that the 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 sample of 100. If the examinee has the option of omitting a word, but in fact attempts the first 100 words 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 previously rejected words (sampling

[4] Usually strenuous efforts are made to make these various forms as identical to one another in content and difficulty as possible. An equating is considered successful if a fully informed examinee is indifferent as to which form she will receive.

[5] The random assignment of forms to individuals makes the assumption of 'missing-at-random' credible.

Table 1. Mean performance scores

	Group A	Group B
Part I	11.7	11.2
Part II		
Question 1	8.2	
Question 2		2.7

from the unselected population), we can reduce uncertainty due to selection.

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

A brief example of the size of the possible effects that need to be adjusted for may be of help. Fremer, Jackson and McPeek (1968) report on one chemistry examination in which there were two parts. Part I consisted of a set of multiple choice questions that everyone was required to answer. Part II allowed the examinee to choose between two large questions. The choice divided the examinees into two groups; (A) those who chose to answer Question 1 on Part II and (B) those who chose to answer Question 2. They found that although there was essentially no difference between these two groups in their performance on Part I there was an enormous difference on Part II. The mean scores are shown in Table 1.

How should we interpret these results? We might believe that whatever is being tested in Part II is quite different than in Part I and that those who chose Question 2 are not as good on it as those who chose Question 1. A second possibility is that Question 2 is more difficult than Question 1. If we believe the latter, fairness requires that we somehow adjust for it. How? An immediate response might be to use the performance on Part I to adjust the scores on Part II. This is sensible only if there is a strong relationship between what is tested in Part I and what is tested in Part II. To the extent that they are different the adjustment will be illegitimate. Yet, if they are not different, why bother with Part II at all? The ironic conclusion seems to be that if choice is justified we have no good way to make comparisons among the groups formed by these choices. We can, however, make fair comparisons among choice sections when the use of choice was unnecessary.

Alas, this sort of argument seems to have fallen largely on deaf ears. Choice options are being implemented within more and more large-scale testing programs. It falls now to us to figure out how to allow choice while at the same time adjusting the scores on choice items of potentially very different difficulty to assure the equivalence (and hence the fairness) of the different test forms built by examinees.

5. Some Statistical Issues Facing NAEP

The National Assessment of Educational Progress (NAEP) is an ongoing, congressionally mandated survey designed to measure educational achievement and changes in that achievement over time for U.S. students of specified ages and grades as well as for subpopulations defined by demographic characteristics and by specific background and experiences. Since its inception in 1969, students have been assessed in the subject areas of reading, mathematics, science, writing, social studies, civics, U.S. history, geography, citizenship, literature, music, career development, art, and computer competence. Many subject areas are re-assessed periodically to measure trends over time. The assessment has always included nationally representative samples of students, drawn via complex multistage probability sample designs. These samples permit the measurement of nationally and regionally defined subpopulations of students but do not allow the reliable reporting of state level results. For the 1990 and the 1992 assessments, congress authorized voluntary state level assessments, in addition to the national assessments. For this purpose, a distinct probability sample is drawn within each participating state to provide individual state representative data. An overview of the NAEP design can be found in Johnson (1992) and Rust and Johnson (1992).

Statistical issues currently facing NAEP fall into two general categories. The first category consists of issues related to the effects of nonresponse on estimates of subpopulation achievement. The second category consists of issues related to the use of assessment methodology other than multiple-choice questions.

5.1. *Effects of nonresponse*

The NAEP design selects schools and then students within selected schools for participation in NAEP. At each of these stages of selection, participation is voluntary. Unfortunately, as testing within schools has become more prevalent, the difficulty in obtaining voluntary participation of the selected schools has increased. In addition to expending considerable effort in attempting to convert refusing schools, NAEP handles school refusals by providing substitutes for nonparticipating schools that could not be converted. However, even though the characteristics of the substitute schools are matched as closely as possible to those of the initially selected schools in terms of minority enrollment, urbanicity, and median household income, substitution does not eliminate bias due to the nonparticipation of the initially selected schools.

In addition to school nonresponse, there is also an issue of student nonresponse. NAEP has handled this type of nonresponse by inflating the sampling weights of the responding students to maintain totals within nonresponse adjustment classes within each primary sampling unit. Nonresponse bias thus exists to the extent that the distributional characteristics of the nonrespondents and the respondents differ within each nonresponse adjustment class. Evidence exists (Rust and Johnson 1992; Rogers, Folsom, Kalsbeek and Clemmer 1977) that the vast majority of nonrespondents to NAEP assessments is the same as the respondents in terms of performance and other characteristics. Nevertheless, there is reason to believe that some proportion of the nonrespondents are less adequately handled by the nonresponse adjustment procedures.

The adequacy of the nonresponse adjustment procedures is an issue, particularly for the state-level assessments where a major goal is the comparison of performance between states. Such a comparison is obviously affected by the level and type of nonresponse, and the stability of nonresponse across states. NAEP is currently considering model-based procedures that attempt to quantify the potential dependence of results on the magnitude and characteristics of nonresponse.

5.2. *Effects of assessment methodology*

NAEP has always been at the forefront of assessment methodologies. For the 1992 assessment, for example, more than 40% of the questions within the subject areas of reading and mathematics are free-response items (including items requiring extended responses, such as essays) and all of the items in the 1992 writing assessment require the writing of an essay. Further non-multiple choice assessment techniques include oral interviews, examinee choice questions, evaluation of school-based writing, and assessments of a student's ability to carry out concrete tasks (see Mullis 1992). Each of these relatively nonstandard assessment techniques present statistical and psychometric issues that need to be solved.

For example, many of the so called authentic tests involve a specific task that the student is to perform coupled with a series of questions. The task might be to read a long passage or to conduct a science experiment. The questions range from multiple choice, to short answer, to extended responses requiring one or more paragraphs. The non-multiple choice questions are scored by trained judges. Since the questions are all related to the same task, a commonly made, and key, assumption that the items are locally independent is likely to be violated. (Local independence means that, conditional on a student's ability level, the response probabilities of any pair of items are independent.)

Since each task could be (perhaps) approached in a variety of ways, and since the mechanism used to solve the problem is of interest in authentic testing, statistical mechanisms are needed to identify subgroups of students who approach a task in similar ways. Because the responses to the tasks are rated by judges, work needs to be done to establish and, it is hoped, account for the effects of variability in the judgment process on the ratings provided.

6. Conclusion

This paper is different than most that find their way onto these pages, in that it is a description of problems rather than the more common structure that includes both a statement of the problem and at least an initial solution. In this way it is in the spirit of Hilbert's (1902) famous paper on "Mathematical Problems." We hope that the response to our statement of these problems is as successful at eliciting solutions from the readers as Hilbert's has been.

7. References

Angoff, W.H. and Dyer, H.S. (1971). The Admissions Testing Program. In W.H. Angoff (Ed.), The College Board Admissions Testing Program. New York: College Entrance Examination Board.

Bock, R.D. and Aitkin, M. (1981). Marginal Maximum Likelihood Estimation of Item Parameters: An Application of an EM-Algorithm. Psychometrika, 46, 443–459.

Box, G.E.P. and Tiao, G.C. (1973). Bayesian Inference in Statistical Analysis. Reading, MA: Addison-Wesley.

Braun, H.I. (1989). Empirical Bayes Methods: A Tool for Exploratory Analysis. In R.D. Bock (Ed.), Multilevel Analysis of Educational Data. San Diego, CA: Academic Press.

Frederiksen, J.R. and Collins, A. (1989). A Systems Approach to Educational Testing. Educational Researcher, 18, 27–32.

Fremer, J., Jackson, R., and McPeek, M. (1968). Review of the Psychometric Characteristics of the Advanced Placement Tests in Chemistry, American History, and French. Internal memorandum. Princeton, NJ: Educational Testing Service.

Glaser, R. (1991). Expertise and Assessment. In M.C. Wittrock and E.L. Baker (Eds.), Testing and Cognition (17–30). Englewood Cliffs, NJ: Prentice Hall.

Gulliksen, H. (1950/1987). Theory of Mental Tests. New York: Wiley. Reprint, Hillsdale, NJ: Erlbaum.

Hilbert, D. (1902). Mathematical Problems. Bulletin of the American Mathematical Society, 8, 437–479.

Holland, P.W. and Rosenbaum, P.R. (1986). Conditional Association and Unidimensionality in Monotone Latent Trait Variable Models. Annals of Statistics, 14, 1523–1543.

Johnson, E.G. (1992). The Design of the National Assessment of Educational Progress. Journal of Educational Measurement, 29, 95–110.

Lesgold, A.M., Lajoie, S., Bunzo, M., and Eggan, G. (1988). SHERLOCK: A Coached Practice Environment for an Electronics Troubleshooting Job. Pittsburgh: Learning Research and Development Center, University of Pittsburgh.

Lewis, C. (1986). Test Theory and Psychometrika: The Past Twenty-five Years. Psychometrika, 51, 11–22.

Lewis, C. (1989). Difficulties with Bayesian Analysis for Random Effects. In R.D. Bock (Ed.), Multilevel Analysis of Educational Data. San Diego, CA: Academic Press.

Lord, F.M. and Novick, M.R. (1968). Statistical Theories of Mental Test Scores. Reading, MA: Addison-Wesley.

Mann, H. (1845). A Description of a Survey of the Grammar and Writing Schools of Boston in 1845. Quoted in O.W. Caldwell and S.A. Courtis (1923). Then and Now in Education, Yonkers-on-Hudson, New York: World Book Company, 37–40.

Mislevy, R.J. (1991). Randomization-Based Inference about Latent Variables from Complex Samples. Psychometrika, 56, 177–196.

Mullis, I.V.S. (1992). Developing the NAEP Content-Area Frameworks and Innovative Assessment Methods in the 1992 Assessments of Mathematics, Reading, and Writing. Journal of Educational Measurement, 29, 111–131.

Muthĕn, B. and Satorra, A. (1989). Multilevel Aspects of Varying Parameters in Structural Models. In R.D. Bock (Ed.), Multilevel Analysis of Educational Data. San Diego: Academic Press.

Novick, M.R., Jackson, P.H., and Thayer, D.T. (1971). Bayesian Inference and the Classical Test Theory Model: Reliability and True Scores. Psychometrika, 36, 261–288.

Rogers, W.T., Folsom, R.E., Jr., Kalsbeek, W.D., and Clemmer, A.F. (1977). Assessment of Nonresponse Bias in Sample Surveys: An Example from National Assessment. Journal of Educational Measurement, 14, 297–311.

Rust, K.F. and Johnson, E.G. (1992). Sampling and Weighting in the National Assessment. Journal of Educational Statistics, 17, 111–129.

Scheuneman, J., Gerritz, K., and Embretson, S. (1991). Effects of Prose Complexity on Achievement Test Item Difficulty. Research Report RR-91-43. Princeton: Educational Testing Service.

Spearman, C. (1904). "General Intelligence" Objectively Determined and Measured. American Journal of Psychology, 15, 201–292.

Thompson, P.W. (1982). Were Lions to Speak, We Wouldn't Understand. Journal of Mathematical Behavior, 3, 147–165.

Thurstone, L.L. (1947). Multiple-Factor Analysis. Chicago: University of Chicago Press.

Wainer, H. and Kiely, G.L. (1987). Item Clusters and Computerized Adaptive Testing: A Case for Testlets. Journal of Educational Measurement, 24, 185–201.

Wainer, H. and Lewis, C. (1990). Toward a Psychometrics for Testlets. Journal of Educational Measurement, 27, 1–14.

Received June 1992
Revised November 1992