Recontextualizing Mental Measurement

Harvey Goldstein
Institute of Education, University of London

Why should we be concerned about "physics envy" as a feature of our approach to educational measurement? What are some limitations of current measurement models? Why should there be greater public participation in decisions regarding testing?

The notion that there can be a *theory* of measurement in fields such as education or psychology is both seductive and elusive: seductive because it conveys the promise of high scientific status, and elusive because it has so far defied achievement.

I should make it clear, of course, that like everyone else in this area I am in favor of having decent theories. Most of us I suspect must admit to at least a touch of what Gould (1981) calls "Physics Envy"—the striving to acquire as much as possible of the natural scientist's tool kit. The problem is that what has been labeled as theoretical is so only in a narrow technical sense. That is to say, it is concerned strictly with only the statistical properties of classes of models for test scores and item responses. These models are used extensively to study the relationships among test scores and item responses, to assist in the construction of measurement instruments and in the summarization of individuals' responses.

Clearly, it is possible to discuss the statistical theory of test responses, as Lord and Novick (1968) do in their classic text, in the same way as one can discuss the statistical theory of linear models. The former theory, like the latter, has no inherent correspondence with any *substantive* theory in the area of application to which the statistical model is applied. Nevertheless, the statistical assumptions of the models used may well constrain the range of substantive theoretical possibilities, whether, for example, by forcing relationships to be linear or latent traits to be unidimensional. A substantive theory must describe and also predict events; it must be amenable to empirical testing and falsifiability and it should not be trivial—that is, it must deal with a useful range of circumstances. A statistical model that describes relationships between events and characteristics may be a necessary concomitant of a theory, but it is hardly sufficient. Thus, for example, the standard item response model (IRM) assumes a single between-individual-subject "dimension," and this is normally treated as an axiom and hence untested. Yet it is such a strong assumption that if any substantive theory is to be based on such a model and given the complexities of social reality, it is difficult to see how such a theory can be more than trivial.

The present article examines how psychometric test models based on certain assumptions have come to be used counterproductively by many practitioners and in ways that severely limit the kinds of conclusions that can be drawn.

One does not have to look very far to find claims that a particular statistical model or class of models provides a form of "objective measurement" or constitutes a "theory" of testing; and those claims are not restricted to the purely technical properties of the models but also carry a substantive message. I shall concentrate on test-score models because these have had the most attention over the recent past. Nevertheless, what I have to say applies at the same level of generality, and with appropriate modifications, to other, often earlier, kinds of models based upon scaling or continuous latent variable techniques.

A historical perspective will help to set the scene.

The Galton Inheritance

The dubious credit for the first serious introduction of the notion of mental traits surely must go to Galton (1884). It was he who supposed, by analogy with anthropometrical measurements, that there were dimensions of the mind; that there was something called intelligence that was supposed to be normally distributed among the population. Much of 20th century psychometrics has been devoted to elaborating this proposition by developing statistical techniques for ever more complex modeling of supposed mental structures. The use of IRMs is merely the most recent example of such activity.

One cultural inheritance of this is a seldom-voiced assumption that there really are mental dimensions waiting for us to attach to them

Harvey Goldstein is a Professor of Statistical Methods in the Department of Mathematics, Statistics, and Computing at the Institute of Education at the University of London, 20 Bedford Way, London, England, WC1H OAL. His specializations are educational assessment and multilevel statistical modeling.

numbers or ordered categories. The process of measuring these dimensions, while presenting problems of definition and interpretation, is assumed to be possible in principle: We simply have to go on trying to refine our instruments. It is possible, however, that there are no real "things" that a measurement can capture in the same manner that height or weight can be captured. Because almost all mental measurement, and certainly the kind with which IRMs are concerned, requires some active participation by the subject, that measurement almost inevitably will alter the current state of the subject. In other words, just as physics acknowledges an uncertainty principle, so psychometrics ought to acknowledge a principle of what might be called "interactive measurement."

Thus, for example, a student taking a reading test is not simply exhibiting knowledge or understanding or test-taking "ability" but is actually participating in a learning procedure whereby at the end of the test her state has been altered. This might be the more pronounced the more "authentic" the measurement becomes. As far as I can tell, current measurement models fail to recognize this possibility. Nor is it trivial. One can envisage many situations where, say, a successful response to an early question in an assessment session creates an increase in confidence and a release of creativity that would boost a subject's performance above that which an early failure would produce. An immediate consequence of this is to lead us to question the standard assumption of "local" or conditional independence and direct us towards a more dynamic modeling whereby the response *sequence* is modeled. After all, subjects actually do perceive the time sequence in responding to a test, and we would therefore expect that a more faithful modeling would incorporate this. Thus, instead of regarding each item response as reflecting factors that remain constant for the duration of the measurement, the *relationship* between responses ought to be studied.

I am not here claiming that Galton was wrong; merely that his views seem too long to have been taken for granted so that assumptions about underlying traits are rarely questioned and have become deeply embedded in almost all current models.

The Elements of Item Response Models

A reasonably general model, a so-called *two-parameter* binary response model with more than one subject-ability dimension, models a function of the response probability π_{ij} as follows:

$$g(\pi_{ij}) = \sum_{k=1}^{p} \alpha_{kj}\theta_{ki} + \beta_j + \sum_{m=1}^{q} \gamma_m z_{mi} \quad (1)$$

where θ_{ki} is the value for subject i on the kth ability dimension; β_j, α_{kj} are facility and discrimination parameters respectively for the jth item, and the z_{mi} are observed covariates (for example, representing group membership), with coefficients γ_m. If

$$p = 1, \quad \alpha_{kj} = 1, \quad g(x) = logit(x), \quad \text{and} \quad \gamma_m = 0$$

then we have the notorious Rasch model. If instead $g(x) = x$, then we have the linear *one-parameter* model. This latter model implicitly underlies traditional item analysis procedures, such as those for estimating reliability and calculating discriminations. From this point of view, the more recent logit item response models are simply a sophisticated development of the same approach, a point that seems to be poorly understood. More detail on all this can be found in Goldstein and Wood (1989). Models that introduce extra parameters (for example, to deal with polytomous responses) do not raise new issues affecting the principles with which this article is concerned.

Equation (1) expresses the relationship between the probability of a correct response to item j from subject i. To be able to provide estimates of the parameters of (1) based on observed data, we have to make some statistical assumptions. The key ones are as follows.

Dimensionality

Consider the simple two-dimensional case of model (1):

$$g(\pi_{ij}) = \alpha_{1j}\theta_{1i} + \alpha_{2j}\theta_{2i} + \beta_j. \quad (2)$$

To obtain estimates for the parameters of (2), the usual procedure is to regard the θ_{ki} as realizations of random variables Θ_k and employ, for example, binary factor analysis techniques to obtain efficient estimates (Bock, Gibbons, & Muraki, 1988). An alternative is to regard the θ_{ki} as fixed parameters and then impose sufficient constraints for estimability; an example of this approach is that given by Goldstein (1980). This latter approach is rather difficult to motivate in many practical situations, and the former approach suffers from the traditional problems associated with factor analysis models, notably the arbitrariness of any particular solution and problems of determining the number of dimensions.

The factor analytic approach has increased in popularity but has a difficulty that is seldom elaborated. This is the issue of the "reference population." Clearly, it is possible to have a model that "fits" well in one subpopulation but not in another; for example, where the subpopulations comprise different ethnic groups. This problem is well recognized in the ordinary linear model analysis of observational data, and there is a large theoretical and applied literature dealing with it.

Although the best psychometric practice does address such complexities and challenges assumptions such as that of unidimensionality, this is not *standard* psychometric practice. The implications of the unidimensionality assumption in particular are considerable, and these are now explored further.

The Self-Fitting Model

In the test construction literature, traditional approaches to fitting unidimensional fixed item parameter models to data have adopted what might be called the *self-fitting* model paradigm. In this, the procedures used for estimating parameters and choosing items to delete or revise are designed to yield a close match between the data and the model assumptions, and thus produce impressive "goodness-of-fit" statistics.

Traditional textbooks on "item analysis" exhort test constructors to

exclude or modify "poorly discriminating" items; in effect, those that do not conform to a common unidimensional model. Similar advice is given for logistic item response models—the most extreme expression coming from many exponents of the Rasch model, where "nonfitting" items are given short shrift. Yet the application of unidimensional models often embodies a strong element of tautology, as can be seen by considering the following example.

Suppose we have the two-dimensional IRM

$$logit(\pi_{ij}) = \alpha_{1j}\,\theta_{1i} + \alpha_{2j}\theta_{2i} + \beta_j \quad (3)$$

and suppose we assume the commonly used two-parameter unidimensional model

$$logit(\pi_{ij}) = \alpha_j\theta_i + \beta'_j. \quad (4)$$

The estimates of α_j, θ_i are complex weighted functions of the responses and hence of the coefficients in (3). By varying the latter, for example by choosing items with particular values of these coefficients, necessarily we will change the values of α_j, θ_i.

It follows that where a two-dimensional (or more generally multidimensional) structure exists, the choice of items to be included in a test will determine the parameters of a unidimensional model fitted to that structure. Because these parameters are complex functions of the parameters of the underlying multidimensional model, in general they will have no separate interpretation of their own.

If we make the reasonable assumption that life is most commonly multidimensional, then where attempts are made to obtain a unidimensional structure by removing "misfits" etc., any resulting unidimensional model estimates will inevitably reflect the original choice of items. In other words, the test constructor's choice of items to represent what is to be measured is crucial. Inasmuch as it pays scant regard to the constructor's original intentions, the use of unidimensional item response modeling to determine the final test content seems a somewhat dubious procedure. This underlines the importance of an interaction between test construction and test analysis.

Other procedures, such as almost all forms of test equating, which assume unidimensionality, are suspect for the same reasons. If there are really several dimensions, and if populations differ along these dimensions, then serious distortions will arise and the techniques will produce results with invalid interpretations.

A similar issue arises with techniques for dealing with item "bias," often referred to as *differential item functioning* (DIF) and sometimes traveling under other labels such as *appropriateness measurement*. Briefly, these lead to a test construction procedure whereby items that exhibit extreme or idiosyncratic patterns in terms of group differences typically are marked for exclusion, or at least modification. Thus, if most items on a test discriminated well between men and women, those that did not would be viewed with suspicion. Shepard, Camilli, & Averill (1981) sum up this view: "An item is biased if two individuals *with equal ability* but from different groups do not have the same probability of success on the item" (italics added). The circularity of such a definition is clear because there is, of course, no way to determine whether "ability" is equal other than from performance on the test itself.

Other ideas, based upon a composite unidimensional score, do not really deal with this difficulty because though they admit multidimensionality, they still depend on a unidimensional summary, possibly based upon a subset of items.

DIF models assume a particular dependency of item responses on traits and grouping factors. Thus, tests for DIF are essentially tests of such a model, and alternative hypotheses could include either modifications to the dimensionality structure or to the grouping structure, possibly including further covariates. Thus, DIF studies are perhaps best viewed as exploratory techniques for model fit. Their problem is that they are nonspecific because they are concerned with *any* departure rather than a specific one.

Response Independence

The properties of the standard estimates from models such as (3) depend upon an assumption that the residual random variables are independently distributed. That is, conditional on the item and the individual parameter values the responses to a set of test items are independent.

McDonald (1979) generalizes this assumption to define a "principle of conditional structure" that allows a more general dependency among the residuals; for example, an autoregressive one. Goldstein (1980) makes a similar point, and Jannarone (1986) introduces a class of models that allow both unidimensionality and local dependency. The standard definition of trait dimensionality assumes local or residual independence, but in terms of what we might term *interindividual dimensionality* we can certainly have complex residual structures.

The situation in reality is that of describing or summarizing a set of item responses in terms of a small number of parameters. We can interpret particular parameters, for example, as interindividual ability dimensions or as covariances. The problem in real life is that the data may not discriminate between, say, a one-dimensional model with nonindependence and a two-dimensional model with independence. This seems to be an inherent problem with latent variable models. In fact, we could have the situation where a one-dimensional covariance model may fit better than, say, a three-dimensional independence model and with fewer parameters. There is a distinct failure in the literature to take seriously nonindependence models.

It is worth saying that if the independence assumption is violated, then inference based on IRMs becomes shaky. To mention one example, the traditional estimates of reliability assume independence, and in particular the formulae for lower bounds are incorrect if independence does not hold.

Other Models

Some authors have begun to question the dominance of current item response models. For example, Masters and Mislevy (1991) advocate latent class models for the allocation of students to "stages" that need not be uniquely ordered. This seems to be an approach worth exploring,

especially in diagnostic assessment. Yet, like factor analysis, it has an attendant set of problems, such as determining the number of classes and interpreting the results. It is not clear in their discussion, however, why any particular form of statistical model should be more appropriate for one measurement rather than another. By the same token, there is nothing inherently more "theoretical" about these models than the ones I have already discussed, although the attempt by these authors to start from a substantive problem and then search for an appropriate model is in welcome contrast to a common tendency to apply simple IRMs to everything in sight. In the context of a renewed interest in other forms of educational assessment, especially in the United States with the exploration of alternatives to multiple choice and related styles of testing, a willingness by the measurement community to think afresh about its models would be very welcome.

Context

Let me now turn to some of the research having to do with contextual effects on achievement test scores.

A large body of research, especially in mathematics and science, has shown how the contextual embedding of a test question can markedly alter the response success rate. The British Assessment of Performance Unit research has shown how the layout of math questions changed the response when the actual mathematics content remained the same (Foxman, Ruddock, & McCallum, 1990). Murphy (1989) has shown how the practical context of performance on authentic test questions tends to favor girls rather than boys when real-life tasks are emphasized as opposed to algorithmic problems. Other research (Wolf, Kelson, & Silver, 1990) has demonstrated how elusive the notion of "skills" can be when one tries to measure them in practical contexts—a finding that casts considerable doubt upon current fashions for "criterion-referenced" assessment.

In the U.S., the so-called National Assessment of Educational Progress (NAEP) reading anomaly (Beaton & Zwick, 1990) demonstrated how sensitive test questions could be to the company they keep. The same small set of questions gave different results depending on how they were embedded in the test. This has led to a search for ways of measuring such effects, and that seems to be a fruitful approach. This is, of course, not the same as trying to find context-independent test questions. Such an activity may have its uses, but is not likely to be helpful in most practical situations. Beside the enormous problems of trying to understand the effects of context, the ever more elaborate development of IRMs seems rather irrelevant.

Social, Ideological, and Other Contexts

I have already mentioned the general problem of context influencing performance. Let me now elaborate on some wider implications.

The case of the Golden Rule Insurance company versus Educational Testing Service (ETS) has been discussed at some length in recent years (Anrig, 1988; Goldstein, 1989). Briefly, the insurance company managed to persuade ETS to adopt a policy of item selection for its entry tests that minimized Black-White differences. It worked by ETS choosing a pool of items, all of which satisfied standard criteria for test inclusion. From this pool the final selection was made by choosing those items that produced the smallest (on average) differences between Blacks and Whites. After some years of this, ETS decided that the whole thing had been a mistake and that they wished to call off the deal. This created something of a stir, one useful consequence being that the issue received some exposure, at least among the measurement profession.

There are certainly some technical difficulties in the procedure, and there is room for debate over the details of its implementation. It is clear, however, that the typical reaction of many psychometricians and measurement experts was that, in the absence of clearly recognized qualitative manifestations of item bias, it should be predominantly technical criteria, and not sociopolitical considerations, that determine test content. One way or another, there existed a technical "fix" for the problem of residual ethnic differences, be it DIF technology or some more elaborate IRM or both. For example, Linn and Drasgow (1987) claim that the Golden Rule settlement was "based on pragmatic rather than sound psychometric principles." In a similar vein, Jaeger (1987), then president of NCME, cites the Shepard et al. (1981) definition of bias to argue against Golden-Rule-like procedures. Such views are natural enough from professionals perhaps, but are not necessarily ones that society at large would wish to accept. Why should we not impose upon our test constructors statistical procedures based on sociopolitical constraints?

Let me give another example from the United Kingdom (Goldstein, 1986). Since the mid-1970s it has been illegal in a number of areas to discriminate between students on the grounds of gender. This was interpreted in the mid-1980s by the Equal Opportunities Commission (EOC) to mean that where tests are used for educational selection they should use the same cut points for each gender. Prior to this, some Education Authorities had been using separate norms for boys and girls to select equal numbers for the academically elite Grammar schools. Because girls tended to score higher than boys on the tests then in use, the EOC interpretation was expected to have the effect of passing more girls to the grammar schools.

Unfortunately, we do not know what the outcome was because the situation was not monitored closely. Nevertheless, there is at least a suspicion that those Education Authorities that were unhappy with the implications simply went away and searched for tests in which the advantage to the girls was reduced or absent. If such tests were not forthcoming, then it would not have been too difficult to devise one to produce such an outcome. Interestingly, when this point was made to the EOC, there was simply no response, the inference being that this would raise too many difficult questions that they would rather not know about.

The notion that tests should seek to minimize or otherwise manipulate group differences is a perfectly

Continued on page 43

participants in their own learning and reflect on their own competencies and achievement. With clarity and thoughtfulness, the individual chapters describe in more detail ways in which to accomplish these goals.

Lastly, in fairness to the authors, in this review I did not provide a summary of the contents for all of the chapters. At times, I tended to highlight issues or assertions that were of most interest or most controversial to me. In so doing, I hope I have not misrepresented the opinions expressed by the authors. As I mentioned previously, I recommend this book to all educators, psychometricians, and policymakers. It provides a thoughtful, reflective discussion on current conceptualizations of assessment and challenges the reader to consider, if not agree with, all the ideas expressed by the authors.

References

Linn, R. L. (1993). Educational assessment: Expanded expectations and challenges. *Educational Evaluation and Policy Analysis, 15*(1), 1–16.

Mehrens, W. A. (1992). Using performance assessment for accountability purposes. *Educational Measurement: Issues and Practice, 11*(1), 3–9, 20.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–104). New York: American Council on Education.

Mislevy, R. J. (1993, April). Test theory reconceived. An invited address presented at the Annual Meeting of the National Council on Measurement in Education, Atlanta.

National Commission on Excellence in Education. (1983). *A nation at risk: The imperative for educational reform* (Publication No. 065-000-00177-2). Washington, DC: U.S. Government Printing Office.

Mental Measurement

Continued from page 19

legitimate *ideological* aim. It can be attacked ideologically, therefore, but it can be attacked technically only if the technical grounds have a theoretical basis. This makes the search for good substantively grounded theory important. Without substantive theoretical support the notion that there is a technical view that can decide this issue (and related ones) is itself an ideological assumption that can and should be challenged. Of course, the search for a theoretical grounding cannot be separated from ideological considerations, but it is better that these are explicit rather than implicit consequences of particular choices of models.

Of course, I am fully aware that by opening up to public participation what has hitherto been a professional process, there may be all kinds of problems and instabilities and uncertainties that could make life uncomfortable. That, however, seems to me to be rather a good thing, although I doubt many in the measurement profession could be persuaded to rally to such a cause. It will be interesting, nevertheless, to see the results of the increasing demands from outside the profession for more openness, accountability, and explanation of some of our more arcane procedures in terms that outsiders are able to understand. I perceive the demystification of item response "theory" as a step in the right direction.

Notes

This article is based upon a paper read at the conference on Modern Theories of Measurement, Montebello, Canada, November 1991.

This article has benefited from the comments of referees and the editor, to whom I am most grateful.

References

Anrig, G. R. (1988). ETS replies to Golden Rule on "Golden Rule." *Educational Measurement: Issues and Practice, 7*, 20–21.

Beaton, A. E., & Zwick, R. (1990). *Disentangling the NAEP 1985–86 reading anomaly.* Princeton, NJ: Educational Testing Service.

Bock, R. D., Gibbons, R., & Muraki, E. (1988). Full-information item factor analysis. *Applied Psychological Measurement, 12*, 261–280.

Foxman, D., Ruddock, G., & McCallum, I. (1990). *APU mathematics monitoring 1984–88 (Phase 2).* London: Schools Examination and Assessment Council.

Galton, F. (1884). *Hereditary genius.* New York: Appleton.

Goldstein, H. (1980). Dimensionality, bias, independence, and measurement scale problems in latent trait test score models. *British Journal of Mathematical and Statistical Psychology, 33*, 234–246.

Goldstein, H. (1986, May). Gender bias and test norms in educational selection. *Research Intelligence: BERA Newsletter*, p. 2–4.

Goldstein, H. (1989). *Equity in testing after Golden Rule.* Unpublished manuscript, University of London, Institute of Education.

Goldstein, H., & Wood, R. (1989). Five decades of item response modelling. *British Journal of Mathematical and Statistical Psychology, 42*, 139–167.

Gould, S. J. (1981). *The mismeasure of man.* New York: W. W. Norton.

Jaeger, R. J. (1987). NCME opposition to proposed Golden Rule legislation. *Educational Measurement: Issues and Practice, 6*, 21–22.

Jannarone, R. J. (1986). Conjunctive item response theory kernels. *Psychometrika, 51*, 357–373.

Linn, R. L., & Drasgow, F. (1987). Implications of the Golden Rule settlement for test construction. *Educational Measurement: Issues and Practice, 6*, 13–17.

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

Masters, G. N., & Mislevy, R. J. (1991). *New views of student learning: Implications for educational assessment* (Tech. Rep. RR-91-24-ONR). Princeton, NJ: Educational Testing Service.

McDonald, R. P. (1979). The structural analysis of multivariate data: A sketch of a general theory. *Multivariate Behavioural Research, 14*, 21–38.

Murphy, P. (1989). Assessment and gender. *National Union of Teachers Education Review, 3*, 37–41.

Shepard, L., Camilli, G., & Averill, M. (1981). Comparison of procedures for detecting test item bias with both internal and external ability criteria. *Journal of Educational Statistics, 6*, 317–375.

Wolf, A., Kelson, M., & Silver, R. (1990). *Learning in context: Patterns of skills transfer and training implications.* London: The Training Agency.