RESPONSES

Assessment for the States—Possibilities and Limitations: A Critical Look at the Duplex Design

Harvey Goldstein, University of London Institute of Education
Leslie McLean, Ontario Institute for Studies in Education

Darrell Bock and Robert Mislevy's comprehensive plan for large-scale assessment of school achievement clearly reflects current demands for testing programs that satisfy a number of separate purposes. The authors identify a hierarchy of levels (students, classes, schools, districts, . . .) for which periodic assessment reports are desired, and for reports that can be compared over time.

In order to meet general constraints of limited resources and testing time, they propose a well-known and often-used scheme *matrix sampling* in which each student responds to a different, relatively short, booklet of achievement items. Using new analysis techniques, however, this procedure is said to be able to provide achievement "scores" not only for districts, schools and classes (which it could all along), but also for students. An additional claim is that it can tell us with precision whether achievement levels are higher or lower from one testing period to another.

The design and analysis can achieve these aims, according to the authors, because a sophisticated *item response model* (IRM) is used to analyze the student responses. (The common use of the phrase item response *theory* to describe such models seems to us a misnomer.) Bock and Mislevy's claim is examined in more detail below, but first we emphasize some aspects of their design and its analysis with IRM, which together they call the *duplex design*.

Matrix Sampling Designs

Any testing procedure which places a light burden on students and teachers is welcome. As Bock and Mislevy point out, all properly designed matrix sampling plans require minimal testing time yet permit us to estimate separate topic, or *domain*, scores for districts, schools and classes. Such plans are in fact special cases of more general multilevel designs which permit us also to take account of background factors such as social group (see e.g., Goldstein, 1987). The duplex design is one particular example of this class of designs, an example for which special advantages are claimed, based upon the use of an IRM.

More general matrix designs do not require IRM's and their strong simplifying assumptions. General designs will be less statistically efficient, however, because they will need to incorporate more terms, more *parameters*, in their models. It is always true that the more strong statistical assumptions one makes, the higher the statistical efficiency one can achieve. If we are not prepared to adopt some of the assumptions, then our results are more general but we will typically need larger samples to improve their statistical efficiency. This will not usually be a problem in a large-scale assessment, and more testing time per student is not required. In our view a debate about the usefulness, efficiency and validity of large scale testing should present a wide framework of alternative approaches so that competing models and assumptions can be evaluated. The "duplex design" paper is a welcome contribution to that debate.

Curriculum Elements and Composite Scales

Bock and Mislevy make another important point when they stress the importance of reporting in narrowly defined curriculum domains. In their mathematics example, reports are available at the class and school level for all 57 varieties of items, making curriculum sensitive assessment possible. Estimation of classroom means can be carried out efficiently by straightforward aggregation procedures and does not require the apparatus of an IRM.

When it comes to calculating student scores, however, difficulties arise. Each student is represented by only one item of each type, that is, one item from each "element pool" or domain (perhaps none, if the student fails to respond to the item). Thus, scores need to be *estimated*, either for content categories (topics) or proficiencies or even mathematics as a whole. This is where the IRM comes in. Item response *scaling* is done within these groupings of items and the scales used to derive student "scores."

Consider, for example, the topic scale *numbers* based on five items (Table 2). The assumption underlying use of an IRM is that there are fewer than five underlying 'dimensions' or

'factors' on which each item will 'load' with different coefficients. Typically, as in Figure 1, only one such dimension is assumed, and on the basis of this assumption scores are derived for the scales. The major problem, however, is that this collapsing of the separate items for reporting at the student level is inconsistent with the per-item or per-curriculum element reporting at class or school level. If it is adequate to report *number* for students, why is it not adequate to do so for classes or schools? If it is important to report on detailed curriculum elements for classes and schools then surely it is also important to do so for students?

Presumably, the reason for reporting in such detail at the classroom level is that we really believe that *number* is at least five-dimensional in terms of curriculum response and the like. If this is so, then *any* summary, at any level, in fewer than five dimensions, (and especially a one-dimensional one) will lose information and the resulting average will to a large extent reflect the implicit weighting of the separate curriculum elements found in the test forms. Thus, for example, a one-dimensional summary using a weighting procedure derived from an IRM may well give roughly equal weightings to each curriculum element simply because these elements are equally represented in the test forms. The decision to include them in such a way is a deliberate choice on the part of the test constructor, who may or may not be a curriculum specialist.

This is not to say, of course, that summary scores for topics on proficiency areas are not useful. The point is that these summaries are based ultimately upon subjective (and hopefully well reasoned) combinations of finer curriculum elements. As we see it, the real importance of these designs lies in the detailed profile reporting at the class and higher levels and in the relationship of these profile responses to curriculum and other factors, rather than in the estimation of individual student scores (McLean, Wolfe, & Wahlstrom, 1987).

Trends Over Time

One of the persistent demands of policy makers and others is for test scores that allow us to compare achievement levels over time. Equally persistent problems make such comparisons difficult. Bock and Mislevy allude to this when they discuss the so called *item parameter drift*. While the issue is not a straightforward one, its essence can be stated as follows.

Over time, curriculum, subject matter and other changes may occur which make some test items inappropriate. For the sake of argument, suppose we are dealing with a truly one-dimensional scale (but the following reasoning applies quite generally). In such a scale an item which becomes inappropriate changes its difficulty. It thus has to be *rescaled* or else replaced with a new item from the same dimension. The duplex design calls for this rescaling to be done with respect to the *base* set of items. Unfortunately, the interpretation of the resulting scale scores becomes problematic.

For simplicity, consider just two items, A and B. Between time 1 and 2, suppose item B becomes "more difficult" and is rescaled (or a new "equivalent" item introduced) so that it maintains a consistent relationship with item A. The problem of interpretation is that we have no way of choosing between two alternative explanations. We have no way of knowing whether the population of students at occasion 2 has the same *ability* as at occasion 1 and it is really item B which has become more difficult, or whether item B has remained unchanged and the population at occasion 2 is less able and at the same time item A has become correspondingly easier. In the latter case, it is item A we should be rescaling so that we can then reflect the changing characteristics of the population.

In other words, there is an inescapable duality between items and students and there can be no absolute *external* reference scale. Thus all statements about item changes are with respect to specific populations and vice versa. This is a fundamental and inescapable property which affects all item replacement or rescaling procedures. It implies that there can be no completely objective method for describing absolute changes over time—only perhaps relative ones. Changes in the *differences between subgroups* might be measurable, although such measures too have their difficulties (see Goldstein, 1983 for a discussion).

Alternatively we might be prepared simply to keep the same unchanging set of items for our reports, describing item difficulties as observed, for example, but this merely shelves the problem of changing item relevance. Of course, we may well decide to use judgmental methods based on known curriculum changes and the like to form reasoned views about changes over time. If so these judgments will become a proper matter for educational opinion and debate, with room for rational disagreement.

School Effectiveness

Bock and Mislevy address themselves briefly to issues of the "effectiveness" of schools. This is an important area which is a topic of much current research interest (see, e.g., Aitkin & Longford, 1986) as well as attention in the popular press. It too is fraught with difficulties and admits no simple solution. To compare schools fairly requires more than an adjustment for economic and social characteristics. Curriculum content is important, but perhaps most important of all is a measure of achievement levels of incoming students, what in England is called school intake achievement. These levels are likely to vary widely within a state or a district, and this variation can seriously distort estimates of school effectiveness—denying recognition to some deserving schools and hiding the shortcomings of others.

Conclusions

Some of the strengths and limitations of the duplex design have been described, especially the limitations of the means proposed for estimating scores for individual students. Reasons were given why the use of IRM's should not encourage legislators and other policymakers to believe that absolute changes over time can be measured objectively

An important positive aspect of the duplex design is the provision of class and school profiles of narrow curricula domains. This represents a real advance over aggregate or global reporting. The stress in this paper on clear reporting methods is also very welcome. Finally, Bock and Mislevy's paper has provided the opportunity to debate issues of considerable importance in the efforts to improve teaching and learning.