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 will allow 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 we emphasize some aspects of their design and its analysis with IRM, which together call the duplex design.

Matrix Sampling Designs

Any testing procedure which places a tight burden on students and teachers is welcome. As Bock and Mislevy point out, if 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 specifications of more general multilevel designs which permit us 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 IRMs 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 routinely 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" or domain (perhaps none, if the student fails to respond to the item). Thus, scores need to be estimated, either for content categories (topics) of proficiency 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 two items (Table 3). The assumption underlying use of an IRM is that there are fewer than five underlying "dimensions" of
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 has become more able. In the latter case, it is item A which has become correspondingly easier. In the latter case, it is item A which we should be questioning so that we can then reflect the changing characteristics of the population

in other words, there is an inseparable duality between items and students and there can be no absolute internal reference scale. Thus, all comparisons about changes are only meaningful in relation to some reference system and some change. This is a fundamental and unavoidable property which afflicts all item replacement or rescaling procedures. It is thus that there can be no completely objective method for distinguishing absolute changes over time—only perhaps relative ones. Changes in the differences between subgroups may be misleading, although such measures too have their difficulties (see Goldstein, 1989 for a discussion).

Alternatively we might be prepared simply to keep the same unchanged 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 may be at risk 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., Ashen & Longbird, 1976) as well as in 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 effects in 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.