DEBATE CORNER

more thoughts on testing

BY HARVEY GOLDSTEIN

THE ARTICLE by Les McLean[1] in a recent issue raises some important general points about large scale testing program, including the International Assessment of Educational Progress (IAEP). I would like to elaborate on three of these points: the secrecy adopted by testing agencies about the items used, the publication of mean scores so that countries or boards or even schools can be ranked, and the failure to exploit properly the information obtained from students.

Item and test disclosure

There is now a wealth of research that demonstrates how even apparently innocuous changes to the wording of a test item, its layout or its position in relation to other items, can markedly change the response pattern. Indeed, Educational Testing Service (ETS) itself has demonstrated this latter point very effectively in its analysis of the so called National Assessment of Educational Progress (NAEP) reading anomaly.[2] They found that changing the position of items in successive tests appreciably altered response rates.

It clearly follows that in order to understand an interpret the responses to individual items, and test scores derived from those responses, we have to know what the items are. We need to know their wording, format, relationship to other items and so forth. It simply is not sufficient to be told by the test analyst that an item can be interpreted as measuring, say, numerical skills, since there are any number of such items, each of which could produce different results. These might be in terms of gender differences, or country differences or differences between curriculum regimes. In other words, without publishing the measuring instrument in its entireity any presentation of results inevitably is incomplete and potentially misleading. Indeed, the secrecy adopted by many professional testers towards their instruments contrasts markedly with the canons of behaviour among survey researchers where it is the common practice to publish the questions asked of respondents to a questionnaire. Why should testers behave differently?

A common response from testing agencies is that they have to keep at least some items secure so that they can be used again, as the basis for relating responses at one time to responses at a later time. This is what was done by NAEP from 1984 to 1986 and led to the so called reading anomaly which demonstrates just how difficult that operation can be. Even if this point is conceded, the other items are still left available for inspection. Yet in the recent IAEP report[3] there is no such presentation of items, nor is there in the technical report, which followed a few months later. Neither is it good enough for an organization such as ETS to say that it will release the items to researchers so long as they do not disclose the contents, since that effectively prevents those researchers conducting an open discussion about the items.

In fact I can see no compelling case for *any* items to be kept secret. This especially is so where items are used to provide an equating of two tests over time since. As with the NAEP reading anomaly, it is precisely those items present at both occasions that are crucial to any interpretation of trends. The most that could be said is that these items be kept secret until they have been used on the second occasion, but then both the tests should be published in detail. Like survey researchers, the testing organizations should get together and draw up a code of conduct for the publication of results.

Ranking the results

The standard method for reporting results of large scale testing programs, especially those that concern country comparisons, is to publish item or domain mean scores, per cent correct or some non-linear transformation of these as in so called item response models. Only secondarily, and usually in a subsequent report, are any more detailed analyses done that attempt to explain the differences found. The IAEP is a good example of a first report limiting itself to reporting average country differences by domain or subtopic in mathematics and science.

Yet surely this is wrong. What is of real importance is some attempt to go behind the crude differences to see if any light can be thrown on the factors associated with those differences. Thus IAEP collected information on opportunity to learn (OTL) — a measure of how much exposure to the test items was had by each student tested. Yet although average values of OTL were reported, no attempt was made to relate this to performance. As Richard Wolfe[4] showed, the performance differences did seem in part to be explained by OTL. This is hardly surprising, but a detailed study of the relationship between OTL and performance would tell us far more of interest than the comparisons we are given. Indeed, there is an important sense in

1 Les McLean, "Let's Call a Halt to Pointless Testing," **Education Canada**, Fall 1990, pp. 10-13

2 A. Beaton, ***The NAEP Reading Anomaly*** (Princeton, N.J.: ETS, 1990)

3 A. Lapointe, N.A. Mead, and G.W. Phillips, ***A World of Differences*** (Princeton, N.J.: ETS, 1989)

4 R. Wolfe, "An Indifference to Differences: Problems with the IAEP 88 Study," **Educational Researcher**, 1990

which the comparisons presented are misleading, as I shall now explain.

A single overall score

If we look at mathematics, it is broken down into several domains such as number skills and problem-solving. As McLean points out, 39% of the maths items belong to the number skills domain and only 13% to the problem-solving domain. Thus, in calculating a total score, number skills receive a high weighting and problem-solving a low one. Yet a total score could be derived in any number of different ways by applying differential weights to the domains. Thus, if problem-solving were given the same weight as number skills, it would improve the overall ranking for a country such as Britain, which did badly in the latter but actually did best in the former. There is no objective reason for choosing a procedure which weights the domains approximately in proportion to the number of items they happen to contain, as IAEP did. The failure to point this out in the report and to refer to a single overall mathematics "proficiency," as if it had some objective reality, even could be deemed irresponsible.

Nor is it an adequate defence for IAEP to say that the detailed domain results are also presented. As is always the case, if a single, simple, summary is presented, it is this that gains public prominence. There seems to have been little overt attempt by IAEP to disown the use of such a single summary.

A better approach

I have already suggested that test organizations should consider carefully how they can make their instruments accessible so that informed comment becomes possible. A recent report on test secrecy discusses this issue in some depth.[5] I have also implied that the publication of simple summaries may do more harm than good, and by implication that more time and trouble should be taken to carry out informative analyses before going public. I have singled out IAEP, not because it is necessarily the worst offender, but because it is one of the most visible. Specifically, let me make a few recommendations.

First, because test results are by their nature difficult to interpret, that interpretation can only gain by being subject to close peer scrutiny and criticism. This activity, in order to be really effective, has to do done *before* results are publicly launched. This can be achieved via seminars and meetings, to which potential critics as well as friends are invited. That way a better product is likely and the temptation to oversimplify should be reduced. Furthermore, with careful planning this should not unduly delay the public appearance of results.

Second, those who fund research and those who would wish to apply it should become more aware of all these issues. They should stop hoping for quick, simple answers and learn to live with what will always be a complex reality. If this means that they become less willing to fund certain kinds of research, then so be it. What researchers should stop doing is trying to promise the earth and then finding that they fall short when they try to deliver it.

Third, let us have an informed debate about these issues. For too long the testing profession has been taciturn about its procedures. Often it has hidden behind obscure mathematical technicalities to justify what are often oversimplified models of educational achievement. On some issues, for example, the choice of a test so as to narrow the difference in performance between majority and minority groups, it has often responded to criticism by retreating again behind irrelevant mathematicization.[6] In short it has attempted to shun the real substantive issues, in many cases where there is considerable social concern. If it fails to put its own house in order, and in particular if the major testing agencies fail to do so, then perhaps it deserves to have regulations imposed upon it. □

5 *J. L. Schwartz and K. A. Viator, **The Prices of Secrecy** (Cambridge, Mass. Educational Technology Center, Harvard Graduate School of Education, 1990)*

6 *See, for example, H. Goldstein, "Equity in Testing After the Golden Rule," paper given at meeting of the American Educational Researchers' Association, San Francisco, April 1989, and R. L. Linn and F. Drasgow, "Implications of the Golden Rule Settlement for Test Construction," **Educational Measurement: Issues and Practice**, Vol. 6 (1987), pp. 13-17*

Harvey Goldstein is a professor in the Department of Mathematics, Statistics and Computing at the Institute of Education in London, England, and an adjunct professor at the Ontario Institute for Studies in Education in Toronto.

PURPOSE AND PRACTICE

Continued from page 31

which is desirable and accomplish change and improvement where necessary.

4. In addition to the commitment to evaluate, boards should demonstrate a corresponding commitment to fair treatment.

5. Through co-operation and collaboration, boards and superintendents should clarify and articulate the role expectations of the superintendent and these expectations should form the criteria that guide the evaluation.

6. The board and superintendent should annually review and revise the role expectations of the superintendent.

7. School boards should prepare themselves to evaluate their superintendent and so board members should seek training to develop and improve their evaluative skills, as well as clarify their directions for education and how they expect the superintendent to further these.

8. Judgement and rating scales should not be used in evaluation, as they are not useful in considerating the realities of human behaviour and intended outcomes.

9. The practice of evaluation should be examined in light of research literature but this should not become the sole basis for evaluation.

10. The practice of evaluation should be examined in light of exemplary models like those in Turtle Mountain and Souris Valley School Divisions.

11. Policies and practices should be division-based, therefore, reflecting the uniqueness of each local jurisdiction.

12. Greater emphasis should be placed on the professional development of the superintendent to enable him or her to be well informed and to offer effective administrative leadership in educational planning.

13. The board should facilitate a professional development plan that addresses the isolation of the superintendent's office.

14. The board should model behaviour that establishes a positive climate, invites growth and development and communicates both expectations of and belief in the superintendent. □

Judith D. Silver, a past president of the Manitoba Association of School Trustees, is a master's degree student in educational administration at the University of Manitoba and a board member of Seven Oaks School Division No. 10. This article is based on a paper prepared under the guidance of Dr. J. Anthony Riffel, Department of Educational Administration and Foundations, University of Manitoba, Winnipeg