February 1999

Education Policy Analysis Archives 07/05

Arizona State University

University of South Florida

9

Education Policy Analysis Archives

Volume 7 Number 5 **February 17, 1999** **ISSN 1068-2341**

A peer-reviewed scholarly electronic journal
Editor: Gene V Glass, College of Education
Arizona State University

Articles appearing in **EPAA** are abstracted in the *Current Index to Journals in Education* by the ERIC Clearinghouse on Assessment and Evaluation and are permanently archived in *Resources in Education*.

Some Comments on the Ad Hoc Committee's Critique of the Massachusetts Teacher Tests

Howard Wainer
Educational Testing Service

"It is a trite but true observation, that examples work more forcibly on the mind than precepts: And if this be just in what is odious and blameable, it is more strongly so in what is amiable and praise-worthy."

Henry Fielding, *Joseph Andrews*, 1742

The critique of the Massachusetts Teacher Tests (MTT) by Haney and his colleagues is deserving of comment, both because of the impact of the MTT and because of the evocative manner in which the tale is told. Their emphasis on examples makes for a forceful argument, and I fear that my reliance on precepts may look meager by comparison. Nevertheless, I hope that some of the observations that follow contribute to the more reasoned assessment of these instruments and their use not just in Massachusetts but in the many other states where similar programs are being developed or contemplated.

It seems clear from some of the data that the Ad Hoc Committee have presented that the MTT is not up to snuff, although it is a bit of a puzzle why its reliability isn't higher. Even a little bit of pre-testing (and Spearman- Brown) would have shown what test length is required for standard reliability. Perhaps time and economic pressures intervened and a less than fully developed instrument was rushed to the field? As the old

saying goes, if it's worth doing, it's worth doing badly.

The reliability in this whole process that is of greatest interest is the reliability associated with the pass/fail decision (i.e., the reliability that emanates from the standard error of measurement around a score of 70). This can be easily calculated from the raw data by noting the inverse of the information function after fitting an Item Response Theory (IRT) model. If the raw data were available, I could do it myself. Considering the high stakes associated with these tests, there is more than the usual obligation on someone's part to make these raw data available.

It can be noted in Figure 1 that 32 of the 66 teachers who failed initially passed on retesting. This does not necessarily speak to unreliability of the tests. It is possible that grit, determination and plenty of quick preparation accounted for the elevation of the scores on the second administration. One wishes to see the other half of the four-fold table (the counts of those who passed the first time and who passed or failed the second time), but such data are never available--passing once is a ticket out of the testing system. One could contrive an approximation (in the spirit of split-halves) by generating two scores (e.g., odd and even item scores) and constructing the 2-by-2 table (pass/fail vs. score-1/score- 2) using 35 as the passing score. Of course, one would want to do this for all candidates, not just those who failed the first time. The extent to which this table is diagonal (i.e., approaching zeroes in the off-diagonals) is the reliability of the test for the decision. Of course it is an underestimate (too short a test) and you need to apply something like the Spearman-Brown "prophecy formula" to estimate the reliability of the decision based on the full test.

A comment made about "Peter" in the examinee vignettes seemed a bit strong. Peter scored in the 91st percentile on the GRE-V and "between the 80th and 85th percentile in Reading". These two results were described as "quite at odds". In fact, they seem fairly consistent, especially considering the somewhat lower reliability of the MTT. We see this by using Kelley's (1947) equation for estimating true score τ from observed score, *X*. Kelley's equation yields

$$\hat{\tau} = \rho_{xx} X + (1 - \rho_{xx}) \bar{X}$$

$$\hat{\tau} = 0.7\,(91) + 0.3\,(50) = 64 + 15 = 79$$

So we would predict (from Peter's observed score of 91) that his true score is 79. Thus on retesting getting 80-85 is not out of line. In fact if reliability was not 0.7 but rather 0.73 the true score estimate is 80. If reliability was 0.85 we would expect a true score of 85. (See also Wainer, 1999.)

One might even say that the Ad Hoc Committee exaggerated the discrepancy between Peter's SAT and MTT scores for effect. (Indeed, a more temperate, less polemical, tone throughout the article would have been more persuasive for me and most readers, perhaps.) This same bit of exaggeration shows itself again in Recommendation 1: "No exam at all is better than an unreliable exam...." Whether or not no exam is better than any exam depends on (i) how unreliable that exam is, (ii) what the selection ratio is, and (iii) the nature of the cost function associated with errors of each type.

For example, suppose one has a test that is 92% accurate (4% that should "pass" actually fail and 4% of those that should "fail" actually pass), and suppose further that one has a selection ratio of 95 to 100--that is one expects to pass 95% of all applicants, and false passes are as bad as false failures. Under these circumstances, the test is a bad idea since if one simply passed everyone the error rate would be 62.5% of what would be obtained if the test were used. But suppose the cost function is different. Suppose a person who fails improperly can take the test again and pass whereas a person who

passes improperly is installed forever to do irreparable damage to our children. Now, using the test with its obvious imperfections seems like a better idea.

What cost structure would one wish to impose on a licensing exam for heart surgeons? Airline pilots?

And what about validity? The ancient Chinese tests had an enormous selection ratio--one in a "gazillion." Consequently, the test did not need much validity to be worthwhile, though it did need at least some. Of course enormous numbers of worthy candidates failed, but the odds of an unworthy one passing were small enough to be ignored for all practical purposes. The same structure manifests itself in such tests as the National Merit Scholarship test in which 1,500 "winners" are chosen from more than 1,500,000 applicants. This 1 in a 1,000 selection ratio means that a very large number of worthy kids do not win, but that all winners are truly wondrous. These three key issues that must be decided before making such a statement as "no exam at all is better than an unreliable exam" where "unreliable" means "reliability in the 0.7 range".

I conclude, then, on the basis of my reading and my own biases, that the MTT needs work, but certainly provides some information to guide decisions. And, if the selection ratio is high enough, the MTT might even be good enough. (I tend to put a high cost on allowing an incompetent teacher into the classroom and a relatively low cost on asking a marginally passing teacher to take the test again.) But that's just me.

I hope that the Massachusetts Department of Education will now make enough data available for a more proper test analysis and, perhaps, publishing the Ad Hoc Committee's critique will hasten that day.

References

Kelley, T. L. (1947). *Fundamentals of statistics*. Cambridge: Harvard University Press.

Wainer, H. (1999). Is the Akebono School failing its best students? An Hawaiian adventure in regression. *Educational Measurement: Issues and Practice, 18*, 26-33.

About the Author

Howard Wainer

Email: hwainer@ets.org

Howard Wainer received his Ph.D. from Princeton University in 1968, after which he was on the faculty of the University of Chicago. He worked at the Bureau of Social Science Research in Washington during the Carter Administration, and is now Principal Research Scientist at the Educational Testing Service. He was awarded the Educational Testing Service's Senior Scientist Award in 1990 and was selected for the Lady Davis Prize. He is a Fellow of the American Statistical Association. His latest book, *Visual Revelations*, was published by Copernicus Books (a division of Springer-Verlag) in 1997.

---

http://epaa.asu.edu

General questions about appropriateness of topics or particular articles may be addressed to the Editor, Gene V Glass, glass@asu.edu or reach him at College of Education, Arizona State University, Tempe, AZ 85287-0211. (602-965-9644). The Book Review Editor is Walter E. Shepherd: shepherd@asu.edu . The Commentary Editor is Casey D. Cobb: casey.cobb@unh.edu .

EPAA Editorial Board

David D. Williams
Brigham Young University

EPAA Spanish Language Editorial Board