BOOK REVIEW

High stakes: Testing for tracking, promotion, and graduation. Edited by J. P. Heubert & R. M. Hauser. 1999. Washington, DC: National Academy Press.

From the 1960s to the 1990s, the field of personnel psychology (a.k.a. industrial-organizational psychology) produced an impressive body of technically advanced research on the costs and benefits of testing for personnel selection. Thousands (yes, thousands) of empirical studies were conducted in the United States alone, demonstrating that a fairly general aptitude or achievement test is the best single predictor of performance for the overwhelming majority of jobs, better than all other factors that employers generally used in hiring. The estimated net benefits of using tests for personnel screening were huge, with costs minuscule and benefits enormous.

In the late 1980s, the U.S. Department of Labor considered providing the federal government's General Aptitude Test Battery (GATB), which is used for hiring in federal jobs, to local employment offices for use in hiring outside the federal government. The test would have been made available to job applicants who wished to take it, and test results would have been made available to employers who wished to review them.

The Labor Department asked the Board on Testing and Assessment at the National Research Council to review the question. Their report is extraordinary. In the face of overwhelming evidence to the contrary, the Board declared the following: there was only negligible evidence to support the predictive power of the GATB, and tests in general provided no benefits in personnel selection. Their conclusions were reached through tortuous illogic and contradiction, and possibly a judicious selection of both committee members and research sources (see Phelps, 1999).

For example, not one of the hundreds of academic psychologists who studied personnel selection was invited to participate in writing the report, whereas several education professors who were well-known opponents of high-stakes testing were. Moreover, only one of the thousands of empirical studies on personnel selection was discussed.

The Board dismissed the benefits of hiring better qualified applicants for jobs by arguing that if an applicant were rejected for one job, the applicant would simply find another somewhere else in the labor market, as all are employed somewhere. (No matter that the other job might be less well-paid, in an undesirable field or location, part time, temporary, or even nonexistent.

Educational and Psychological Measurement, Vol. 60 No. 6, December 2000 992-999

In the view of the report, "unemployment is a job.") The Board continued with the astounding contradiction that, whereas selection (and allocation) effects should be considered nonexistent because all jobs can be considered equivalent, general tests, like the GATB, cannot be any good as predictors because these tests do not account for the unique character of every job.

The New Book

Now, we have this new report (Heubert & Hauser, 1999) from the National Research Council's (NRC) Board on Testing and Assessment on a very similar topic: high-stakes student testing. Is this report any more objective than its ancestor? Not by much. Many of the same people were involved in producing it and it betrays some of the same biases, both procedural and ideological.

Does the NRC Board support the use of high-stakes tests? Well, yes, seemingly provided that the tests can be administered without any student failing them or dropping out of school or feeling bad because of them, and provided that no ethnic or gender groups or any disabled or second language students do any better or worse on average than any others.

The present report includes over 40 recommendations. With some exceptions, any one of them taken alone seems reasonable. Taken together, they would impose a burden on the states that none could feasibly meet, particularly if the regulators at every tiny step in the testing process that the NRC Board wants to see regulated shared the Board's premises. In my view, under these recommendations, test approval would be held up for years, probably until such a time as test content becomes outdated. If the onerous, burdensome regulatory regime the NRC Board wants to impose on the states indeed were to be imposed, we would never see another high-stakes test.

The report even floats a proposal to require that tests be pretested, before they can be used for high-stakes purposes, using a new, very general standard of predictive validity. Because testing proponents argue that high-stakes tests promote more learning or better employment, the NRC Board argues that we should hold off certifying the use of any particular high-stakes test until it can be proven that over time, the test does increase learning (say, in college) and improve employment outcomes. It would take years, of course, to conduct such an experiment, even if the experiment were feasible. But, of course, it is not. One cannot test the effects of high-stakes tests when the stakes are not high, as they would not be during the life of the experiment.

The report comes down solidly on the side of more radical egalitarians and radical constructivists in the high-stakes testing debate. This is evident in many ways, but none is more telling than the source material that is cited (and not cited). Sources were included that buttress the Board's views and hundreds of sources that do not were ignored.

With huge resources at its disposal (a budget of over $1 million), the NRC Board minimized its research effort. On issue after issue, it threw its lot in

with a single or a single group of researchers. The chapter on tracking is really about the work of just one person. The strong counterevidence and counterarguments on that issue are kept completely hidden from the reader. Again, the early childhood, readiness testing, and promotion and retention sections also feature only one person's point of view. Chapter 10 has only two sources, the NRC itself and the chair of the NRC Board. Chapter 11 cites substantially only two sources. Two thirds of the citations refer to less than a dozen research sources.

Although virtually no room was found in the report for the considerable evidence and many reasonable arguments in favor of high-stakes testing, ample space was found to remind us of the racial overtones of IQ testing in the 1920s and 1930s and to suggest that such sentiments among testing proponents are still widespread; after all, witness publication of *The Bell Curve*.

What Is Left Out

The most glaring aspect of the report, then, is what is not there. What is left out? Pretty much everything. Everything, that is, that could have made the report more useful and balanced. For example,

- The NRC report repeatedly implores us to make a proper assessment of the costs and benefits of tests, but never brings any numbers to the issue, even though the numbers are widely available. The costs are minuscule and the benefits huge.
- Although most benefits of high-stakes testing are not even mentioned, the costs of not having high-stakes testing (e.g., the implicit encouragement of lower standards, less learning, and social promotion) are ignored as well.
- The research on the huge reliability, validity, and fairness problems associated with the alternatives to high-stakes tests, like high school grade point averages, is completely left out.
- The report was written as if the United States is the only country on Earth. One would think in reading the report that we are the first country in the world to consider administering high-stakes tests. Surprise, the majority of other advanced industrialized nations in the world have been administering large-scale national-level or state-level high-stakes tests for decades. The NRC Board cautions time and again throughout the report that we should not develop high-stakes tests until we know what will happen if we do this or that. Instead, doing more research is endorsed. It is never mentioned that with a few overseas telephone calls one can learn exactly what happens when one does this or that.

  This lack of familiarity with other countries leads to some embarrassing assumptions. For example, it is stated based on Third International Mathematics and Science Study (TIMSS) data that, "the difference in average achievement of students in different classes in the same school is far greater in the United States than in most other countries" (p. 93). This is used as an argument *against* tracking. What the report does not seem to acknowledge is that most other advanced countries, including all but one of the countries performing better than the United States on the TIMSS, practice much *more* selectivity and *more*

tracking in their education systems than we do, but they do it by school rather than by classroom.

- Does the NRC Board again ignore the studies of psychologists on the topic of high-stakes selection? Yes. There are 10 citations (out of 400) from psychology journals, but they pertain only to a discussion of assessment standards and theoretical concepts of validity. The report avoids, in its entirety, the huge mass of accumulated *empirical* evidence on high-stakes selection from psychology journals. The report looks only at education journals, even then, only the work of a dozen researchers, all opponents of high-stakes testing.
- Given the NRC Board's interest in the benefits and costs of high-stakes student testing, it is only natural that the work of economists on the matter would be reviewed. Nonetheless, the report included only three citations (out of 400) to economists' studies. The work of economists Robert Costrell and John Bishop on one tangential side point is mentioned. Robert Costrell has demonstrated with elegant mathematical models that local school districts face strong disincentives to enforce high standards (they will be punishing their own students in the college and labor markets if other school districts maintain lax standards), thus standards enforcement must occur at the state level, or higher.

  But, perhaps the works of the Cornell labor economist, John Bishop, could not be completely ignored, and were not. For those unfamiliar with Bishop's research, he has for more than a decade used data sets from large national and international test administrations (the Scholastic Aptitude Test [SAT], the International Assessment of Educational Progress [IAEP], the TIMSS, the National Assessment of Educational Progress [NAEP]) to show that students from states, countries, or provinces with high-stakes testing regimes perform better on these neutral, common tests, controlling for background variables. Conclusion: high-stakes testing induces more learning.

  How is Bishop's work received? Only one of his many studies is cited. Second, Bishop's findings are discounted:

  > He (Bishop) found that countries with demanding exit exams outperformed other countries at a comparable level of development. He concluded, however, that such exams were probably not the most important determinant of achievement levels and that more research is needed. (p. 174)

  So if high-stakes exams are not the *strongest of all* predictors of academic achievement, they do not count at all. No matter that they are a positive and strong predictor of achievement. No matter that their net benefits are huge.
- The opinions of the general public, parents, teachers, and students are seemingly ignored just as effortlessly. The report acknowledges that public support for high-stakes testing is overwhelming. But, what does the public know?

  > Despite some evidence that the public would accept some of the potential tradeoffs, it seems reasonable to assume that most people are unaware of the full range of negative consequences related to . . . high-stakes test use. Moreover, it seems certain that few people are aware of limits on the information that tests provide. No survey questions, for example, have asked how much measurement error is acceptable when tests are used to make high-stakes decisions about individual students. The support for testing expressed in polls might decline if the public understood these things. (pp. 44-45)

> Then, again, the support might not decline. Almost all adults have been through at least 10 years of education. During that time each of them took many tests. They know tests are not perfect. They also know that the alternatives to tests are not perfect either. They are very familiar with tests and they are familiar with the alternatives, and they want more high-stakes testing.
>
> Granted, most members of the public are not psychometric experts. But, that does not disqualify them from setting policy in our democracy. As a group, they are not experts in the details in any other policy arena either, such as criminal justice, environmental regulation, or nuclear arms control.

The logic of the NRC report is carried by means of some convenient assumptions and some contradictions. Readers of the report should not miss either the assumptions or the contradictions.

Convenient Assumptions

All Classrooms Equally Wonderful

The report assumes that what happens in the classroom, absent standardized testing, is just wonderful. Because it is wonderful and because enforced standards and high-stakes standardized tests encourage changes in classroom behavior, testing opponents classify test-induced changes as bad, as deviations from wonderfulness. These deviations are labeled *corruptions* of the natural order and standardized test scores as *pollution* of the natural evaluations of students, which can only be done by teachers.

It is not entertained as possible that what teachers teach in the absence of common standards could be ideal, but also could in some cases be less than ideal. Moreover, what happens in the school as a whole, absent common and enforced standards, is presumed to be just as wonderful. Schools that value sports or social achievement more than academic achievement or that employ social promotion as their chief academic standard, for example, are employing natural, "uncorrupted" structures on their students.

All Students Equal

The report does not entertain the politically incorrect proposition that some students might work harder than others. The students who work harder should not be rewarded or encouraged, as that would make the other students feel bad. The high achievers should, rather, be punished for their bad behavior and made to pay a penance of community service. That is, they should not be allowed to skip a grade or take advanced classes. They should be held back with the remedial students where they will be assigned to work as tutors, trying to help students who do not work as hard to pass. It feels as though, because all students have equal ability and work equally hard, any achievement

differences shown by tests only prove that the tests themselves must be flawed or the curriculum or instruction is bad or biased.

Accelerated Curricula

A major finding of the report is that students in lower track classrooms get easier work to do. Those of us who are not "technical experts" might naively think that this makes perfect sense—students who are not mastering the regular material are given easier material or taught at a slower pace.

The report suggests that in reality, tracking is a conspiracy to hold certain types of students back from success. It is mentioned that lower-track students tend to do less well on high-stakes tests. That comes as no surprise because these students were doing less well with their regular school work at the outset. The report, however, argues that these students do less well on high-stakes tests because they were given easier material in their lower track. The problem is the low tracking, not what put students in the low track in the first place.

The report advocates that lower track students be taught material *more* difficult than average and at a *faster* than average pace, even though they were put in the lower track in the first place because the regular material was too difficult or covered too quickly. Accelerated instruction might be considered one plausible remedy by some, but the plausibility that acceleration is not a cure, or at least not a cure for everyone, should have been given more consideration.

We all know that slower students can catch up if they get extra help (and if they want to do the extra work). One could well argue that social justice demands they be given the opportunity. So, should we not all celebrate a system that guarantees these students this extra help? Where do we find such school systems?

In states with high-stakes testing, that is where. In states *without* high-stakes testing, standards do not matter, meeting standards does not matter, and students are promoted and graduated no matter how little they learn. There is no need for extra help, after-hours tutoring, summer school, or Saturday classes. In high-stakes testing states, students in academic trouble get the extra help they need. Remedial instruction *must* be offered to students demonstrably requiring assistance, and that demonstration is high-stakes testing.

Contradictions

National Tests Versus National Boards

National tests are judged to be bad because education should be a state and local issue in the United States, but the report encourages establishing a na-

tional board to review the "technical quality" of high-stakes state tests. Would this board be composed of the country's most strident testing opponents?

Pro– and Anti–Parents' Control

State tests are judged to be bad because they reduce local control and parents' influence in their children's schools, yet the report encourages the reader to dismiss the public's opinions on high-stakes testing (which are overwhelmingly in favor) because the public is ignorant about the "technical" aspects of testing.

Pro– and Anti–Federal Control

The report opposes President Clinton's Voluntary National Tests (VNTs) ostensibly because the tests would reduce local control (even though they are voluntary). Then the report complains that, with the VNTs, "the federal government would be unable to regulate how states and local districts would use the test results" (p. 41).

Use of Hard Versus Easy Standards

Although the report evaluates high-stakes tests by seemingly perfectionist standards, much of the research it cites to buttress its arguments does not meet minimal standards of scientific quality. Much of the research consists of little more than ideological treatment with a veneer of research-like prose.

Conclusion

If high-stakes standardized testing were to be abolished, as the report authors might like, American society as a whole would be much worse off, and so would many individual students. Probably the most unfairly affected would be the high achievers among the poor. Wealthy families who value academics have the choice of moving to a school district where their high achieving children can excel, or sending their children to a private school. It is a waste of money and otherwise too bad that they feel they must move, but they can. Poor families are not so mobile.

High achieving students who cannot leave a school district where academic achievement is undervalued face varied pressures that impede them: pressure to fit in and be popular; to excel at sports; to work at low-pay, dead-end jobs to earn money for cars and parties; and so on. If they study hard and excel at academics, they will be taunted; disliked; called "nerd," "geek," "dork"; or be accused of "acting White."

The report spends considerable effort worrying about the feelings of students who might fail high-stakes tests, but little if any effort worrying about the social fallout of abandoning high academic standards. High-achieving students among our poor should be considered our country's most precious human resources.

For a variety of reasons, our society very badly needs these students to prosper; so their gifts and ambitions should be nurtured, not discouraged. The report, however, in the effect of its recommendations, would have these students treated as pariahs and have them feel guilty for wanting to work hard and succeed. After all, if these students work hard and succeed, won't that make other students who do not want to work hard look and feel bad?

Abandoning the enforcement of high academic standards will not eliminate pressures and hurt feelings among our youth, however. Pressure and hurt feelings are facts of life. Abandoning academics just means the pressures will come from and the hurt feelings will be caused by nonacademic aspects of these students' lives.

Is that really what we want? In radical egalitarian bliss, there will be no high-stakes tests, no academic standards enforced in any meaningful way, and no academic tracking. Academic progress in every school and for every student will be slowed to the preferred pace of the least-motivated student.

—Richard P. Phelps
Washington, D.C.

Reference

Phelps, R. P. (1999). Education establishment bias? A look at the National Research Council's critique of test utility studies? *The Industrial-Organizational Psychologist*, *36*(4), 37-49.