Appendix D:

The National Research Council's Testing Expertise*

Richard P. Phelps

Imagine this scenario if you can: A new surgical technique has been developed by medical surgeons that is estimated will provide health and longevity benefits to U.S. citizens on the scale of tens of billions of dollars. Over a thousand controlled studies have been conducted to date, and the aggregate results are overwhelmingly positive. Moreover, many of the studies, and their meta-analyses, have been conducted by some the world's most respected medical professors and surgeons.

The U.S. Department of Health and Human Services, which would have to pay for the new surgical procedure if it were to be approved for reimbursement under Medicare, wishes to conduct one final evaluation of the efficacy of the new technique. So, they contract with the scientific "court of last resort", the National Research Council (NRC), to evaluate. The NRC agrees and, as usual, sets about recruiting experts to serve on a committee that will conduct the study and produce a final evaluative report.

But, the NRC does not recruit any of the medical professors who have been studying the new technique and publishing their results in medical journals, nor does it recruit any of the medical surgeons who have tried the new technique on human subjects. It does recruit two medical doctors to serve on the committee, but they have no experience with the new technique or its research literature.

Instead, the NRC recruits five veterinarians to serve on the committee, none of whom has any experience with the new technique or familiarity with its research literature. Moreover, they are holistic veterinarians, who accept the use of surgery only for exploratory purposes and condemn the use of invasive techniques that might decisively alter an animal's, or a patient's, condition.

This scenario is fictional. But, change the four elements of medicine, surgery, veterinarians, and the U.S. Department of Health and Human Services, respectively, to standardized testing, personnel psychology, education professors, and the U.S. Department of Labor, and the story is factual.

In the late 1980s, the Labor Department proposed allowing state employment offices to provide job candidates' results on the General Aptitude Test Battery (GATB) to private employers (Hartigan & Wigdor). They asked the National Research Council to critique the proposal.

Hundreds of predictive validity studies had been performed on data sets incorporating GATB scores, usually correlated with one or more job performance measures, such as supervisor ratings, output-per-time period, promotions, earnings increases, and so on. Some readers may be familiar with utility analysis and the fairly predictable conclusions of hundreds of related studies: scores on general ability tests are better predictors of performance in most jobs than are other single predictors (Schmidt & Hunter).

Grade point averages, for example, only tell an employer how well a potential employee performed in school relative to other students at that particular school, if that. Grade point averages are normed at the school level. Given the enormous variety in school quality and standards, it is no wonder that general ability test scores explain a large amount of additional variance in regressions of job performance on groups of predictor variables that also include school grade point averages.[1]

The NRC Report

The NRC appointed a Committee on the General Aptitude Test Battery to write a report, the tone of which is not particularly respectful of the rich tradition of research by industrial-organizational psychologists. The Committee criticized the validity studies of the GATB in several ways, driving down the predictive validity coefficient through a variety of rationales. They conceded a coefficient of 0.22, half the level of the highest, unadjusted predictive validity claimed for the GATB (Hartigan & Wigdor, 1989, pp. 134–171). Cutting the estimates in half, however, still left a net present value estimate of about $10,000 per worker lifetime (or, a total social net benefit into perpetuity beyond $13 trillion), an enormous benefit by comparison with the meager cost of a standardized test.

Then, in their chapter addressing the economic claims made for the GATB, the Committee claimed flatly that there are no job selection benefits to testing because the U.S. labor market is a zero-sum game: if one employer selects better workers by using GATB scores, the Committee argued, other employers will get the remaining workers and it's all a wash. All workers work somewhere in the economy.

Analyzing the NRC Report

Several aspects of the NRC report, *Fairness in Employment Testing*, are striking, in addition to its bitter tone: (a) the odd composition of the committee; (b) the repeated insistence of the committee that there was only meager evidence for the benefits of testing, in the face of thousands of studies in personnel psychology research demonstrating those benefits; (c) the theory of the zero-sum labor market; and (d) the logical contradiction in the report's primary assertions that: all jobs are unique so general ability tests will be invalid for each, but there is no benefit from selection because any worker's abilities will be equally useful anywhere they work, no matter what their training and no matter what the field of work.

The Odd Composition of the Committee

Several years ago, I telephoned Alexandra Wigdor, the NRC study director and report co-editor, to ask why only two personnel psychologists (and none who had conducted research in the issue under study) were recruited for a study panel on personnel testing, whereas five education researchers (four of whom were well known to be opponents of high-stakes testing) were. She asserted that there was no deliberate effort to exclude personnel psychologists or include education researchers. They had sought out the best researchers they could find.[2]

Out of curiosity, I calculated the probabilities of picking only education school professors at random from a large pool of testing experts. Let's assume that personnel testing experts are equally distributed across places where personnel psychologists or education school faculty work. That's a big IF, but I want to be conservative. There were about 1,000 members in the National Council for Measurement in Education and about 3,900 in the wider-scope Measurement and Research Methodology division of the American Educational Research Association. Similarly, there were about 1,500 members in the American Psychological Association's Evaluation and Measurement division and over 5,000 in the Society for Industrial and Organizational Psychology.

Depending upon whether one circumscribes the fields narrowly or broadly, personnel measurement experts outnumber school testing experts by a ratio of about 5.7 to 4.3. What are the odds that five educational testing experts, by chance, were the five best qualified to serve on the committee? Less than two percent, assuming that school testing experts are just as qualified, on average, to judge personnel testing issues as are personnel testing experts. Add the eminently

reasonable assumption that personnel testing experts are more qualified, and the odds diminish toward zero.

Would the NRC hire microeconomists to evaluate a macroeconomic problem? Would it hire inorganic chemists to study an issue in organic chemistry? Would it hire personnel psychologists to evaluate school curricula? Why did the NRC hire education researchers to evaluate personnel testing issues? ...especially given that the United States boasts some of the world's most advanced research and dozens of the world's most respected researchers in personnel testing?[3]

The Meager Evidence of Benefits Argument

Consider the following quotes from the NRC report, *Fairness in Employment Testing*:

- It is also important to remember that the most important assumptions of the Hunter-Schmidt models rest on a very slim empirical foundation....Hunter and Schmidt's economy-wide models are based on simple assumptions for which the empirical evidence is slight (1989, p. 245).
- Some fragmentary confirming evidence that supports this point of view can be found in Hunter et al. (1988)... We regard the Hunter and Schmidt assumption as plausible but note that there is very little evidence about the nature of the relationship of ability to output (1989, p. 243).
- There is no well-developed body of evidence from which to estimate the aggregate effects of better personnel selection...we have seen no empirical evidence that any of them provide an adequate basis for estimating the aggregate economic effects of implementing the VG-GATB on a nationwide basis (1989, p. 247).
- ...primitive state of knowledge... (1989, p. 248).

Was the NRC Committee correct about the paucity of research? Not remotely. From the 1960s on, thousands of studies have been conducted by dozens of researchers in personnel psychology affirming positive net benefits to the use of general ability testing in employee hiring.

There are so many studies it is easier to count just the meta-analyses. A 1988 meta-analysis by John Boudreau covered 87 such studies. A 1984 meta-analysis by Schmitt, Gooding, Noe, and Kirsch covered over 300. A 1998 article by Schmidt and Hunter presented the validity of 17 different selection procedures over 85 years. Hunter and Hunter conducted a meta-analysis of 23 meta-analyses in 1984, summarizing thousands of validity studies.

The Zero-Sum Labor Market Argument: Part 1

The NRC Committee asserted that, contrary to the claims of personnel psychologists, there are no job selection benefits to testing; the U.S. labor market is a zero-sum game. If one employer becomes more efficient in selecting good workers by using job applicants' GATB scores in making selection decisions, the Committee argued, some other employer will end up with those less efficient workers and it's all a wash. All workers work somewhere in the economy (Hartigan & Wigdor, 1989, pp. 241–242).

The zero-sum labor market argument is erroneous in several respects. For example, the unemployed comprise about 5% of the labor force. The Committee cited the fact that the unemployment rate is fairly stable over time as evidence that the unemployed population is stable (Hartigan & Wigdor, 1989, pp. 235–248). While the rate may vary only within a narrow band, the labor market churns people through the ranks of the unemployed and marginally employed over and over.

Using figures from the Bureau of Labor Statistics for the average duration of unemployment (16.6 weeks) and the average number unemployed in 1995 (7.4 million), I estimate the number

of individual "spells" of unemployment for 1995 at 23.2 million.[4] That totals to 17.5% of the labor force unemployed at some time during the year (U.S.BLS, 1997, tables 2, 31, 35).

Another 3.3% of the labor force in 1995 were "economic part-time" employees. That is, they wanted to work full time but could not find full-time employment.[5] Add them to the 17.5% above for a proportion of the labor force close to 21% (U.S.BLS 1968–96).

Then, there are "contingent workers," whose number is difficult to estimate. Anne E. Polivka (1996) calculated estimates ranging from 2.7 million workers in jobs less than a year, who expected the jobs to last no longer than 1 year more, to 6 million workers who simply did not expect their jobs to last. If we subtract the subpopulation of persons classified as "independent contractors or self-employed" from her upper bound, for the reason that those people had chosen a necessarily contingent occupation, we calculate 5.3 million workers who believed their jobs were temporary and probably did not want them to be. That total comprises 4% of the labor force.

These three subpopulations above—unemployed at some time during the year, economic part-time, and contingent workers—were 25% of the labor force.

That still does not include the large number of workers employed outside their field of training, like philosophy Ph.D.s who work as computer programmers, college graduates in international affairs who work as secretaries, and so on. These workers had jobs that required lower level, or different, credentials for entry. These workers were "underemployed."

Finally, an estimated 8.6% of the adult population who quit looking for work out of discouragement for their prospects remained out of the labor force entirely.

The NRC Committee assumed that if a worker didn't get selected for a job, she would get selected for a different job and that other job would be equivalent in the most important ways to the job she didn't get. That assumption is untenable. The person not selected for the first job could end up unemployed (p=.175), unwillingly working part time (p=.033), working in contingent employment (p=.04), underemployed (p=?), working in a field outside their training, or out of the labor force entirely. This is a large group of adults.

The Zero-Sum Labor Market Argument: Part 2

Let's pretend that two college students graduate at the same time from different colleges with degrees in organizational psychology and enter the job market as Worker A and Worker B (see Figure 1). They have approximately the same grade point averages, but Worker A attended a college with higher standards, followed courses of more rigor, studied more, and studied harder than Worker B. Thus, while both workers A and B accumulated human capital in the field of organizational psychology and in general abilities, Worker A accumulated more than did Worker B, a human capital surplus. This surplus is not detectable from the college transcripts, however, or letters of recommendations, or work experience, which are the same for both A and B. The surplus is detectable only through testing.

In strong economic conditions, both employers have jobs available; in poor economic conditions, only one employer has a job available.

If only one employer tests, that employer will become aware of Worker A's human capital surplus and will want to hire Worker A, but will only have to offer a slightly higher salary than the other employer offers Worker B. This is because the other employer is ignorant of Worker A's surplus and so sees workers A and B as equally qualified. The employer knowledgeable of Worker A's surplus will, thus, capture Worker A's surplus in the form of higher quality work, without having to pay more than a nominal amount for it. If both employers test, and both are aware of Worker A's surplus, then Worker A can bid them against each other up to the point

where the anticipated benefit of her surplus is fully incorporated in her salary offer. With full information, Worker A is compensated for her surplus. If an employer's job is in the graduates' field of study, they should be more willing to pay for Worker A's surplus because she has more need of it.

Of the 12 possible outcomes of these various permutations, three produce benefits that can be ascribed to job selection or allocation effects.

In the outcome that produces job selection benefits, Employer X is the only one with a job available in a poor economy; she tests the two job applicants; and Worker A is hired after scoring higher on the test. Because Worker A must take whatever salary is offered, the employer gets to pocket Worker A's human capital surplus. Without the test, however, employer would have hired Worker A with only a .5 probability and, thus, only a .5 probability of capturing the surplus. The test increases Employer X's probability of putting Worker A's surplus to use from .5 to 1.0.

In another outcome that produces job allocation benefits, Employers X and Y both have jobs available, but Employer X's job is in the same field so she needs Worker A's surplus more than Employer Y does. In this case, both employers test, and Employer X hires Worker A at a salary somewhere above Worker B's, and shares Worker A's surplus with Worker A. If Employer X did not test, her probability of capturing part of Worker A's surplus would be only .5, while the probability that some of Worker A's surplus would be wasted (if Worker A worked at the job outside her field) would also be .5. Thus, by testing, employer A increases the probability of putting the human capital surplus to use from .5 to 1.0.

In another outcome producing job allocation benefits, only Employer X tests and becomes aware of Worker A's surplus. She hires Worker A for only a slightly higher salary than was offered Worker B. By testing, Employer X increases the probability of hiring Worker A (and putting her surplus to use) from .5 to 1.0.

Each of the three outcomes increases the probability, through testing, of putting Worker A's human capital surplus to use rather than letting it waste. Productive assets are employed, rather than left unused.

Logical Contradiction of Homogenous Jobs and Unique Tests

The NRC Committee claimed no job selection benefits to employment testing:

> Employment Service use of the VG-GATB will not improve the quality of the labor force as a whole. If employers using the Employment Service get better workers, employers not using the Employment Service will necessarily have a less competent labor force. One firm's gain is another firm's loss... The economy as a whole is very much like a single employer who must accept all workers. All workers must be employed (Hartigan & Wigdor, 1989, pp. 241–242).

Essentially, the NRC argued that skills measured by employment tests are equally useful in all jobs. That, of course, assumes that general intellectual aptitudes or abilities are equally valuable in all lines of work, and co-vary equally with all other relevant skills, say those used in brain surgery or street sweeping.

At the same time, in its chapter 8, "GATB Validities," the NRC Committee asserted that "Validities vary between jobs... GATB validities have a wide range of values over different jobs." In order for pre-employment tests to be beneficial, they must be uniquely tailored to unique jobs (Hartigan & Wigdor, 1989, pp. 170–171).

The two assertions are contradictory. The NRC Committee tried to have it both ways—declaring the GATB to be invalid in predicting job performance because every job is unique and,

at the same time, declaring selection effects moot because any worker not getting one job will get another and equal value will be provided to society overall.

Conclusion and Discussion

I spoke with three persons intimately familiar with the activity of the National Research Council's Committee on the General Aptitude Test Battery. One claimed that the Committee was deliberately set up to be a hostile committee. Another claimed that the Committee considered only one personnel testing study from among hundreds in existence, yet made claims that implied they had considered all of them. The third claimed that the Committee refused to consider some of the most basic and relevant evidence pertaining to personnel testing issues, such as: the ways in which the Hunter and Schmidt estimates of utility *underestimated* the benefits of testing; the true magnitude of the effect of range restriction on the utility estimates (for which the Committee refused to correct); the true value of average inter-rater reliability of ratings of .50 (they assumed .80, thus under-correcting for criterion unreliability); and (pertaining to the NRC assertion that Hunter and Schmidt did not adjust their estimates for the time value of money, incremental validity, or what have you) the substantial research in personnel psychology that has explicitly considered all those issues (and found little difference in the direction or magnitude of the resulting utility estimates).

These are serious charges and imply that those at the National Research Council responsible for the evaluation of testing issues were biased and, further, that the NRC Board on Testing and Assessment had been "captured" by education interests.

Endnotes

* This Appendix excerpts from 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), April, 37–49.

References

Bishop, J. H. (1994). Schooling, learning and worker productivity. In R. Asplund (Ed.). *Human capital creation in an economic perspective*. Helsinki: Physica-Verlag.

Boudreau, J. W. (1988). Utility analysis for decisions in human resource management (Working Paper #88–21), Ithaca, NY: Cornell University, School of Industrial and Labor Relations.

Farkus, S, Johnson, J, & Duffet, A. (1997). *Different drummers: How teachers of teachers view public education*. New York: Public Agenda.

Hartigan, J. A., & Wigdor, A. K. (1989). *Fairness in employment testing: Validity generalization, minority issues, and the General Aptitude Test Battery*. Washington, DC: National Academy Press.

Hunter, J. E., & Hunter, R. F. (1984). Validity and utility of alternative predictors of job performance. *Psychological Bulletin, 96 (1).*

Polivka, A. E. (1996). A profile of contingent workers. *Monthly Labor Review*, Washington, DC: U.S. Department of Labor.

Schmidt, F.L., & Hunter, J.E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implication of 85 years of research findings. *Psychological Bulletin, 124,* 262–274.

Schmitt, N., Gooding, R. Z., Noe, R. D., Kirsch, M. (1984). Meta-analysis of validity studies published between 1964 and 1982 and the investigation of study characteristics. *Personnel Psychology, 37,* 407–422.

U.S. Department of Labor, Bureau of Labor Statistics (1997). *Household data: Annual averages*. Washington, DC: Author, Tables 2, 31, 35, and unpublished tabulations.

---

[1] John Bishop (1994) has written much about how good high school students only get paid what they're worth after several years of having to prove themselves all over again in the workplace because there exits no good means of signaling their competence to employers at the outset.

[2] Only two members of the committee had any background in personnel psychology: one worked as an executive in a large corporation; the other worked in an administrative position at a university. Neither of them, however, was intimately familiar with the research on test utility, the studies of the GATB, and employee hiring. Several very well known personnel test utility researchers were included in the "Liaison Group," but that group was seldom consulted and kept separate from the secret deliberations of the committee.

[3] Do education professors, in general, have policy preferences similar to the general public's, qualifying them to make policy decisions for the rest of us? Not on testing issues (see Farkus, Johnson, & Duffet).

[4] At first thought, one might think that I am calculating the number of persons who are unemployed at some time during the year. While the estimate probably brings us close to that number, the estimate probably also subsumes a small number of spells that are shared by individuals. In other words, some persons may have more than one spell of unemployment in a year.

[5] There is no average duration figure with which to calculate the number of persons who go through "economic part time" spells during the year. We have to settle for this lower-bound number for the number of workers who are at some time during the year forced to accept part-time employment when they would prefer full-time employment.