The Rot Festers:
Another National Research Council Report on Testing

Hout, M., & Elliott, S.W. (2011). *Incentives and test-based accountability in education.* Committee on Incentives and Test-Based Accountability in Public Education. Board on Testing and Assessment, Division of Behavioral and Social Sciences and Education, National Research Council. Washington, DC: The National Academies Press.

reviewed by Richard P. Phelps[1]

In research organizations that have been "captured" by vested interests, the scholars who receive the most attention, praise, and reward are not those who conduct the most accurate or highest quality research, but those who produce results that best advance the interests of the group. Those who produce results that do not advance the interests of the group may be shunned and ostracized, even if their work is well-done and accurate.

The prevailing view among the vested interests in education does not oppose all standardized testing; it opposes testing with consequences based on the results that is also "externally administered"—i.e., testing that can be used to make judgments of educators but is out of educators' direct control. The external entity may be a higher level of government, such as the state in the case of state graduation exams, or a non-governmental entity, such as the College Board or ACT in the case of college entrance exams.

One can easily spot the moment vested interests "captured" the National Research Council's Board on Testing and Assessment (BOTA). BOTA was headed in the 1980s by a scholar with little background or expertise in testing (Wise, 1998). Perhaps not knowing who to trust at first, she put her full faith, and that of the NRC, behind the anti-high-stakes testing point of view that had come to dominate graduate schools of education. Proof of that conversion came when the NRC accepted a challenge from the U.S. Department of Labor to evaluate the predictive validity of the General Aptitude Test Battery (GATB) for use in unemployment centers throughout the country.

[1] Richard P. Phelps is the author of *Standardized Testing Primer* (Peter Lang) and co-author and editor of *Correcting Fallacies about Educational and Psychological Testing* (American Psychological Association).

***Fairness in Employment Testing*, 1989**[2]

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 GATB, which was 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 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. Many of the world's most-respected personnel and GATB testing experts were appointed to a "Liaison Committee", but it was never consulted; their names, however, were then published in the final report, as if to imply they approved of the report.

They did not. Members of the Liaison Committee accused the NRC of choosing deliberately a committee they knew would be hostile toward the GATB research.

Moreover, only one of the thousands of empirical studies on personnel selection was discussed. In the face of thousands of predictive validity studies on general aptitude tests in employment, the study committee wrote: "very slim empirical foundation", "the empirical

---

[2] 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. This section borrows from Phelps, R. P. (2008/2009c). The National Research Council's Testing Expertise, Appendix D in R. P. Phelps (Ed.), *Correcting fallacies about educational and psychological testing*, Washington, DC: American Psychological Association. < http://supp.apa.org/books/Correcting-Fallacies/appendix-d.pdf > and Phelps, R. P. (1999, April). Education establishment bias? A look at the National Research Council's critique of test utility studies. *The Industrial-Organizational Psychologist, 36*(4), 37–49.

evidence is slight", "fragmentary confirming evidence", "very little evidence", "no well-developed body of evidence", and "primitive state of knowledge."

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).

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.

Constants on NRC testing study committees for the past quarter century have been the multiple participation of members of the federally-funded Center for Research on Educational Standards and Student Testing (CRESST), headquartered at UCLA, and members of an even more radical (anti-) testing research center at Boston College.[3] Committee memberships are then rounded out with scholars known in advance to support CRESST biases and a few others with recognizable names and ideological sympathies, but little familiarity with the study topic. The many scholars who disagree with CRESST's point of view are neither invited to participate nor cited in the study reports.

***High Stakes*, 1999[4]**

The most revealing aspect of the National Research Council's 1999 report, *High stakes: Testing for tracking, promotion, and graduation* (Heubert & Hauser) is its choice of source material. Sources were included that buttressed the views of the BOTA and hundreds of sources that did not were ignored. The majority of citations went to CRESST research and CRESST researchers. At the time, NRC's Board was chaired by a CRESST director. The "Committee for Appropriate Test Use", the entity responsible for the particular study, included three CRESST grandees and one from Boston College.

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 (UCLA's Jeannie

[3] The Lynch School of Education at Boston College is large and diverse. It houses, for example, a U.S. Education Department center for analyzing international test results and a large higher education research center. The group I refer to here comprises several testing and measurement scholars who work on the regular faculty and have at times called themselves The National Commission on Testing and Public Policy, the Center for the Study of Testing, Evaluation, and Educational Policy (CSTEEP), or The National Board on Educational Testing and Public Policy.

[4] Heubert, J. P., & Hauser, R. P. (Eds.). (1999). *High-stakes: Testing for tracking, promotion, and graduation*. Washington, DC: National Research Council. This section borrows from Phelps, R. P. (2000, December). High Stakes: Testing for Tracking, Promotion, and Graduation, Book review, *Educational and Psychological Measurement, 60*(6), 992–999.

Oakes). The counterevidence and counterarguments on that issue are kept completely hidden from the reader. The early childhood, readiness testing, and promotion and retention sections also feature only one person's point of view (that of Lorrie Shepard of CRESST). Chapter 10 cites only three sources (an earlier NRC report and Shepard and Linn of CRESST). Chapter 11 cites, essentially, only two sources, George Madaus and Walt Haney of Boston College. Two thirds of the citations in the report refer to less than a dozen research sources.

For a book on a psychometric topic, the NRC report strangely ignores psychology research. Only ten citations out of 400 come from psychology journals, and 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 refers exclusively to research in education journals and reports and, even then, only the work of a small group.

The opinions of the general public are dismissed just as effortlessly. The report (pp. 44–45) acknowledges the high level of public support for high-stakes, but discounts it thusly:

> "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."

Then again, it might not. Almost all adults are experienced former students. It so happens that they know something about school.

High Stakes 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. The report even floats a proposal to require that tests be pre-tested, 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, presumably, they would not be during the life of the experiment.

High stakes was released at a propitious time, just before the debate over and design of the No Child Left Behind (NCLB) Act. For those who regarded the National Research Council's work to be objective and trustworthy, it would serve as a caution, and nothing more. A century's

worth of program evaluations and experimental research on the optimal design of high-stakes test-based accountability systems was ignored, relegated to an information abyss. When the nation needed the information most and was most ready to use it, the National Research Council suppressed it.

In response to the NRC's deliberate neglect of the research literature, I began to study it myself. As I lacked the NRC's considerable resources, it took me some time—a decade, as it turned out—to reach a satisfactory stage of completion. I hedge on the word "completion" because I do not believe it possible for one individual to collect all the studies in this enormous research literature …that CRESST officials claim does not exist.

To date, I have read over 3,000 studies, found about a third of them to be appropriate for inclusion in a summary of qualitative studies and meta-analyses of quantitative and survey studies. Most had been available to the NRC study group as well, but were implied not to exist. A summary of the study is published in the *International Journal of Testing* (Phelps, 2012). Source lists can be found here:

http://npe.educationnews.org/Review/Resources/QuantitativeList.htm
http://npe.educationnews.org/Review/Resources/SurveyList.htm
http://npe.educationnews.org/Review/Resources/QualitativeList.htm

Perhaps not surprisingly, a review of a great expanse of the research literature, rather than just the selective, tiny bit covered by the NRC report, leads to quite different conclusions and policy recommendations.

***Common Standards for K–12 Education*? 2008**[5]

Almost two decades ago, while working at the U.S. General Accounting Office (GAO, now called the Government Accountability Office), I managed a study to estimate the extent and cost of standardized testing in the United States. At the time, then-president George H. W. Bush had proposed a national testing program, and the U.S. Congress wanted to know how much it might cost and the effect it might have on then-current state and local testing programs.

On every quality indicator (e.g., survey response rates, fact-checking) the study exceeded GAO norms. A Who's Who of notables in the evaluation, statistical, and psychometric worlds reviewed various aspects of the study. Nothing like it in quality or scale had been done before—it included details from all 48 states with testing programs and from a representative sample of over 500 U.S. school districts. One would think the education research community would have been interested in the results (U.S. General Accounting Office, 1993).

[5] Beatty, A. (2008). *Common Standards for K-12 Education?: Considering the Evidence: Summary of a Workshop Series*. Committee on State Standards in Education, Washington, DC: National Research Council. This section borrows from Phelps, R. P. (2000, Winter). Estimating the cost of systemwide student testing in the United States. *Journal of Education Finance, 25*(3) 343–380, and Phelps, R. P. (1996, Spring). Mis-conceptualizing the costs of large-scale assessment, *Journal of Education Finance, 21*(4) 581–589.

I left the GAO for other employment before the report was actually released, however, and, apparently, the pressure to suppress the report and its findings (essentially, that standardized testing is not that burdensome and does not cost that much) descended even before it was released. Over the ensuing months, I became gradually aware of more efforts to suppress the report's findings. Panels were held at CRESST conferences—panels to which I was not invited—eviscerating it and suggesting that better studies were needed.[6] The characterizations of the GAO report were completely false: the critics claimed that information was left out that, in fact, was not, and that information was included that, in fact, was not. But, reasonable people, allowed to hear only one version of the story, believed it, and the GAO report, along with the most thorough and detailed data base on testing practices ever developed, faded into obscurity.

In its place, other reports were written and presented at conferences, and articles published in mainstream education journals, purporting to show that standardized tests cost an enormous amount and were overwhelming school schedules in their volume. The studies were based on tiny samples, a single field trial in a few schools, a few telephone calls, one state, or, in some cases, the facts were just made up. The cost studies among them that actually used some data for evidence tended to heap all sorts of non-test activities into the basket and call them costs of tests.

I contacted the researchers making the erroneous claims and the CRESST directors by email, postal letter, and telephone.[7] In a few cases, I received assurances, first, that the matter would be looked into—it was not—and, second, that an *erratum* would be published in the CRESST newsletter—it never was.

I submitted articles based on the GAO study to mainstream education journals and they were rejected for outlandish and picayune reasons, or because "everyone knows" that the GAO report was flawed.

Ultimately, a summary of the GAO report won a national prize and was published in a finance journal (Phelps, 2000, Winter). I suspect, however, that if the GAO report had arrived at "correct" conclusions (i.e., that standardized tests are enormously expensive and otherwise bad) any article derived from it could easily have been published several years earlier in most any mainstream education journal.

---

[6] For example, 1993 CRESST Conference: "Assessment Questions: Equity Answers: What Will Performance Assessment Cost?" Monday, September 13; 1994 CRESST Conference: "Getting Assessment Right: Practical and Cost Issues in Implementing Performance Assessment", Tuesday, September 13; 1995 CRESST Conference: "Assessment at the Crossroads: What are the Costs of Performance Assessment?", Tuesday, September 12. CRESST report #441 still contains mostly erroneous claims related to the GAO report, on pages 5 and 64–66, and mostly erroneous claims about CRESST's work on the issue, in the first seventeen pages.

[7] Among the researchers directly contacted (by email, letter, and telephone) were: Picus, Monk, and Tralli. Organizations directly contacted were the federally funded, taxpayer-supported CRESST and CPRE (for Center for Policy Research in Education, based at U. Wisconsin and U. Penn), and the U.S. Education Department. Individuals at those organizations directly contacted included: Baker, Dietel, Linn, Resnick, Odden, and Sweet. With the exception of Lauren Resnick, who treated the matter in a professional way, my appeals were met with years of inaction and animosity.

One would think that the assault on the GAO study might have ended in the 1990s, given how successful it was. But, perhaps, the report's quality or the GAO name is so durable that education insiders feel the need to condemn it even fifteen years later, as they have in the National Research Council report *Common Standards for K–12 Education?*

To my observation, the CRESST and NRC preferred practice for information suppression is to ignore or declare nonexistent any research that contradicts theirs. There are several advantages to this practice, the "dismissive review" (see Phelps, 2009):

first, it is easier to win a debate with no apparent opponent;

second, declaring information nonexistent discourages efforts to look for it;

third, because it is non-confrontational, it seems benign and not antagonistic; and

fourth, there is plausible deniability, i.e., one can simply claim that one did not know about the other research.

When only one side gets to talk, of course, it can say pretty much anything it pleases. With no counterpoint apparent, "facts" can be made up out of thin air, with no evidence required. Solid research supportive of opposing viewpoints is simply ignored, as if it did not exist. It is not mentioned to reporters, it is not cited in footnotes or reference lists. It is treated as if it was never done.

Dismissive reviews are not credible to outsiders, however, when contradictory research is widely known to exist. Thus, the research that remains—that which cannot credibly be dismissed as nonexistent—must, instead, be discredited. In such cases, the preference for dismissive reviews must be set aside in favor of an alternate strategy: misrepresent the disliked study and/or impugn the motives or character of its author.

And, so it has been with the GAO report on testing costs. The GAO manifests a prominent profile and a reputation that is not easily demeaned. Nonetheless, the federally-funded research center CRESST has worked tirelessly for two decades now to do exactly that. And, in that quest, CRESST has had a distinct advantage: it is mandated and funded to disseminate its findings, whereas the GAO is not. Once a GAO report is released and a GAO official testifies to its only client—the U.S. Congress—no further agency effort promotes the work. By contrast, CRESST's mission and funding include promotion of its work through marketing and conferences.

This 2008 NRC report, released fifteen years after the GAO report on testing costs, asserts, again, that the GAO report left something out and so underestimated the cost of testing.[8] And,

[8] On pp. 8–9 of the background paper "The Resource Costs of Standards, Assessments, and Accountability" (Harris & Taylor, 2008) one reads "On the other hand, neither Phelps nor the GAO study ascribes any costs to standard setting...."

again, the assertion is false. This time, the NRC accused the GAO of neglecting to consider the cost of standard-setting during test development; in fact, this cost was fully counted in the GAO estimate.

But, yet again, claiming a void in others' calculations is used as an excuse to bulk up their own cost estimates massively. Here are just a few ways that the NRC report, *Common Standards for K–12 Education?*, overestimates the cost of testing:

- One-time-only start-up costs—e.g., standard setting—are counted as annual recurring costs.

- Educator travel and lodging expenses for serving on standard-setting and other test development panels are counted twice, both as direct educator expenses and in the budget of the state education agency (which, in fact, reimburses the educators for these expenses).

- The full duration of all testing activities at a school—said to be 3–5 days—is allotted to each and every educator participating. So, for example, the time of a fifth grade teacher who administers a one-hour math exam on Tuesday of testing week, and who otherwise teaches regular class that week, is counted as if s/he were involved in administering each and every exam in every subject area and at every grade level throughout the entire 3–5 days. Moreover, the time of each and every teacher in the school is counted as if each and every teacher is present in each and every testing room for all subject areas and grade levels. By this method, the NRC overestimates the amount of educator time spent directly administering tests about twenty-fold.

  Another way of looking at it is to ignore the fact that a school administers a series of one-hour tests across the tested subject areas and grade levels over the span of 3–5 days but, instead, assume that all classes in all subject areas and grade levels are sitting for 3–5 days doing nothing but taking 3–5-day-long exams, which, in fact, is not what happens.

- The NRC calculates the number of teachers involved by using a federally-estimated average pupil-teacher ratio, rather than an average class size estimate. Pupil-teacher ratios underestimate class sizes because they include the time of teachers when they are not teaching. By this method, the NRC overestimates the number of teachers involved in directly administering tests by another 50%.

- The NRC counts all teachers in a school, even though only those in certain grade levels and subject areas are involved in testing—usually amounting to fewer than half a school's teachers. By this method, the NRC overestimates the number of teachers involved in directly administering tests by another 50% or more.

- In calculating "data administration costs" of processing test data in school districts and states, the NRC classifies all who work in these offices as "management, business, and financial" professionals who make $90,000/year. Anyone who has worked in state and local government data processing departments knows that this would grossly overestimate the real wages of the majority of these employees who, essentially, work as clerical and, oftentimes, contingent staff.

- The NRC is told by one school district that their average teacher spends 20 hours every year in professional development related to assessment and accountability. Despite how preposterous this number should sound, this one piece of hearsay is used by the NRC to estimate the amount of time all teachers everywhere, whether involved in testing or not, spend annually in related professional development.

- Moreover, professional development related to testing and accountability is assumed to be unrelated to regular instruction and, so, is counted as a completely separate, added-on (i.e., marginal) cost.

- The NRC counts educator time working on standard-setting and other test development panels as "two or three days" which, as anyone who has worked in test development knows, is a high estimate. One to two days is more realistic.

Finally, the NRC studies testing and accountability in only several school districts in only three states. But, according to them, the GAO report which analyzed details from all 48 states involved in testing and over 500 school districts …is the study that left stuff out. In the end, the NRC estimates for testing and accountability costs, are in their own words "about six times higher" than previous estimates.

***Incentives and Test-Based Accountability in Education*, 2011**[9]

[9] Hout, M., & Elliott, S.W. (2011). *Incentives and test-based accountability in education*. Committee on Incentives and Test-Based Accountability in Public Education. Board on Testing and Assessment, Division of Behavioral and Social Sciences and Education, National Research Council. Washington, DC: The National Academies Press. This section borrows from Phelps, R. P. (2008/2009b). The rocky score-line of Lake Wobegon. Appendix C in R. P. Phelps (Ed.), *Correcting fallacies about educational and psychological testing*, Washington, DC: American Psychological Association. < http://supp.apa.org/books/Correcting-Fallacies/appendix-c.pdf > ; Staradamskis, P. (2008, Fall). Measuring up: What educational testing really tells us. Book review, *Educational Horizons, 87*(1). Available at: http://www.nonpartisaneducation.org/Foundation/KoretzReview.htm ; Phelps, R. P. (2010, July). The source of Lake Wobegon [updated]. *Nonpartisan Education Review / Articles, 1*(2). Available at: http://www.nonpartisaneducation.org/Review/Articles/v1n2.htm ; Phelps, R. P. (2011). Extended Comments on the Draft Standards for Educational & Psychological Testing (But, in particular, Draft Chapters 9, 12, & 13). *Nonpartisan Education Review/Essays, 7*(3). Available at: http://www.nonpartisaneducation.org/Review/Essays/v7n3.htm ; Phelps, R. P. (2011). Educators cheating on tests is nothing new; Doing something about it would be. *Nonpartisan Education Review/Essays, 7*(5). Available at: http://www.nonpartisaneducation.org/Review/Essays/v7n5.htm

"The celebrity professor is a new phenomenon and not a good one. In celebrity-driven academia, 'getting ahead' means beating other people, which means establishing a personal reputation and denying it, to the extent possible, to rivals." — Harry Lewis

By coincidence, a draft copy of this most current National Research Council study was released about the same time I was writing up the results of my decade-long research summary and meta-analyses of the effect of testing on student achievement. So, naturally, I was interested to see how much more the resource-rich NRC could do with the same material. As it turns out, they have remained true to form—this report covers only a highly-selective, tiny fraction of the research literature—but implies that it is all there is to be found—and ignores or declares nonexistent the vast majority of relevant research.

For my work, I looked at studies published in English between 1910 and 2010 that I could obtain and review before a self-imposed deadline in 2010. My coverage of the research literature is far from complete. It includes 244 qualitative studies (e.g., direct observations, site visits, interviews, case studies), 813 individual item-response group combinations from 247 survey studies (e.g., program evaluation surveys, opinion polls), and 640 separate measurements of effects from 177 quantitative research studies (e.g., regression analysis, structural equation modeling, pre-post comparison, experimental design, or interrupted time series design). In total, I analyzed 1,671 separate effects from 668 studies.

The domain of coverage for the NRC study is nominally larger than mine because they purported to analyze non-test incentives and effects on outcomes other than achievement. Nonetheless, they include but 18 studies in their analysis.

They whittle down the number by eliminating from consideration all qualitative and survey studies and studies conducted prior to the past two decades. For "pre-NCLB" studies from the 1990s to the mid-2000s, the NRC simply expropriates a meta-analysis conducted by Jaekyung Lee (2008) that covered only 14 "cross-state", hugely large-scale studies, the type least likely to find a strong effect. Unlike small-scale studies and, particularly, experiments that can focus on the factor of interest, empirical studies of large-scale testing programs comprise hundreds of factors for programs with multiple goals and objectives. In statistical lingo, such studies are full of "noise".

Lee calculated an average effect size across the 14 cross-state studies of 0.08, a positive, but very weak effect. And, that is what the NRC goes with. They conclude that, prior to the date the NCLB Act's stakes kicked in (about 2006) studies of the effect of testing on student achievement found a 0.08 average effect size.

I include all of Lee's studies, and his calculations, in my own analysis of quantitative studies that meet the NRC criteria for inclusion, but I also include a multitude of studies the NRC deliberately leaves out. Counting relevant studies only from the same time period—1990 to 2005—the mean effect size is 0.82, ten times larger than the NRC's. Moreover, that is the simple, "bare-bones" effect size, unadjusted for measurement artifacts—adjustments that would make it even larger.

For the post-NCLB period, the NRC includes seven other large-scale studies that accumulate feeble effect sizes, and 13 studies published between 2002 and 2010 of “incentive experiments using rewards” from India, Israel, Kenya, and the United States.

On the whole, the NRC selection of studies is quite odd, and ridiculously unrepresentative of the research literature it purports to summarize. One consistency in the selection is apparent, however. Only studies finding very small effects are included.

In my review of the 1999 study, *High Stakes*, I criticized the NRC for restricting its literature survey to U.S. education research, ignoring relevant research conducted in other countries and in other disciplines, such as psychology and economics.

Behold. This newer report mentions some relevant work conducted overseas and by psychologists and economists. Still, the sample is highly selective and excludes the most seminal work in the field. The NRC finds a small group of work that reflects the in-group bias, and the larger world of research and researchers is ignored as if it did not exist.

Most notably, in addition to the usual education researchers, this NRC study covers the work of a crew of young economists, who reach the preferred conclusions (of feeble effect sizes). It turns out that this group of economists shares another characteristic in common with the NRC veterans from CRESST and Boston College—dismissive reviews.

One economist whose work is discussed at length in *Incentives and Test-Based Accountability in Education* claims to have conducted the first systematic empirical study of teacher cheating (in the early 2000s), the first case study of an urban school district comprehensive accountability system (in 2003), one of the first studies of school-based accountability utilizing individual student data (in 2002), one of the first studies of high-stakes testing (in 2002), and one of the first studies of the effect of grade-promotion testing (in 2002). The same fellow declared (in 2001) there to be no empirical research on minimum competency testing programs or high school graduation exams. The research literature fluffed off by just this one person includes several hundred studies dating back to the 1910s.

Another economist well-regarded by the NRC also declared (in 2002) there to be little to no evidence of the effects of testing or accountability systems. A third economist declared (in 1999) there to be little to no empirical work on school-based incentive programs. A fourth declared (in 1996) "Virtually no evidence exists about the merits or flaws of MCTs [minimum competency tests]". A fifth claimed (in 2005) “there is almost no research on the impact of remediation on student outcomes.” A sixth claimed (in 2000) that a paper he had just written “…provides the first empirical evidence on the effects of grading standards, measured at the teacher level."

Whereas all but a trivial amount of the great mass of relevant research is ignored, the work of NRC study committee members is cited liberally. Daniel Koretz wins the prize for the most citations and references with twelve and nine. Overall, 48 citations and 40 references (out of

200) go to committee members' work. Over 30 references cite CRESST work. The bulk of the rest cite the work of close friends and colleagues, or earlier NRC studies.

At the same time, a Who's Who of the leading researchers in the field, past and present, goes missing, names such as: John Hattie, Roddy Roediger, John Bishop, Frank Schmidt, W.J. Haynie, Harold Wenglinsky, Linda Winfield, C.C. Ross, E.H. Jones, Mike McDaniel, Lorin Anderson, J.R. Nation, J.H. Block, Carol Parke, S.F. Stager, Arlen Gullickson, Lynn Fuchs, Douglas Fuchs, Kathy Green, Max Eckstein, Harold Noah, Benjamin Bloom, Jeffrey Karpicke, Michael Beck, Stephen Heynemann, David Driscoll, William D. Schafer, Francine Hultgren, Willis Hawley, James H. McMillan, Elizabeth Marsh, Susan Brookhart, Gene Bottoms, Gordon Cawelti, Lorna Earl, Mike Smoker, David Grissmer, Arthur Powell, Harold Stevenson, Hunter Boylan, Elana Shohamy, Aletta Grisay, Chris Whetton, Steve Ferrara, Glynn Ligon, Micheline Perrin, Thomas Fischer, A. Graham Down, Nigel Brooke, John Oxenham, Caroline Gipps, Arthur Hughes, D. Pennycuick, John Poggio, Anthony Somerset, John O. Anderson, Noel McGinn, Anne Anastasi, Nick Theobald, David Miller, Linda Bond, Nancy Protheroe, Floraline Stevens, Thomas Corcoran, Clement Stone, Suzanne Lane, Frank Dempster, and state agencies in Massachusetts, Florida, and South Carolina.

And, those are just names of some folk who have conducted one or more individual studies. Others have summarized batches of several to many studies in meta-analyses or literature reviews, for example (in chronological order): Panlasigui (1928); Ross (1942); Kirkland (1971); Proger & Mann (1973); Jones (1974); Bjork (1975); Peckham & Roe (1977); Wildemuth (1977); Jackson & Battiste (1978); Kulik, Kulik, Bangert-Drowns, & Schwalb (1983–1991); Natriello & Dornbusch (1984); Dawson & Dawson (1985); Levine (1985); Resnick & Resnick (1985); Guskey & Gates (1986); Hembree (1987); Crooks (1988); Dempster (1991); Adams & Chapman (2002); Locke & Latham (2002); Roediger & Karpicke (2006); and Basol & Johanson (2009).

Long lists of many more relevant names and studies that, in most cases, accumulated results unwanted by CRESST and NRC researchers can be found in Phelps 2003, 2005, 2007, and 2008/2009a.

You will find none of this research and none of these researchers mentioned in *Incentives and Test-Based Accountability in Education*. Yet, the report claims to summarize the relevant literature. Meanwhile, as in earlier NRC reports, this one declares that important research questions remain unanswered, the implication being that these dismissive reviewers should be given millions of dollars to do the research they have declared nonexistent.

Finally, this NRC report advances its pet theory of "test-score inflation", while excluding the full abundance of counterevidence, thus recommending exactly the wrong policy to address a very serious and very topical problem (see Phelps, 2011a, 2011b).

In the 1980s, a young medical resident working in a high-poverty region of West Virginia heard local school officials claim that their children scored above the national average on standardized tests. Skeptical, he investigated further and ultimately discovered that every U.S.

state administering nationally-normed tests claimed to score above average, a statistical impossibility. The phenomenon was tagged the "Lake Wobegon Effect" after Garrison Keillor's "News from Lake Wobegon" radio comedy sketch, in which "all the children are above average".

The West Virginia doctor, John Jacob Cannell, M.D., would move on to New Mexico and, eventually, California, but not before documenting his investigations in two self-published books with titles, "How All Fifty States Are above the National Average" (1987) and "How Public Educators Cheat on Standardized Achievement Tests" (1989).

As usually happens after newsworthy school scandals, policy makers and policy commentators expressed disapproval, wrote opinion pieces, formed committees, and, in due course, forgot about it. Deep dives into the topic were left to professional education researchers, the vast majority of whom worked then, as now, as faculty at graduate schools of education, where they shared a vested interest in defending the *status quo*.

Dr. Cannell cited educator dishonesty and lax security in test administrations as the primary culprits in the Lake Wobegon Effect, also known as "test score inflation" or "artificial test score gains". It is easy to understand why. Back then, it was common for states and school districts to purchase nationally-normed standardized tests "off the shelf" and handle all aspects of test administration themselves. Moreover, to reduce costs, it was common to reuse the same test forms (and test items) year after year. Even if educators did not intentionally cheat, over time they became familiar with the test forms and items and could easily prepare their students for them. With test scores rising over time, administrators and elected officials could claim credit for increasing learning.

Test security was so lax because they were diagnostic and monitoring tests that "didn't count"—only one of the dozens of state tests Cannell examined was both nationally-normed and "high-stakes"—involving direct consequences for the educators or students involved.

Regardless the fact that there were no stakes attached to Cannell's tests, however, prominent education researchers blamed "high stakes" for the test-score inflation he found (Koretz, et al, 1991; Koretz, 2008). Cannell had exhorted the nation to pay attention to a serious problem of educator dishonesty and lax test security, but education insiders co-opted his discovery and turned it to their own advantage (Staradamskis, 2008; Phelps, 2008/2009b, 2010).

"There are many reasons for the Lake Wobegon Effect, most of which are less sinister than those emphasized by Cannell," said the co-director of CRESST (Linn, 2000, p. 7). After Dr. Cannell left the debate and went on to practice medicine, this federally-funded education professor and his colleagues would repeat the mantra many times—high stakes, not lax security, cause test-score inflation.

It is most astonishing that they stick with the notion because it is so obviously wrong. The SAT and ACT are tests with stakes—one's score on either helps determine which college one attends. But, they show no evidence of test-score inflation. (Indeed, the SAT was re-centered in the 1990s because of score deflation.) The most high-stakes tests of all—occupational licensure tests—show no evidence of test-score inflation. Both licensure tests and the SAT and ACT, however, are administered with tight security and ample test form and item rotation.

Spot the causal factor

	High security	Lax security
High stakes	No test-score inflation e.g., SAT, ACT, licensure examinations	Test-score inflation possible e.g., some internally administered district and state examinations
No/Low stakes	No test-score inflation e.g., NAEP, other externally administered examinations	Test-score inflation possible e.g., some internally administered district and state examinations, such as those Cannell investigated

Current test cheating scandals in Washington, DC, Atlanta, and Pennsylvania once again draw attention to a serious problem, and this time there is no doubt that stakes are involved. With the No Child Left Behind Act, schools can be rewarded with cash, or punished through reconstitution or closure, depending on their students' test scores. So, as they have now for over two decades, most educators blame the stakes and alleged undue pressure that ensues for the cheating. Their recommendation: drop the stakes and reduce the amount of testing.

Meanwhile, twenty years after J. J. Cannell first showed us how lax security corrupts test scores, regardless the stakes, test security remains cavalierly loose. We have teachers administering tests in their own classrooms to their own students, principals distributing and collecting test forms in their own schools. Security may be high outside the schoolhouse door, but inside, too much is left to chance. And, as it turns out, educators are as human as the rest of us; some of them cheat and not all of them manage to keep test materials secure, even when they aren't cheating.

The furor over educator cheating scandals in Atlanta and Washington, DC could lead to real progress on test security reform so long as the vested interests do not continue to control the debate and determine the policy outcome as they have with Dr. Cannell's legacy.

And, they are trying to. In *Incentives and Test-based Accountability in Education*, the National Research Council again asserts a causal relationship between stakes and test-score inflation and ignores test security's role. Their solution to the problem is not to increase

security, but to administer no-stakes "audit tests" to shadow the high-stakes test administration over time, under the presumption that any no-stakes test's scores are trustworthy and incorruptible. Thus, resources that could be used to bolster the security of the test that counts will be diverted instead toward the development and administration of a test that doesn't. That other test that doesn't count will almost certainly be administered with little security by school officials themselves.

With any high-stakes test subject to audit by any low-stakes test, its perceived quality will be determined entirely by the low-stakes test. Indeed, those who oppose high-stakes testing could add an easily manipulated and unmonitored low-stakes test and tailor it to discredit score gains on their jurisdiction's externally-mandated and monitored high-stakes test.

Even worse, the same education researchers who have co-opted federally-funded and National Research Council work on educational testing are attempting to compromise the *Standards for Educational and Psychological Testing* which, after more than a decade is currently being revised. The *Standards* is a set of guidelines for developing and administering tests. In the absence of any good alternative it has been used by the courts as a semi-official code of conduct. Thus, it has profound impact beyond the boundaries of the relatively tiny community of testing professionals. The education insiders have incorporated into the draft revision of the *Standards* their notion that stakes, not lax security, cause test-score inflation and audit tests are the way to control it. Meanwhile, in over 300 pages, the draft *Standards* says next to nothing about test security.

The most fundamental issues in these school scandals are neither cheating, nor pressure, nor testing; they are power and control. Standardized test scores will be trustworthy if responsible external authorities control their administration. It is that simple. Leave control of testing, or "audit testing", to school administrators themselves, and wide-scale institutionalized cheating on educational tests will be with us forever.

Conclusion

The latest report sponsored by the Board on Testing and Assessment, Division of Behavioral and Social Sciences and Education, of the National Research Council faithfully continues a quarter-century tradition of bias, obfuscation, misinformation, and dissemblage. The National Academies describes its study process reassuringly:

> "The reports of the National Academies are viewed as being valuable and credible because of the institution's reputation for providing independent, objective, and non-partisan advice with high standards of scientific and technical quality. Checks and balances are applied at every step in the study process to protect the integrity of the reports and to maintain public confidence in them."

This description may validly describe reports that the National Academies produce on other topics. Since 1989, however, the several reports under their nameplate, the National Research Council (NRC), on standardized testing have been anything but—neither independent,

objective, nor balanced. Rather, they have been partisan reports, with no checks on rampant, self-interested bias.

But, bias isn't the only problem; the process is corrupt. This particular type of corruption does not involve money. The currency of scholars is attention, with the "richest" among them achieving the most—genuine fame—celebrity status that floods a confluence of honors, awards, and remuneration streams.

The NRC reports mentioned above are not just used to proselytize and mislead; more emphatically, they are expropriated to showcase the careers of those involved. At the same time the report authors declare the work of other researchers nonexistent, they liberally cite their own work and that of their close friends, and package the combination as if it were all that anyone who matters should care for.

The behavior is arrogant. It is also unethical, dishonest, and cowardly. Nonetheless, it has worked efficiently to gloriously advance the professional careers of the few researchers inside the NRC tent and to relegate massive research literatures to oblivion.

Journalists, unfortunately, simply assume that those who get the most attention in the research world are also the most deserving of that attention. They simply assume that education research dissemination is objective and fair. They couldn't be more wrong.

But, some journalists step further into an ethical abyss—they help promote dismissive reviews. No journalist has the time to validate such claims; it can take years to learn a research literature. So, every time a journalist writes "there is a paucity of research on this topic", or the like, they're just taking one self-interested person's word for it. Every time a journalist writes "there is little research in this area" or "so-and-so's study is the first of its kind" they are complicit in the corruption.

The capture of the National Research Council's BOTA by vested interests and the tragic results illustrate how federal money can concentrate power to achieve exactly the opposite result from that intended. For a quarter-century, U.S. taxpayers have funded just one research center to study educational testing, the Center for Research on Educational Standards and Student Testing (CRESST). Its mandate is to objectively review all the research available on the topic; instead it promotes its own and declares most of the rest nonexistent. Its mandate is to serve the interests of all the U.S. taxpayers who fund its operations; instead it serves the interests of its own members and that of the education *status quo*.

Few experts in education research or testing are willing to criticize the work of CRESST grandees, even when the flaws are obvious. CRESST officials are too powerful, and can too easily wreck a career. Several CRESST officials have been elected president of the American Educational Research Association (AERA) and most CRESST researchers are well represented on powerful and well-funded boards, commissions, and committees, like those at the National

Research Council. The current arrangement works very well for them; they are unlikely to initiate any effort to change it.

Until those in positions of responsibility who can distinguish right from wrong are willing to take a stand, CRESST folk will continue to eradicate the vast bulk of a century's worth of research on educational testing and accountability, and replace it with the very warped bit of their own divining. And, we taxpayers will pay them to do it.

References

Beatty, A. (2008). *Common Standards for K-12 Education?: Considering the Evidence: Summary of a Workshop Series*. Committee on State Standards in Education, Washington, DC: National Research Council.

Cannell, J.J. (1987). *Nationally normed elementary achievement testing in America's public schools: How all fifty states are above the national average*. (2nd Ed.), Daniels, WV, USA: Friends for Education.

Cannell, J.J. (1989). *How public educators cheat on standardized achievement tests*. Albuquerque, NM, USA: Friends for Education.

Harris, D.N., & Taylor, L.L. (2008, March 10). *The Resource Costs of Standards, Assessments, and Accountability: A Final Report to the National Research Council*.

Hartigan, J.A. & Wigdor, A.K. (1989). *Fairness in Employment Testing: Validity Generalization, Minority Issues, and the General Aptitude Test Battery*. Washington, D.C.: National Academy Press.

Heubert, J.P. & Hauser, R.M. (1999). *High Stakes: Testing for Tracking, Promotion, and Graduation*. Washington, D.C.: National Academy Press.

Hout, M., & Elliott, S.W. (2011). *Incentives and test-based accountability in education.* Committee on Incentives and Test-Based Accountability in Public Education. Board on Testing and Assessment, Division of Behavioral and Social Sciences and Education, National Research Council. Washington, DC: The National Academies Press.

Koretz, D.M. (2008). *Measuring up: What educational testing really tells us*. Cambridge, MA: Harvard University Press.

Koretz, D.M., Linn, R.L., Dunbar, S.B., & Shepard, L.A. (1991). The effects of high-stakes testing on achievement: Preliminary findings about generalization across tests. Paper presented in R.L. Linn (Chair), Effects of High-Stakes Educational Testing on Instruction and Achievement, symposium presented at the annual meeting of the American Educational Research Association, Chicago, April 5.

Lee, J. (2008). Is Test-driven External Accountability Effective? Synthesizing the Evidence from Cross-State Causal-Comparative and Correlational Studies. *Review of Educational Research, 78*(3), 608–644.

Linn, R.L. (2000). Assessments and accountability. *Educational Researcher*. March, 4–16.

Phelps, R.P. (1996, Spring). Mis-conceptualizing the costs of large-scale assessment, *Journal of Education Finance, 21*(4) 581–589.

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

Phelps, R.P. (2000, December). High stakes: Testing for tracking, promotion, and graduation, Book review, *Educational and Psychological Measurement, 60*(6), 992–999.

Phelps, R.P. (2000, Winter). Estimating the cost of systemwide student testing in the United States. *Journal of Education Finance, 25*(3) 343–380.

Phelps, R.P. (2003). *Kill the messenger: The war on standardized testing*. New Brunswick, NJ, USA: Transaction Publishers.

Phelps, R.P. (2005). The rich, robust research literature on testing's achievement benefits. Chapter 3 in Phelps, R. P., Ed. *Defending standardized testing*. Mahwah, NJ, USA: Lawrence Erlbaum.

Phelps, R.P. (2007). *Standardized testing primer*. New York, NY, USA: Peter Lang.

Phelps, R.P. (2008/2009a). Educational achievement testing: Critiques and rebuttals. In R. P. Phelps (Ed.), *Correcting fallacies about educational and psychological testing*, Washington, DC: American Psychological Association.

Phelps, R.P. (2008/2009b). The rocky score-line of Lake Wobegon. Appendix C in R. P. Phelps (Ed.), *Correcting fallacies about educational and psychological testing*, Washington, DC: American Psychological Association. Available at: http://supp.apa.org/books/Correcting-Fallacies/appendix-c.pdf

Phelps, R.P. (2008/2009c). The National Research Council's Testing Expertise, Appendix D in R. P. Phelps (Ed.), *Correcting fallacies about educational and psychological testing*, Washington, DC: American Psychological Association. Available at: http://supp.apa.org/books/Correcting-Fallacies/appendix-d.pdf

Phelps, R.P. (2009, November). Worse than plagiarism? Firstness claims and dismissive reviews. (slide show). *Nonpartisan Education Review / Resources.* Available at: http://www.nonpartisaneducation.org/Review/Resources/WorseThanPlagiarism.htm

Phelps, R.P. (2010, July). The source of Lake Wobegon [updated]. *Nonpartisan Education Review / Articles, 1*(2). Available at: http://www.nonpartisaneducation.org/Review/Articles/v1n2.htm

Phelps, R.P. (2011a). Extended Comments on the Draft *Standards for Educational & Psychological Testing* (But, in particular, Draft Chapters 9, 12, & 13). *Nonpartisan Education Review/Essays, 7*(3). Available at: http://www.nonpartisaneducation.org/Review/Essays/v7n3.htm

Phelps, R.P. (2011b). Educators cheating on tests is nothing new; Doing something about it would be. *Nonpartisan Education Review/Essays, 7*(5). Available at: http://www.nonpartisaneducation.org/Review/Essays/v7n5.htm

Phelps, R.P. (2012). The effect of testing on student achievement, 1910–2010 *International Journal of Testing, 12*(1), 21-43, International Test Commission.

Staradamskis, P. (2008, Fall). Measuring up: What educational testing really tells us. Book review, *Educational Horizons, 87*(1). Available at: http://www.nonpartisaneducation.org/Foundation/KoretzReview.htm

U.S. General Accounting Office. (1993). *Student testing: Current extent and expenditures, with cost estimates for a national examination*. GAO/PEMD-93-8, U.S. Congress.

Wise, L. (1998). Personal communication, Alexandria, VA.