The Rot Spreads Worldwide: The OECD—Taken In and Taking Sides

OECD encourages world to adopt failed U.S. education programs

by Richard P. Phelps

In the aftermath of Hurricane Katrina, one of the United States' largest insurers refused to honor damage claims from Gulf Coast customers who submitted hurricane insurance claims, asserting that their property had been damaged not by hurricane, but by flooding. Only a high-stakes, high-profile class-action lawsuit ultimately pried the insurance payments loose. Currently, this large insurance company, with its own trust issues, is running a series of television commercials poking fun at an institution that

it assumes is trusted by the public even less—the Internet. "They wouldn't put it on the Internet if it wasn't true," says the naïve foil who purchased allegedly inferior insurance after believing the promises in an Internet advertisement, presumably eliciting off-screen laughter in millions of living rooms.

Now suppose that you are responsible for learning the "state of the art" in the research literature on an important, politically sensitive, and hotly contested public policy topic. You can save money by hiring master's-level public policy students or recent graduates, though none with any particular knowledge or experience in the topic at hand—a highly specialized topic with its own doctoral-level training, occupational specializations, and vocabulary. You give your public policy masters a computer with an Internet browser and ask them to complete their reports within a few months. What would you expect them to produce?

You can see for yourself at the website of the Organisation for Economic Cooperation and Development (OECD).[1] In 2009 the OECD launched the Review on Evaluation and Assessment Frameworks for Improving School Outcomes. Apparently the "Review" has not claimed an acronym. In my own interest, then, I give it one—REAFISO.[2]

In its own words, REAFISO was created to provide analysis and policy advice to countries on the following overarching policy question:

> "How can assessment and evaluation policies work together more effectively to improve student outcomes in primary and secondary schools?"

To answer this question, the OECD intended to

> look at the various components of assessment and evaluation frameworks that countries use with the objective of improving the student outcomes produced by schools. . . .

and

> extend and add value to the existing body of international work on evaluation and assessment policies.

I was interested in REAFISO's work for two reasons. First, I once worked for the OECD, on two fixed-length consulting contracts

accumulating to sixteen months. I admired and respected their education research work and thoroughly enjoyed my time outside work hours. (The OECD is based in Paris.) I particularly appreciate the OECD's education (statistical) indicators initiatives.

Second, I have worked myself and on my own time to address the overarching question they pose, ultimately publishing a meta-analysis and research summary of the effect of testing on student achievement. Because I lacked the OECD'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.

To date, after reading more than three thousand studies, I have found about a third of them appropriate for inclusion in a summary of qualitative studies and meta-analyses of quantitative and survey studies. 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, which is far from complete, 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.

A summary has been published in the *International Journal of Testing* (Phelps, 2012b). Source lists can be found at these three virtual locations:

http://www.nonpartisaneducation.org/Review/Resources/QuantitativeList.htm

http://www.nonpartisaneducation.org/Review/Resources/SurveyList.htm

http://www.nonpartisaneducation.org/Review/Resources/QualitativeList.htm

All but a couple of these several hundred sources were available to REAFISO as well. Yet despite having many times the resources at their disposal, they managed to find just a few percent of what I found. Granted, the search parameters (as best I can discern theirs) were not exactly the same, but were far more alike than different. Not surprisingly, a review of a great expanse of the research literature,

rather than just the selective, tiny bit covered by REAFISO, leads to quite different conclusions and policy recommendations.

Deficiencies of the OECD's REAFISO research reviews include:

- overwhelming dependence on U.S. sources;
- overwhelming dependence on inexpensive, easily-found documents;
- overwhelming dependence on the work of economists and education professors;
- wholesale neglect of the relevant literature in psychology (the social science that invented cognitive assessment) and from practicing assessment and measurement professionals; and
- wholesale neglect of the majority of pertinent research.

Moreover, it seems that REAFISO has fully aligned itself with a single faction within the heterogeneous universe of education research—the radical constructivists. Has the OECD joined the U.S. education establishment? One wouldn't think that it had the same (self-) interests. Yet canon by canon by canon, REAFISO's work seems to subscribe to U.S. education establishment dogma. For example, in her report "Assessment and Innovation in Education," Janet Looney writes

> Innovation is a key driver of economic and social programs in OECD economies. If absent, innovation growth stalls; economies and communities stagnate. . . . (p. 6)
>
> Teaching and learning approaches considered as innovation, on the other hand, are generally characterized as being "student-centered," or "constructivist." (p. 8)
>
> This report has focused on [the] impact of high-stakes assessment and examinations on educational innovation. It has found significant evidence that such assessments and examinations undermine innovation. (p. 23)

First, Looney equates national economies and school classrooms. Then she adds the famous economist Joseph Schumpeter's definition of innovation as "creative destruction." For radical constructivists, and apparently for Looney, each teacher is a unique craftsperson in a unique classroom, and anything done to standardize their work inhibits their potential to guide each unique student in his or her

own unique, natural discovery of knowledge. To radical constructivists, there are no economies of scale or scope in learning.

Whereas innovation is a holy commandment for the U.S. education professoriate, critics charge that it leads to a continuous cycle of fad after fad after fad. After all, if innovation is always good, then any program that has been around for a while must be bad, no matter how successful it might be in improving student achievement. Moreover, if the pace of today's-innovation-replacing-yesterday's-innovation proceeds fast enough, evaluation reports are finished well after one program has been replaced by another, become irrelevant before they are published, and end up unread. Ultimately, in a rapidly innovating environment, we learn nothing about what works. Some critics of the radical constructivists suspect that that chaotic, swirling maelstrom may be their desired equilibrium state.

A case in point is the sad and expensive 1990s saga of the New Standards Project in the United States and the most deliberate effort to implement its assessment formula in practice, the State of Maryland's MSPAP (for Maryland School Performance Assessment Program). REAFISO writer Allison Morris (p. 16) cites Thomas Toch's (2006) erroneous assertion that cost considerations reversed the 1980s–1990s U.S. trend toward more performance testing. Not so, the MSPAP and similar programs (e.g., CLAS [California Learning Assessment System] and KIRIS [Kentucky Instructional Results Information System]) failed because test reliability was so low, test scores were too volatile to be useful, feedback was too late and too vague to be useful, and parents thought it unfair when their children's grades depended on other students' efforts (in collaborative group activities).

Resounding public disgust killed those programs. But ten years is a long time in the ever-innovating world of U.S. education policy, long enough for the young REAFISO writers to be unaware of the fiascos. REAFISO now urges us to return to the glorious past of New Standards, MSPAP, CLAS, and KIRIS, dysfunctional programs that, when implemented, were overwhelmingly rejected by citizens, politicians, and measurement professionals alike.

What happened at the OECD?

REAFISO relies on staff generalists and itinerant workers to compose its most essential reports. I suspect that the REAFISO writers started out unknowing, trusted the research work they found most

easily, and followed in the direction those researchers pointed them. Ultimately, they relied on the most easily and inexpensively gathered document sources.

I believe that REAFISO got caught in a one-way trap or, as others might term it: a bubble, echo chamber, infinite (feedback) loop, or myopia. They began their study with the work of celebrity researchers—dismissive reviewers—researchers who ignore (or declare nonexistent) those researchers and that research which contradicts their own (Phelps, 2012a)—and never found their way out. Dismissive reviewers blow bubbles, construct echo chambers, and program infinite loops by acknowledging only research and those researchers they like or agree with.

The research most prominently listed in Internet searches for REAFISO's topics of interest is, as with most topics on the Internet, that produced by groups with the money and power to push theirs ahead of others'. When librarians select materials for library collections, they often make an effort to represent all sides of issues: such is ingrained in their professional ethic. Internet search engines, by contrast, rank materials solely by popularity, with no effort whatsoever to represent a range of evidence or points of view. Moreover, Internet popularity can be purchased. In education research, what is most popular is that which best serves well-organized and wealthy interests.

The research literature on educational assessment and accountability dates back to the late nineteenth century, after Massachusetts' Horace Mann and his Prussian counterparts, earlier in the century, had initiated the practice of administering large-scale versions of classroom examinations across large groups of schools to compare practices and programs (Phelps 2007b). So-called "scientific" assessments were invented around the turn of the century by several innovators, such as Rice and Binet, and their use was already widespread by the 1920s. The research literature on the effects of these and more traditional assessments had already matured by the 1940s. Some assessment and evaluation topics had been researched so heavily in the early and middle decades of the twentieth century that their researcher counterparts (in psychology) in more recent times have felt little compulsion to "re-create the wheel." If one limits one's search to recent research, one may not find the majority of it, nor the most seminal.

Dismissive Reviews Lead into One-Way Traps*

In scholarly terms, a review of the literature or literature review is a summation of the previous research that has addressed a particular topic. With a dismissive literature review, a researcher assures the public that no one has yet studied a topic or that very little has been done on it. A firstness claim is a particular type of dismissive review in which a researcher insists that he is the first to study a topic. Of course, firstness claims and dismissive reviews can be accurate—for example, with genuinely new scientific discoveries or technical inventions. But that does not explain their prevalence in nonscientific, nontechnical fields, such as education, economics, and public policy.

Dismissive reviewers typically ignore or declare nonexistent research that contradicts their own. Ethical considerations aside, there are several strategic advantages:

- 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 journalists nor cited in footnotes or reference lists.

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.

Dismissive reviewing can be effective and profitable. The more that dismissive reviewers cite each other (and neglect to cite others), the higher they rise in academe's status (and salary) hierarchy. In the scholarly world, acknowledgment is wealth and citations are currency.

By contrast, researchers with contrary evidence whose work is ignored are left in the humiliating position of complaining about being left out. If those responsible for their ostracism can claim higher status—by teaching at more prestigious universities, serving on more prestigious commissions and panels, and receiving larger grants—naïve outsiders will equate the complaints with sour grapes. After all, everything else being equal, an ordinary observer is more likely to trust the research pronouncements of, say, the chemistry professor from Harvard than the chemistry professor from No-name State College. One has faith that the community of chemistry researchers has properly designated its authorities. Is the same faith warranted for professors in U.S. education schools?

*See Phelps 2007a, 2008/2009a, 2012a

I made no extra effort to find older sources in my literature review of several hundred sources on the effect of testing on student achievement. As a result, my search was biased toward more recent work. It is easier to obtain—more likely to be available in electronic form, more likely to be available at no cost, and so on. Still, half of my sources were written prior to 1990.

Of the 900+ references contained in the eight OECD staff and contractor reports I reviewed, only 19 were produced before 1990, and just 112 between 1991 and 2000. More than 800 sources were written after 2000. Why this complete neglect of a century's worth of information in favor of that from just the past decade or so? Does the OECD believe that human nature fundamentally changed around the year 2000? Probably not, but consider this: the World Wide Web came online in the 1990s.

To conduct my literature searches, I spent hundreds of hours inside academic libraries reading microfiche and accessing expensive on-line databases or remote archives. Had I wanted to be more thorough, I would have paid for interlibrary loan access, even international library loan access. As it was, the work was plenty tedious, time-consuming, and expensive. I suspect that OECD researchers eschew doing research that way, and it shows in the myopia of their product.

In fairness to the OECD, one particular assessment method, to my knowledge, was rarely studied prior to the past couple of decades—using student test scores to evaluate teachers. But this was only one of several research literatures REAFISO claims to have mastered. For the others, its claims of thorough coverage are grossly exaggerated.

The OECD on educational testing and accountability

Officially founded in 1948, the Organisation for Economic Co-operation and Development (OECD) is the stepchild of other post–World War II transnational economic unions, such as the International Monetary Fund (IMF) and the World Bank. Formed by eighteen non-Communist European countries, it was originally purposed to manage American and Canadian financial aid for the continent's reconstruction. It has since entrepreneurially shaped itself into a "rich country club" with thirty-four members from all continents (save Africa and Antarctica).

The OECD's growth strategy has by necessity been opportunistic; the IMF, World Bank, and other organizations had already laid claim to the more obvious roles for multinational economic unions.

Relative to other sectors, however, education had been given little attention.[3,4]

For a quarter-century, OECD has published its now-annual *Education-at-a-Glance* collection of "indicators," comparing education systems at the country level on a wide variety of statistics related to enrollment, level of educational attainment, finance, and staffing. For the past ten years, the OECD's Programme for International Student Assessment (PISA) has tested fifteen-year-olds around the world in reading, science, and mathematics and published the results for comparison.

All along, the OECD has also conducted research reviews on various education topics and organized country-level consulting visits. Typically, country-level reviews gather several experts from among OECD staff, the education ministries of member countries, other international organizations, and university faculty to spend a week or two meeting a full range of responsible officials in a single host country. Afterward, a report full of recommendations is written, reviewed, and written again.

Americans can be rather jaded and parochial regarding international organizations. Most countries hosting OECD study teams, however, take them quite seriously. The structure of a country-level review is negotiated between country and OECD and costs are shared. Reviewers are invited in and busy officials are required to give freely of their time and resources to aid the evaluation.

In the OECD's own words,

> The OECD Review on Evaluation and Assessment Frameworks for Improving School Outcomes, launched in late 2009, is designed to respond to the strong interest in evaluation and assessment issues evident at national and international levels. It will provide a description of design, implementation and use of assessment and evaluation procedures in countries; analyse strengths and weaknesses of different approaches; and provide recommendations for improvement.
>
> The Review looks at the various components of assessment and evaluation frameworks that countries use with the objective of improving student outcomes. These include student assessment, teacher appraisal, school evaluation, and system evaluation. The analysis focuses on primary and secondary levels of education.[5]

Housed within the OECD's Directorate for Education's Early Childhood and School Division, REAFISO seems typical of OECD efforts—committing only a tiny (N=5) full-time staff, while expecting to leverage expertise and other resources from member country personnel and institutions, or short-term contract workers, as it goes along. Its web pages exude a sense of temporariness; the review was scheduled for 2009 to 2012, but not necessarily any longer than that.

Note that the OECD rather confusingly tags the word "Review" with multiple meanings:

- a particular office and its employees within the OECD's Directorate for Education;
- a research project on the topic of testing and accountability scheduled for 2009 through 2012; and
- a 1- or 2-week visit to study a host country's education testing and accountability programs by OECD-assembled teams of experts.

By the end of 2012, the efforts of the Review on Evaluation and Assessment Frameworks for Improving School Outcomes (REAFISO) had amassed a large cache of documents. There are several research summaries on student standardized testing, teacher and school evaluation, alignment, and "assessment and innovation" (about 50 pages each); a dozen or so country reviews (at 150 to 200 pages); equally long "background reports" written by host countries prior to OECD study team visits; minutes of OECD-wide meetings of national experts and country representatives; an occasional newsletter; and miscellaneous other missives.

I have read everything I could find on the Web on the background and training of the OECD staff and the contractors they hired to write the REAFISO research summaries. As far as I can tell, none of them has had any training or experience working in assessment and measurement. Rather, they are smart people with general training and experience in research and specific knowledge of other education topics. In such situations, the initial literature reviews can determine the direction of the entire project, either keeping it on course or, as it turned out in REAFISO's case, steering it off its intended course. Though it may be the least-interesting (and typically is the most tedious) task in the research process, the literature search and review can be the most important, particularly in hotly contested, bitterly disputed venues such as education policy.

The Internet Has Changed How Some Conduct Research . . . Unfortunately**

One common approach to research, employed by many journalists and university students, is to instigate a World Wide Web search by typing all the relevant keywords that come to mind into a Web search engine (e.g., Yahoo! Search, Bing, Google). One obtains a long reference list from a few clicks of the mouse. Only fuddy-duddies would complain; it's so much more convenient, it must be better than traditional search methods.

What would the fuddy-duddies say? They might say: most information is not available on the Web and that which is, is ranked by "popularity," which often means that those with money and power can push their point of view ahead of others'.

In the old days of library card catalog searches, references on a topic were delivered unordered, burdening the searcher with the responsibility of considering various points of view. The Web has relieved us of this burden by ranking all references. A researcher working on a deadline, confronted with a few thousand sources from a World Wide Web search on a particular topic, is forced by time limitations to choose only some of them. What is a naïve, unknowledgeable researcher to do when presented with a thousand choices? You and I would probably do the same—pick from among the highest-ranking sources and assume that they represent the whole.

In card-catalog days, the choosing method might have been random sampling among all the seemingly equal possibilities. These days, the choices are made for us; search engines rank sources by popularity, which is often, if inappropriately, equated with importance, prestige, and accuracy. Compare, for example, the outcome of the publications of researchers X and Y. Researcher X is an unaffiliated individual who submits his work to a scholarly journal (or journals), does what he is told to do no matter how arbitrary by the anonymous reviewers, and, if successful, gets his work published in a journal—one among thousands of journals—in due course, that being a year to several years after first submitting his manuscript for review.

Meanwhile, Researcher Y is a brand-new Ph.D. graduate from Harvard University, liked by faculty with Washington, D.C., connections. He is hired to work at the Brookings Institution, the American Enterprise Institute, or some other made-for-D.C. institution with ample funding; you pick. Unlike Researcher X's report, forced to tread the often-arbitrary, often-haphazard gauntlet of peer review, and to rely on the penurious benevolence of a scholarly journal's meager budget for publication, his is reviewed only by those who already like him and are sympathetic to his views. The Public Affairs Office of the Think Tank arranges a conference for "the release" of the publication: ordering *hors d'oeuvres*, inviting the media, and broadcasting press releases to the masses. Researcher Y's publication is professionally edited and formatted and freely available for download off the Internet, whereas Researcher X's journal article is available only with an annual subscription or an exorbitant one-time fee.

If the research conclusions of Researcher X and Researcher Y differed diametrically, and you were forced by time and expense limitations to choose, which would you choose to believe? Consider that Researcher Y's publication will rank very high in search engine rankings and Researcher X's will rank very low.

**See Herring 2001; Phelps 2005a, 2007a, 2008/2009a, 2012a; Stevens-Rayburn 1998.

In his literature review of high-stakes testing, Morten Rosenkvist (2010) asks:

> What is the empirical evidence on the effects of using student test results for accountability and improvement? (p. 5)

He lists about 165 references and he claims (p. 6):

> Empirical evidence has been identified through a broad search in the literature using databases and search engines such as *ScienceDirect, Jstore, Google Scholar, ERIC,* and *SpringerLink*. The search has been conducted in English, the Scandinavian languages, and to a lesser degree German and Spanish.

Sounds thorough, and Rosenkvist's discussion of his search method is the most thorough among the eight REAFISO research summaries. But he did not follow the standard meta-analyst's recipe. We do not know which keywords he used, in which time periods, how he decided which studies were relevant and which were not, how many studies he reviewed, or anything else. He basically says he cast a wide net, and we're supposed to trust that he trawled in a section of the sea representative of the entire research ocean.

Surveying his reference list, I found a dozen sources, among the 165, that pre-date the year 2000, and only one that pre-dates 1990, this for an enormous research literature that dates back to the nineteenth century. Ergo, simply by chronology, hundreds of relevant studies are ignored.

By far the majority of the research on this topic has been conducted by psychologists, the folk who invented cognitive assessment in the first place and, to this day, who continue to produce, manage, administer, and score almost all of them. Yet among his 150-odd sources Rosenkvist includes zero psychology journals. By contrast, 43 references lead to political science and economics journals.

The majority of the REAFISO research summaries' references lead to U.S. sources, and two groups of researchers are cited most frequently—those affiliated with the federally funded Center for Research on Education Standards and Student Testing (CRESST)—a consortium of the education schools at UCLA, U. Colorado, U. Pittsburgh, and the Rand Corporation—and a small group of economists (mostly) affiliated with Republican Party-oriented think tanks, such as the Hoover Institution and the Manhattan Institute. Most of the CRESST researchers have testing and measurement training, but espouse a particular doctrine of assessment policy that, contrary to

their claims, is not shared by most in the profession. This particular group of economists has no practical training or experience in testing and measurement and has, for reasons unknown to me, fallen head-over-heels for CRESST doctrine.

Adding internal references to other OECD documents to those for these two groups, one can account for about half of all the references. At the same time, REAFISO blanks on the majority of the research in the field.[6]

Based on my review of the REAFISO reviews, I conclude the following:

- The OECD conducted arguably the most important aspects of the project—the literature searches and reviews—on the cheap, with smart but young and inexperienced researchers with no particular understanding of the highly technical topic of educational assessment.[7]
- The OECD did not follow standard meta-analysis protocols in structuring its literature searches and reviews. Therefore, it is not possible to know why they chose the research literature they chose and ignored the larger body of relevant research.
- Given that it is impossible to know, at least from reading what they have written, how they conducted their literature searches, it is easy to speculate that their searches were determined by opportunity, convenience, and professional biases (e.g., toward economics, away from psychology).

Six different writers drafted REAFISO's eight research summaries.[8] One possible advantage to doing it this way might have been to diversify information sourcing. With several different individuals, working independently, the probability of a narrow review of the literature should have diminished.

Perusing the reference lists of the various reports, however, one can see that they largely ladled from the same soup pot. The same references and the same research groups appear frequently across different reports. Likewise, major source omissions, obvious to anyone deeply familiar with the research literature, are common across all.

The eight research summaries collectively contain more than 900 references. As is to be expected, many reference each other, other OECD documents, or the documents of closely related institutions, such as the Education Information Network of the European Commission (EURYDICE) and the United Nations Educational,

Scientific, and Cultural Organisation (UNESCO). I count 79 references of this type. For comparison, I count 144 references to CRESST documents and articles by CRESST-affiliated researchers.[9] For contrast, I count no references to the substantial research literature that contradicts CRESST doctrine or disputes its evidence and methods.

By time period, the eight reports' references are: 19 pre-1990; 112 from 1991 to 2000; and 803 post-2000. For those references to journal articles, 206 lead to journals in education, 99 to economics journals, 29 to testing and measurement journals, and a paltry 16 to psychology journals.

Still, not all eight of the OECD staff and contractor reports are the same. They range in quality from sort of OK to just awful. For example, Stephanie Dufaux's review of upper-secondary level (i.e., high school) assessment programs and research suffers from some of the same biases as the others, relying too much on education and economics professors' work conducted only in the past decade, and neglecting older work and that conducted by psychologists (though with five, she includes more psychology references than any of her colleagues). Though still dominated by U.S. sources, she at least makes a concerted effort to widen her search geographically. Beyond the slants, however, she more or less just reports what she finds; she doesn't try to overreach with her conclusions.

"Student Standardised Testing: Current Practices in OECD Countries and a Literature Review," *OECD Education Working Paper No. 65*

Dufaux's colleague Allison Morris is described as a master's graduate (from *Sciences Po*) with a specialty in human security and "research experience in the areas of microfinance, education in emergencies, and economic development" (Morris, p. 3). Her report "aims to synthesise the relevant empirical research on the impact of standardised testing on teaching and learning and to draw out lessons from the literature on aspects of standarised tests that are more effective in improving student outcomes."

That goal well matches that for my meta-analyses and research summary described earlier (Phelps, 2012b). Of the several hundred studies I found to help answer the question, however, Morris includes exactly three in her review. Her report claims that "key debates concerning standarised testing are identified throughout," but only one side in those debates seems to be represented in her review.

Disseminating Misinformation

The REAFISO writers cite as solid and unchallenged the conclusions from several studies that are misrepresentations of the evidence at best and frauds at worst:

Boaler (2002)
Jo Boaler conducts quasi-experimental studies comparing student performance among schools she refuses to identify with data she refuses to release. Despite the fact that students in her constructivist classrooms end up performing worse on all standardized tests administered to them, they perform better on a test she designed herself. So, she reasons, constructivist learning must be superior. SOURCES: Bishop, Clopton, and Milgram 2012; Bishop and Milgram 2012; Milgram 2012

Haney (2000)
Haney's famous study allegedly disproving the advertised success of Texas's "education miracle" in the 1990s contains an extraordinary number of misleading analyses: he sometimes uses different numbers than claimed; surreptitiously alters the definitions of common terms; frequently makes calculation errors; misrepresents data; misrepresents laws, procedures, and events; neglects to consider confounding factors; and sometimes just makes things up. I checked dozens of Haney's "evidence-based" assertions and found none that stood up to any scrutiny. Moreover, every one of his "mistakes" led in the same direction, strongly suggesting willfulness. In its number of factual misrepresentations, Haney's book-length study is the most substantial collection of research fraud I have ever studied. SOURCES: Grissmer, Flanagan, Kawata, and Williamson 2000; Phelps 2003, pp. 127–144, Toenjes, Dworkin, Lorence, and Hill 2000.

Hout and Elliot (2011)
Whereas all but a trivial amount of the great mass of relevant research is ignored, the work of U.S. National Research Council (NRC) study committee members is cited liberally. Daniel Koretz wins the prize for the most citations and references. Overall, forty-eight citations and forty references (of two hundred) go to committee members' work. More than thirty 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.

Also, 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. SOURCE: Phelps 2012c.

Klein, Hamilton, McCaffrey, and Stecher (2000)
In the "October Surprise" of the 2000 U.S. presidential campaign, these CRESST researchers debunk Texas' gains on the U.S. National Assessment of Educational Progress by disaggregating the data to the lowest level possible, then running a separate hypothesis test on each disaggregation. Their method is obviously invalid, as their conclusion is about Texas' gain scores on all the NAEP segments together—the "pooled" data. Overall, Texas' gains on the NAEP exceeded those of all other U.S. states but one in the 1990s. SOURCE: Phelps 2003, pp. 122–127

Koretz (2005a, 2005b, 2008)
J. J. Cannel's "Lake Wobegon Effect" studies showcased the causes of test-score inflation—educator dishonesty and conflicts of interest, lax security, and outdated norms. Koretz identified high stakes as the main culprit, even though all but one of Cannell's score-inflated tests were national norm-referenced monitoring (i.e., no-stakes) tests. Koretz cites a study he and CRESST colleagues conducted around 1990 in an unidentified school district, with unidentified tests, as evidence that high stakes cause test score inflation. But he controlled for none of the other factors—such as lax security—that could have explained the results. Nor, apparently, did the test genuinely have high-stakes. SOURCES: Fraker 1986/1987, Phelps 2008/2009b, 2010; Staradamskis 2008.

Linn (1998, 2000)
Linn further argued that the pre-post testing requirement (or, Title I Evaluation and Reporting System [TIERS]) of the Title I Compensatory Education [i.e., anti-poverty]) program from the late 1970s on offered more evidence of the high-stakes-cause-test-score-inflation theory. His study had no controls, however, and the test involved did not carry any stakes—it was merely a reporting requirement (with no actual consequences). Linn argued that tests administered on a fall-spring schedule (presumably by the same teacher) averaged higher gain scores than those administered on a fall-fall schedule (presumably by different teachers). The average summer learning loss, as found in meta-analyses on the topic, however, entirely explains the difference in average test scores gains. SOURCES: Sinclair and Gutman 1992; Cooper, Nye, Charlton, Lindsay, and Greathouse, 1996; Phelps 2008/2009b, 2010.

Shepard (1989)
Shepard published a table that lists, allegedly, all the possible causes of the Lake Wobegon Effect that research to that date had suggested. Conspicuously absent from the table were Cannell's chief culprits: lax security and educator dishonesty and conflicts of interest. With these causal factors eliminated from consideration, she was free to attribute causation to high-stakes. SOURCES: Phelps 2003 (chapter 4), 2008/2009a,b, 2010, 2011a,b

Morris lists fifty-nine references, but visits and revisits certain individual sources repeatedly in the text. She cites five CRESST researchers ninety-one times.[10] She cites a report from the CRESST satellite National Research Council Board on Testing and Assessment nine times.[11] Citations for the cuckolded group of economists allied with CRESST exceed fifty. One must squint to see how Morris synthesizes the relevant empirical research and identifies key debates when she cites a single, sparsely informative book chapter by Figlio and Loeb (2011) *thirty-six times*.

Among the more egregious of Morris's erroneous statements:

> While being highly reliable and comparable, multiple choice questions can be limiting in that they do not test critical thinking or problem solving skills and it is argued such

> questions encourage surface learning and rote recollection, rather than deep, cognitive processes. Rather than testing thinking skills, multiple choice or other close-ended questions test content only (p. 16). . . . [D]ue to the nature of a standardised test, the tests often cannot test for critical thinking, analytical or problem solving skills (p. 21).

This is dead wrong if a multitude of better-supported studies are to believed (see, for example, Bridgeman 1991; Feinberg 1990; Rudman 1992; Traub 1993; Powers and Kaufman 2002; Goodman and Hambleton 2005; Roediger and Marsh 2005; Struyven, Dochy, Janssens, Schelfhout, and Gielen 2006), and had the author taken the time to peruse some of the many freely available retired tests online and read their items, she could have seen so for herself.

> According to the literature, validity of large-scale, standardised tests—specifically those used to assess program effectiveness—is increased through matrix sampling. (p. 23)

"The literature" turns out to be a single source written by CRESST authors. The obvious question is: valid to whom? With most matrix-sample tests, no results are valid at the student, teacher, classroom, or school level, and so are responsibly not reported at those levels.

> [P]lacing a "premium" on student test performance in the form of rewards or sanctions for teachers increases the risk of instruction being reduced to test preparation, which in turn limits the depth of the student experience and reduces the skill needed by teachers. Additionally, incentives such as bonuses can lead to strategic actions by teachers that distort or manipulate data. These include cases of teacher cheating, exclusion of students in assessments, and teaching to the test, all of which are reviewed in greater detail below. (p. 29)

Had REAFISO widened its literature search just a little, it might have learned: when teachers teach standards-based subject matter they are properly teaching to the test (as it is aligned with the standards); when they spend more than a smidgen of time drilling on test format they hurt, not help, their students' scores on the upcoming test; when they see in advance the specific content of an upcoming test, the problem is lax test security, not improper incentives. By the way, test developers know that drilling on test format does not work and discourage it (see, for example, Messick and Jungeblut 1981; DerSimonian and Laird 1983; Kulik, Bangert-Drowns, and Kulik 1984; Fraker 1986/1987; Whitla 1988; Snedecor 1989; Smyth 1990;

Becker 1990; Moore 1991; Powers 1993; Tuckman 1994; Tuckman and Trimble 1997; Powers and Rock 1999; Robb and Ercanbrack 1999; Camara 1999, 2008; Briggs 2001; Palmer 2002; Briggs and Hansen 2004; Crocker 2005; Roediger and Karpicke 2006a, 2006b; Allensworth, Correa, and Ponisciak 2008).

It is the researchers REAFISO has chosen to trust who broadcast the erroneous and destructive suggestion that it works.

> Research from the United States has shown that if national tests are considered to be 'high stakes' for teachers and schools, teaching to the test can easily lead to an artificial over-inflation of results and thus render the results useless as a measure of real progress. (p. 37)

If CRESST researchers were correct that high-stakes caused test-score inflation, we should expect to find test-score inflation with all high-stakes tests, such as the hundreds of occupational licensure tests and U.S. college admission tests (e.g., SAT, ACT), but we do not. We do not because these tests are administered with high levels of security and frequent form and item rotation. The source of test-score inflation is lax test security, not high-stakes. (See, for example, Phelps 2010; Staradamskis 2008.)

"Using Student Test Results for Accountability and Improvement: A Literature Review," *OECD Education Working Paper No. 54*

For his research review, Morten Anstorp Rosenkvist, the Norwegian civil servant on loan to the OECD for a few months in early 2010, read about student testing without stakes for students but sometimes stakes for teachers or schools. More so than the other REAFISO writers, Rosenkvist read surveys and opinion poll reports to better gauge the attitudes and preferences of non-researchers toward testing. This interested me personally since I have been studying the same for a couple decades, ultimately publishing a meta-analysis of 813 individual item-response group combinations from 247 program evaluation surveys and opinion polls conducted between 1960 and 2010 (Phelps 2012b).

Alas, 'tis a pity, Rosenkvist did not happen upon my work. He mentions the results of three surveys of local public officials (p. 16), three of school administrators (p. 16), six of teachers (p. 18), seven of parents (pp. 19–20), and three of students (p. 20). According to

Rosenkvist (p. 20), "Students generally dislike high stakes assessments." But my meta-analysis counted twenty student surveys regarding high-stakes tests that accumulated an average, rather large effect size of +1.03. For the other response groups, Rosenkvist concludes generally positive support for high-stakes testing, roughly matching the results from my meta-analysis. Whereas he bases his conclusions on three, three, six, and seven cases, however, mine emerge from seven, thirty-four, eighty-five, and fifty-five, as well as several from university faculty and hundreds from the general public, and average effect sizes can be calculated precisely for each group. All this information was available to Rosenkvist had he asked for it (Phelps 2005b, 2012b).

I felt similar frustration reading Rosenkvist fumbling around with summarizing the research on the effect of high-stakes tests on student achievement (pp. 22–24) and the cost of assessment (p. 27). He would have encountered far more evidence, and reached more reliable conclusions, had he been willing to search outside the CRESST-U.S. think tank bubble.

A wider search might have smoothed out the inevitable contradictions, too. At several points, Rosenkvist encourages readers always to use a variety of measures and multiple tests, because no one test can be perfect or cover a domain of interest sufficiently. Yet he also recommends Daniel Koretz's method of judging the validity of one test score trend by comparing it with that of another, even with completely different topical coverage. In his final concluding paragraph (p. 35), Rosenkvist asserts, reasonably, that "student test results must be reliable, valid and fair." Then, in the next sentence, he recommends that "several assessments should be used to measure [each] student outcome." Good luck with that.

"Assessment and Innovation in Education," *OECD Education Working Paper No. 24*

The worst REAFISO research in most respects, however, is that conducted by the one American, Janet W. Looney (2009, 2011, 2013), the most prone to overreach. Unfortunately, whereas her colleagues wrote one report each, Looney wrote three. Looney is such an ardent, devout missionary for CRESST doctrine, she could be mistaken as their publicist.[12] Using the same comparison metric as before, Looney's references include twenty-two to OECD and related sources, and more than three times as many—a whopping seventy—references to

CRESST and CREST-affiliate sources (and zero references to sources that dispute CRESST's evidence or methods).

> While only a small number of studies have focused on the validity of test score gains in high-stakes testing, they have usually found evidence of score inflation. (p. 30)

This statement is wrong on both counts. There exist a large number of studies, and they did not find score inflation where test security was tight and form and item rotation frequent. (See, for example, Becker 1990; Moore 1991; Tuckman 1994; Tuckman and Trimble 1997; Powers and Rock 1999; Briggs 2001; Palmer 2002; Crocker 2005; Roediger and Karpicke 2006a, 2006b; Camara 2008.) The studies conducted by Koretz and Linn that, allegedly, found evidence of score inflation involved low-stakes tests (erroneously labeled as high-stakes). Neither study employed any controls. Linn's result is easily explained away by the summer-learning-loss effect. Koretz's study remains shrouded in secrecy two decades later, with a still-unidentified school district with still-unidentified tests. (See, for example, Phelps 2010; Staradamskis 2008.)

As mentioned earlier, Looney aggressively promotes innovation—specifically the type espoused by U.S. radical constructivists and CRESST researchers. It seems not to matter that the public has expressed little interest, or that the programs have failed miserably when tried. She writes (p. 17):

> Most new programmes experience an "implementation dip"—that is, student performance gets worse before improving. Improvements in student achievement may take as long as five years in primary schools, and longer in secondary schools. Teachers working in innovative programmes will need extra support to understand where they may need to adjust practices.

Looney advocates sticking with programs she favors even if the evidence of their effectiveness is negative for five years (with primary graders) or even longer (with older students). But experience shows that most innovations fail; certainly they do not succeed simply for the fact that they are innovative. We should rest steadfast through five, six, seven, or more years of negative results before even considering pulling the plug on a program? And what of the children whose education has been stunted in the meantime? Do they matter?[13]

Continuing, Looney writes:

> Teachers may find it impossible to balance the pressures of implementing new and innovative programs and high-stakes tests." (p. 18) Moreover, tight alignment . . . tends to undermine innovative programmes. . . . (p. 20)

The implication is, of course, that it is the standards and tests that must be at fault. They should either be pushed aside to make way for the innovations, or be radically reconstructed to fit the new order.

Summary

My judgment of REAFISO's efforts should be apparent at this point. But REAFISO's efforts should be judged unfavorably even by its own standards. In the *Design and Implementation Plan for the Review* (OECD, 2009), REAFISO promised to, among other goals:

> . . . extend and add value to the existing body of international work on evaluation and assessment policies. (p. 5)

> Synthesise research-based evidence on the impact of evaluation and assessment strategies and disseminate this knowledge among countries. Identify policy options for policy makers to consider. (p. 4)

> . . . take stock of the existing knowledge base within the OECD and member countries as well as academic research on the relationship between assessment and evaluation procedures and performance of students, teachers and schools. It will look at the quantitative and qualitative evidence available on the different approaches used to evaluate and assess educational practice and performance. (p. 16)

To the contrary, REAFISO has not synthesized the existing body of research-based evidence on evaluation and assessment policies, much less extended it. By telling the world that a small proportion of the existing body of research is all that exists, they have instead hidden from the world most of the useful and relevant information (or implied that it is not worth considering).

The ordinary Citizen Joe knows that one shouldn't trust everything one finds on the Internet, nor assume that Internet search engines rank documents according to their accuracy. So naturally, scholarly researchers who are trained to be skeptical, systematic, thorough, aware of biases, and facile with statistical sampling methods would be too. After all, scholarly researchers have spent several

more years in school, often prestigious schools. They should "know how to know" as well or better than the average citizen.

Yet REAFISO's reviews repeatedly offer one or a few examples of research from their favored sources to summarize topics, even though thorough reviews of dozens, hundreds, or thousands of sources were to be found had they simply looked widely enough. In some cases, REAFISO writers conclude a policy recommendation on the basis of one or a few studies, when a reading of the whole of the research literature on the topic would suggest exactly the opposite policy.

In its document, *Evaluation and Assessment Frameworks for Improving School Outcomes: Common Policy Challenges* (2011), written two years after the *Design and Implementation Plan*, REAFISO claims to have completed "a thorough analysis of the evidence on evaluation and assessment."

Ironies

Ideas matter, as do their censorship and suppression. For all their hawking of recipes from the radical constructivist cookbook—promoting "critical" and *"higher-order"* thinking, discovery learning, and innovation for its own sake —the REAFISO writers neither construct their own knowledge, discover knowledge with unique learning styles, nor evaluate what they read critically. The six writers read and came to believe the same, unquestioningly parroting a single dogma.

For all of REAFISO's adulation of innovation, there's little to be found in their reviews. Apparently, they started with the most accessible and heavily promoted reports from the U.S., CRESST, and CRESST's cuckolded cabal of economists, and followed their well-worn path like a workhorse with side blinders. REAFISO's writers looked where they were told to look and conspicuously avoided looking in any of the directions they were not told to.

REAFISO criticizes traditional tests for "narrowing the curriculum" and "teaching to the test." But REAFISO itself narrowed its focus in the relevant research literature to a tiny aperture, reducing its search to that seen within the perimeter. Then, over several hundred pages, REAFISO repeatedly, relentlessly drills its confirmation bias into its readers.

As a result, the OECD now recommends to all its members the wisdom of U.S. education research, certainly the world's least effective and perhaps the world's most corrupt—responsible for producing

one of the world's least successful education systems (as measured by outcomes over inputs). U.S. students continue to underperform on the OECD's own PISA assessment, despite U.S. taxpayers spending more on education per capita than all but a few other countries. The U.S. public and politicians see their education system in a perpetual state of crisis, as having largely failed. The OECD now suggests the rest of the world copy it.

In 2011, REAFISO wrote:

> The effectiveness of evaluation and assessment relies to a great extent on ensuring that both those who design and undertake evaluation activities as well as those who use their results possess the proper skills and competencies. This is crucial to provide the necessary legitimacy to those responsible for evaluation and assessment.

If only they had practiced what they preach.

References

Allensworth E., Correa, M., Ponisciak, S. (2008, May). *From High School to the Future: ACT Preparation–Too Much, Too Late: Why ACT Scores Are Low in Chicago and What It Means for Schools.* Chicago: Consortium on Chicago School Research at the University of Chicago.

Becker, B. J. (1990, Fall). Coaching for the Scholastic Aptitude Test: Further synthesis and appraisal. *Review of Educational Research, 60*(3), 373–417.

Bishop, W., Clopton, P., and Milgram, R. J. (2012). A Close Examination of Jo Boaler's Railside Report. *Nonpartisan Education Review / Articles, 8*(1). http://www.nonpartisaneducation.org/Review/Articles/v8n1.pdf

Bishop, W., and Milgram, R. J. (2012). A Response to Some of the Points of: "When Academic Disagreement Becomes Harassment and Persecution". *Nonpartisan Education Review / Essays, 8*(4). http://www.nonpartisaneducation.org/Review/Essays/v8n4.htm

Bridgeman, B. (1991, June). Essays and multiple-choice tests as predictors of college freshman GPA. *Research in Higher Education,* 32(2), 319–332.

Briggs, D. C. (2001, Winter). The effect of admissions test preparation. *Chance.*

Briggs, D., and Hansen, B. (2004, May). Evaluating SAT test preparation: Gains, effects, and self-selection. Paper presented at the Educational Testing Service, Princeton, N.J.

Boaler, J. (2002). *Experiencing School Mathematics: Traditional and Reform Approaches to Teaching and their Impact on Student Learning*, Lawrence Erlbaum Associates, Mahwah, N.J.

Camara, W. (1999, Fall). Is commercial coaching for the SAT I worth the money? *College Counseling Connections 1*(1). New York: The College Board.

————. (2008). College admission testing: Myths and realities. In R. P. Phelps (Ed.), *Correcting fallacies about educational and psychological testing*, Washington, D.C.: American Psychological Association.Cooper, H., Nye, B., Charlton, K., Lindsay, J., and Greathouse, S. (1996, Fall). The effects of summer vacation on achievement test scores: A narrative and meta-analytic review. *Review of Educational Research. 66*(3).

Crocker, L. (2005). Teaching FOR the test: How and why test preparation is appropriate. In R. P. Phelps (Ed.), *Defending Standardized Testing* (pp. 159–174). Mahwah, N.J.: Lawrence Erlbaum.

DerSimonian and Laird. (1983). Evaluating the effect of coaching on SAT scores: A meta-analysis, *Harvard Educational Review* 53, 1-5.

Dufaux, S. (2012), "Assessment for Qualification and Certification in Upper Secondary Education: A Review of Country Practices and Research Evidence," *OECD Education Working Papers,* No. 83, OECD Publishing: Paris. http://dx.doi.org/10.1787/5k92zp1cshvb-en

Faubert, V. (2009). "School Evaluation: Current Practices in OECD Countries and a Literature Review". *OECD Education Working Papers,* No. 42, OECD Publishing: Paris. http://www.oecd.org/officialdocuments/displaydocument/?doclanguage=en&cote=edu/wkp(2009)21

Feinberg, L. (1990, Fall). Multiple-choice and its critics: Are the 'alternatives' any better? *The College Board Review*, 157, 13–17, 30–31.

Figlio, D., and Loeb, S. (2011). "School Accountability," in E. Hanushek, S. Machin and L. Woessman (eds.), *Handbooks in Economics, Vol. 3*, North-Holland, The Netherlands, pp. 383–421.

Fraker, G. A. (1986–1987, Winter). The Princeton Review reviewed. *The Newsletter*. Deerfield, Mass.: Deerfield Academy, Winter.

Goodman, D., and Hambleton, R. K. (2005). Some misconceptions about large-scale educational assessments. In R. P. Phelps, (Ed.), *Defending standardized testing* (pp. 91–110). Mahwah, N.J.: Lawrence Erlbaum.

Grissmer, D. W., Flanagan, A., Kawata, J., Williamson, S. (2000, July). *Improving Student Achievement: What NAEP State Test Scores Tell Us*, Rand Corporation. Haney, W. (2000). The Myth of the Texas Miracle in Education, *Education Policy Analysis Archives, 8*(41). http://epaa.asu.edu/ojs/article/view/432

Hout, M., and Elliott, S. (eds.) (2011). *Incentives and Test-Based Accountability in Education,* National Research Council, The National Academies Press, Washington, D.C. http://www.nap.edu/catalog.php?record id=12521

Herring, M. Y. (2001, April). Ten reasons why the Internet is no substitute for a library. *American Libraries*.

Isoré, M. (2009). "Teacher Evaluation: Current Practices in OECD Countries and a Literature Review". *OECD Education Working Papers,* No. 23, OECD Publishing: Paris. http://www.oecd.org/officialdocuments/displaydocument/?doclanguage=en&cote=edu/wkp(2009)2

Klein, S., Hamilton, L., McCaffrey, D., and Stecher, B. (2000). *What do test scores in Texas tell us?* Issue paper, Rand Education.

Koretz, D. (2005a). *Alignment, High Stakes, and the Inflation of Test Scores*, University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST), Los Angeles.

————. (200b5). Alignment, High Stakes, and the Inflation of Test Scores, in J. Herman and E. Haertel (eds.), Uses and Misuses of Data in Accountability Testing, *Yearbook of the National Society for the Study of Education, 104, Part 2*, Malden, MA: Blackwell Publishing.

————. (2008). *Measuring up: What Educational Testing Really Tells Us*, Cambridge, Mass.: Harvard University Press.

Kulik, J. A., Bangert-Drowns, R. L., and Kulik, C-L. C. (1984). Effectiveness of coaching for aptitude tests, *Psychological Bulletin* 95, 179-188.

Linn, R. (1998). *Assessments and Accountability,* University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST), Los Angeles.

Linn, R. L. (2000). Assessments and Accountability, *Educational Researcher,* 29, pp. 4– 16.

Looney, J. W. (2009). "Assessment and Innovation in Education." *OECD Education Working Papers,* No. 24, OECD Publishing: Paris. http://www.oecd.org/officialdocuments/displaydocumentpdf/?cote=edu/wkp(2011)9&doclanguage=en

————. (2011). "Integrating Formative and Summative Assessment: Progress Toward a Seamless System?" *OECD Education Working Papers,* No. 58, OECD Publishing: Paris. http://www.oecd.org/officialdocuments/displaydocumentpdf/?cote=edu/wkp(2011)4&doclanguage=en

————. (2013). "Alignment in Complex Education Systems: Achieving Balance and Coherence". *OECD Education Working Papers,* No. 64, OECD Publishing: Paris. http://www.oecd.org/officialdocuments/displaydocumentpdf/?cote=edu/wkp(2011)9&doclanguage=en

Messick, S., and A. Jungeblut (1981). Time and method in coaching for the SAT, *Psychological Bulletin* 89, 191–216.

Milgram, R. J. (2012). Private Data—The Real Story: A Huge Problem with Education Research, *Nonpartisan Education Review / Essays, 8*(5). http://www.nonpartisaneducation.org/Review/Essays/v8n5.htm

Moore, W. P. (1991). Relationships among teacher test performance pressures, perceived testing benefits, test preparation strategies, and student test performance. Ph.D. dissertation, University of Kansas, Lawrence.

Morris, A. (2011). "Student Standardised Testing: Current Practices in OECD Countries and a Literature Review." *OECD Education Working Papers,* No. 65, OECD Publishing: Paris. http://www.oecd.org/officialdocuments/displaydocument/?cote=EDU/WKP(2011)10&doclanguage=en

OECD Review on Evaluation and Assessment Frameworks for Improving School Outcomes. (2009, October). *Design and Implementation Plan for the Review.* Education Policy Committee, OECD Directorate for Education, Organisation for Cooperation and Development, Paris. http://www.oecd.org/edu/preschoolandschool/44568070.pdf

OECD Review on Evaluation and Assessment Frameworks for Improving School Outcomes. (2011). *Common Policy Challenges.* Education Policy Committee, Directorate for Education, Organisation for Co-operation and Development, Paris. http://www.oecd.org/edu/preschoolandschool/46927511.pdf

Palmer, J. S. (2002). Performance incentives, teachers, and students: Ph.D. dissertation. Columbus: The Ohio State University.

Phelps, R. P. (2003). *Kill the messenger*. New Brunswick, N.J.: Transaction Publishers.

————. (2005a, February). Educational testing policy: Stuck between two political parties, *Yale Politic*. http://www.nonpartisaneducation.org/Foundation/YalePoliticArticle.htm

————. (2005b). Persistently positive. In R. P. Phelps (Ed.), *Defending standardized tes*ting (pp. 1–22). Mahwah, N.J.: Lawrence Erlbaum.

————. (2007a, Summer). The dissolution of education knowledge. *Educational Horizons*. 85(4), 232–247. http://www.nonpartisaneducation.org/Foundation/DissolutionOfKnowledge.pdf

————. (2007b). *Standardized testing primer*. New York: Peter Lang.

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

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

———. (2010, July). The source of Lake Wobegon [updated]. *Nonpartisan Education Review / Articles, 6*(3). Available at: http://www.nonpartisaneducation.org/Review/Articles/v6n3.htm http://www.nonpartisaneducation.org/Review/Articles/v6n3.htm http://www.nonpartisaneducation.org/Review/Art icles/v6n3.htm

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

———. (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

———. (2012a). Dismissive reviews: Academe's Memory Hole. *Academic Questions*, Summer. http://www.nas.org/articles/dismissive reviews academes memory hole

———. (2012b). The effect of testing on student achievement, 1910–2010, *International Journal of Testing, 12*(1), 21-43, International Test Commission. http://www.tandfonline.com/doi/abs/10.1080/15305058.2011.602920#preview

———. (2012c, Summer). The Rot Festers: Another National Research Council Report on Testing, *New Educational Foundations, 1*(1), pp. 30–52. http://www.newfoundations.com/NEFpubs/NewEduFdnsv1n1Announce.html

Powers, D. E. (1993). Coaching for the SAT: A summary of the summaries and an update. *Educational Measurement: Issues and Practice.* 39, 24–30.

Powers, D. E., and Rock, D. A. (1999). Effects of coaching on SAT I: Reasoning test scores. *Journal of Educational Measurement,* 36(2), 93–118.

Powers, D. E., and Kaufman, J. C. (2002). Do standardized multiple-choice tests penalize deep-thinking or creative students? *Research Report RR-02-15*. Princeton, N.J.: ETS.

Robb, T.N., and Ercanbrack, J. (1999, January). A study of the effect of direct test preparation on the TOEIC scores of Japanese university students. *Teaching English as a Second or Foreign Language. 3*(4).

Roediger, H. L., and Marsh, E. J. (2005). The positive and negative consequences of multiple-choice testing. *Journal of Experimental Psychology: Learning, Memory and Cognition,* 31, 1155–1159.

Roediger, H. L. and Karpicke, J. D. (2006a). The power of testing memory: Basic research and implications for educational practice. *Perspectives on Psychological Science,* 1, 181–210.

————. (2006b). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science,* 17, 249–255.

Rosenkvist, M. A. (2010). "Using Student Test Results for Accountability and Improvement: A Literature Review". *OECD Education Working Papers,* No. 54, OECD Publishing: Paris. http://www.oecd.org/officialdocuments/displaydocument/?cote=EDU/WKP(2010)17&doclanguage=en

Rudman, H. C. (1992, Fall). *Testing for learning* (book review). *Educational Measurement: Issues and Practice*, 31–32.

Shepard, L. A. (1989, March). Inflated test score gains: Is it old norms or teaching the test? Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Sinclair, B. and Gutman, B. (1992). *A Summary of State Chapter 1 Participation and Achievement Information for 1989–90*. Prepared for the U.S. Department of Education, Office of Policy and Planning, 1992.

Smyth, F. L. (1990). SAT coaching: What really happens to scores and how we are led to expect more. *The Journal of College Admissions,* 129, 7–16.

Snedecor, P. J. (1989). Coaching: Does it pay—revisited. *The Journal of College Admissions.* 125, 15–18.

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

Stevens-Rayburn, S. (1998). If it's not on the Web, it doesn't exist at all: Electronic information resources¯myth and reality. *Library and Information Services in Astronomy III, ASP Conference Series V,* 153.

Struyven, K., Dochy, F., Janssens, S., Schelfhout, W., and Gielen, S. (2006). The overall effects of end-of-course assessment on student performance: A comparison between multiple choice testing, peer assessment, case-based assessment and portfolio assessment. *Studies in Educational Evaluation,* 32, 202–222.

Toenjes, L. A., Dworkin, A. G., Lorence, J., and Hill, A. N. (2000, August). *The Lone Star Gamble: High Stakes Testing, Accountability, and Student Achievement in Texas and Houston*. Department of Sociology, University of Houston.

Traub, R. E. (1993). On the equivalence of the traits assessed by multiple choice and constructed-response tests. In R. E. Bennett and W. C. Ward (Eds.), *Construction versus choice in cognitive measurement*. Hillsdale, N.J.: Lawrence Erlbaum.

Tuckman, B. W. (1994, April). Comparing incentive motivation to metacognitive strategy in its effect on achievement, Paper presented at the annual meeting of the American Educational Research Association, New Orleans (ERIC ED368790).

Tuckman, B. W. and Trimble, S. (1997, August). Using tests as a performance incentive to motivate eighth-graders to study, Paper presented at the annual meeting of the American Psychological Association, Chicago. (ERIC: ED418785).

Whitla, D. K. (1988). Coaching: Does it pay? Not for Harvard students. *The College Board Review.* 148, 32–35.

Notes

1. The OECD's English-language publications use British, rather than American, spellings.
2. The OECD is on the Web at www.oecd.org www.oecd.org www.oecd.org and REAFISO, http://www.oecd.org/edu/preschoolandschool/oecdreviewonevaluationandassessmentframeworksforimprovingschooloutcomes.htm.
3. By the turn of the century, the World Bank had conducted seemingly thousands of "returns to education" studies that calculated the long-term economic effect of varying amounts of time spent in school. As we all know, however, time spent in school can vary quite a lot in efficacy. To know what works inside the schoolhouse, more information was needed.
4. The United Nations Educational, Scientific, and Cultural Organization (UNESCO) enables cooperation among member countries' education ministries. Economists, and economic analysis, have never occupied more than a small proportion of its activities.
5. See project home page home page home page home page.
6. Names missing from all the REAFISO reports include: Roddy Roediger, 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, Jeffrey Karpicke, Michael Beck, Stephen Heynemann, William D. Schafer, Francine Hultgren, Willis Hawley, James H. McMillan, Elizabeth Marsh, Susan Brookhart, Gene Bottoms, Gordon Cawelti, 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, Arthur Hughes, D. Pennycuick, John Poggio, Anthony Somerset, John O. Anderson, Noel McGinn, Anne Anastasi, Nick Theobald, David Miller, Nancy Protheroe, Floraline Stevens, Ted Britton,

Senta Raizen, Thomas Corcoran, Clement Stone, Frank Dempster, and state agencies in Massachusetts, Florida, and South Carolina.

Those are just names of some folks 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 and Mann (1973); Jones (1974); Bjork (1975); Peckham and Roe (1977); Wildemuth (1977); Jackson and Battiste (1978); Kulik, Kulik, Bangert-Drowns, and Schwalb (1983–1991); Natriello and Dornbusch (1984); Dawson and Dawson (1985); Levine (1985); Resnick and Resnick (1985); Guskey and Gates (1986); Hembree (1987); Bangert-Drowns, Kulik, and Kulik (1991); Dempster (1991); Adams and Chapman (2002); Locke and Latham (2002); Roediger and Karpicke (2006); and Basol and Johanson (2009).

7. Janet W. Looney, responsible for three reports, has a master's degree in public administration (U. Washington) and apparently once worked at the OECD. She now works as a freelance writer and editor. Morten Anstorp Rosenkvist worked for a few months at the OECD on *secondment* (essentially, temporary loan) from the Norwegian Ministry of Education and Research. Aside from his OECD report, he has written "Mobility of Teachers across Education Sectors in Norway," but I could find nothing else about him on the Web or in the Norwegian Ministry of Education website. Allison Morris, Violaine Faubert, and Marlène Isoré are identified by the OECD as graduate students at the *Institut d'Etudes Politiques de Paris* (*Sciences Po*). Two of the three have apparently written economics papers on financial accounting and international trade, but I could find nothing else about them through Web searches. Stephanie Dufaux works at the OECD as a "Carlo Schmid Fellow," and that's all I could find on her.

8. Dufaux, S. (2012), "Assessment for Qualification and Certification in Upper Secondary Education: A Review of Country Practices and Research Evidence," *OECD Education Working Papers,* No. 83, OECD Publishing: Paris. http://dx.doi.org/10.1787/5k92zp1cshvb-en

   Faubert, V. (2009). "School Evaluation: Current Practices in OECD Countries and a Literature Review." *OECD Education Working Papers*, No. 42, OECD Publishing: Paris. http://www.oecd.org/officialdocuments/displaydocument/?doclanguage=en&cote=edu/wkp(2009)21

   Isoré, M. (2009). "Teacher Evaluation: Current Practices in OECD Countries and a Literature Review". *OECD Education Working Papers*, No. 23, OECD Publishing: Paris. http://www.oecd.org/officialdocuments/displaydocument/?doclanguage=en&cote=edu/wkp(2009)2

   Looney, J. W. (2009). "Assessment and Innovation in Education". *OECD Education Working Papers,* No. 24, OECD Publishing: Paris.

http://www.oecd.org/officialdocuments/displaydocumentpdf/?cote=edu/wkp(2011)9&doclanguage=en

Looney, J. W. (2011). "Integrating Formative and Summative Assessment: Progress Toward a Seamless System?" *OECD Education Working Papers,* No. 58, OECD Publishing: Paris. http://www.oecd.org/officialdocuments/displaydocumentpdf/? cote=edu/wkp(2011)4&doclanguage=en

Looney, J. W. (2013). "Alignment in Complex Education Systems: Achieving Balance and Coherence." *OECD Education Working Papers,* No. 64, OECD Publishing: Paris. http://www.oecd.org/officialdocuments/displaydocumentpdf/? cote=edu/wkp(2011)9&doclanguage=en

Morris, A. (2011). "Student Standardised Testing: Current Practices in OECD Countries and a Literature Review". *OECD Education Working Papers,* No. 65, OECD Publishing: Paris. http://www.oecd.org/officialdocuments/displaydocument/? cote=EDU/WKP(2011)10&doclanguage=en

Rosenkvist, M. A. (2010). "Using Student Test Results for Accountability and Improvement: A Literature Review." *OECD Education Working Papers,* No. 54, OECD Publishing: Paris. http://www.oecd.org/officialdocuments/displaydocument/?cote=EDU/WKP(2010)17&doclanguage=en

9. Included in the CRESST counts are those for the US National Research Council's Board on Testing and Assessment which has, essentially, been run as a CRESST satellite since the late 1980s (see Phelps, 2012c).
10. CRESST researcher (number of citations): Laura Hamilton (32); Brian Stecher (17); Stephen Klein (17); Daniel Koretz (15); Robert Linn (7).
11. Hout and Elliot, 2011.
12. The Center for Research in Education Standards and Student Testing, a federally funded consortium including the Rand Corporation, and the Schools of Education at UCLA, U. Colorado, and U. Pittsburgh (and formerly Arizona State U., as well).
13. Looney is a graduate of a master's in public administration program. Presumably, she studied project planning, discounting, returns on investment, and the like. She should know that programs guaranteed to have negative returns for five or more years seldom pay off. Moreover, she says nothing about the size of the assumed future payoff of these innovative programs.

Richard P. Phelps is a member of the *New Educational Foundations* editorial board, the founder of the *Nonpartisan Education Review* (nonpartisaneducation.org) and co-author and editor of *Correcting Fallacies about Educational and Psychological Testing* (APA Books, 2008/9).