

Down the Memory Hole: Evidence on Educational Testing

Richard P. Phelps



© Springer Science+Business Media, LLC, part of Springer Nature 2020

What happens to the research evidence in a scientific field when the professionals in that field do not like it?

Some naively believe, as I once did, that all scientific research is somehow accumulated and preserved. Some of it is, even if its preservation may be obscure. Many scholarly journal indexes, for example, date back to the early twentieth century, and their earliest journal contents can still be found in some dusty academic libraries or on microfiche. Other scientific research is not deliberately preserved, or even indexed, and can more easily be forsaken and forgotten.¹

Research on educational testing, its uses and effects, should greatly interest the American public. A standardized test, when administered by objective third parties, is one of the few instruments available to measure what happens inside our schools, which is not controlled by those who run our schools. For several decades, most U.S. states have incorporated systemwide testing in their

¹In the field of educational testing, for example, standardized test use proliferated in schools throughout the country starting in the 1920s, often with the assistance of a school district or state level “research bureau.” Often, the research professionals evaluated the uses and effects of those testing programs. But, far more often than not, those evaluations ended up on shelves in bureau offices, scattered across the country, and may or may not still exist.

Many school districts and state education agencies managed systemwide standardized testing programs in the latter half of the twentieth century, and in most cases those programs were evaluated, as was typically required by law. Where can one find those evaluation reports now? They may or may not rest on shelves in district or state archives.

Meta-analysts—scholars who specialize in summarizing research literatures—classify such reports as “grey literature,” along with masters and doctoral theses, the proprietary reports of commercial test developers and foundations, and other odds and ends research documents not found in journal indexes or internet searches.

Richard P. Phelps is founder of the Nonpartisan Education Review (<https://nonpartisaneducation.org>); mail@nonpartisaneducation.org. His most recent books are *Defending Standardized Testing* (2005), *Standardized Testing Primer* (2007), and *Correcting Fallacies about Educational and Psychological Testing* (2009).

education programs. Then, starting in the early 2000s, the federal government intervened with system wide testing requirements in most states in seven grade levels. Those requirements continue today. To many, testing seems omnipresent in our public schools.

It is no secret, however, that education professors tend to be less enthusiastic than the general public about testing mandates or externally administered standardized tests.² Nonetheless, by default our graduate schools of education, their libraries, and the scholarly journals they manage serve as the primary repositories of research on the uses and effects of educational testing.

In my "spare" time, I read research on the effect of testing on student learning. Over the years, I have reviewed thousands of studies and found several hundred that fit the requirements for a statistical meta-analysis, including hundreds of randomized controlled experiments—the "gold standard" in social science research—dating back to the 1910s. Among the many sources I found helpful were a 233-page *Bibliography of Educational and Psychological Tests and Measurement* from 1923 and a 1942 book by C. C. Ross, *Measurement in Today's Schools*—a source that led me to many other sources.

The "scientific" study of school testing—that is, the statistical analysis of test use and its effects—dates back to the 1890s. In 1923, standardized educational tests were still relatively new, but had already proliferated widely. The *Bibliography*, conducted for the U.S. Department of Interior, lists several hundred different tests and cites several hundred more reports of their implementation.

By 1942, many testing programs had been evaluated and dozens of experimental studies conducted. C. C. Ross, a former student of the testing and measurement pioneer Edward Lee Thorndike, references some of those studies in *Measurement in Today's Schools*. In the book's preface, he writes:

The rapid increase in the number of tests and scales published has made it impossible to keep the books [about educational testing] either complete or up to date. Fortunately, in recent years the appearance of rather complete and frequently revised bibliographies of published tests, together with

²For example, in a 1997 study by the survey firm Public Agenda, a national sample of education school faculty expressed substantially less support for high-stakes standardized testing than did other groups:

"[O]nly 49 percent believe raising the standards of promotion from grade school to junior high and letting kids move ahead only when they pass a test showing they've reached those standards, would do a great deal to improve academic achievement. In sharp contrast, the percentage reaches 70 percent among the general public (and 62 percent among teachers)."

"[W]hile supporting standards in concept, professors of education seem reluctant to put into place concrete, high-stakes tests that would signal when kids are meeting the standards."

"Fully 78 percent want less reliance on multiple-choice exams in the schools."

critical evaluations, has made detailed lists and descriptions of available measuring instruments in textbooks no longer necessary. . . .

At the same time the enormous expansion of the experimental literature relating to measurement has had to be considered in any course that is at all adequate. . . .

It appears to the author that the time has come for a critical appraisal of measurement in today's school, and for a careful search for generalizations to guide both theory and practice. The experimental evidence supporting these generalizations has been examined, and wherever possible reported in the language of the original author.

In 597 pages, *Measurement in Today's Schools* is both a how-to guide for developing tests and testing programs and a systematic review of the abundant research literature on test use from the first four decades of the twentieth century. In addition to more than a thousand footnotes and citations, most of Ross's several dozen chapters end with a section entitled "Selected References for Further Reading," in which the author provides bibliographic detail to help the reader find other books relevant to educational testing research—hundreds of books in all, including a few that one might have considered competitive rivals to his.³

Most of Ross's exhaustive coverage of the subject remains relevant today. To be sure, today's testing and testing research differ; there were no computer-delivered tests in 1942, for example. But, in most essential aspects, the use of tests, and how students and teachers relate to them, remain the same.

Fast forward to 1971, and much had changed. By then, for example, most of the old teacher apprenticeship "normal" schools had evolved into graduate schools of education, producing their own research and researchers.

In that year, the profession's flagship review journal *Review of Educational Research* published a literature review by one Marjorie C. Kirkland, then working at a military base branch campus of Alabama's Troy State University. "The Effects of Tests on Students and Schools" contains 234 references, many of

³Just some among the dozens of sources cited (these from pages 53–59) in Julian C. Stanley's 1954 third edition of Ross's *Measurement in Today's Schools*:

W., A. McCall, *How to Measure in Education* (New York: The Macmillan Company, 1922).

G.M. Ruch, G.D. Stoddard, *Tests and Measurements in High School Instruction* (Yonkers, NY: World Book Company, 1927).

C.W. Odell, *Educational Measurement in High School* (New York: D. Appleton-Century Company, 1930).

Other testimonies to the abundance of educational testing and empirical research in the first half of the twentieth century can be found in Lincoln & Workman 1936, 4, 7; Butts 1947, 605; Monroe 1950, 1461; Holman & Docter 1972, 34; Tyack 1974, 183; and Lohman 1997, 88.

which lead one to sources in her professional subfield of counseling and guidance, including a large number of articles on intelligence tests. Fewer lead to genuine research studies of the more quotidian "effects of tests on students and schools" as promised by the article's title.

Instead, one finds within Kirkland's forty-seven page article numerous cautions for and criticisms of educational test use, plus several bold declarations that little empirical research existed. For example:

Since these issues affect the lives of so many, and since so much has been written about tests, one would expect to find a great deal of empirical research in this area. However, a review of the literature revealed only a few small-scale and somewhat peripheral empirical studies. (306)

The search period for my meta-analysis encompasses Kirkland's—1910 to 1970. Yet, I found thirty-one empirical studies—mostly randomized experiments—on test use in the schools that she failed to mention.⁴ And, note that I was only looking within a subset of a larger research literature that Kirkland claimed to know intimately.⁵ Moreover, all but a few of the thirty-one studies that she overlooked were published in journals that she had allegedly included in her search. Of the 600+ authors mentioned in C.C. Ross's 1942 tome, Kirkland cited only sixteen. It bears remembering: anyone can claim to know a research literature; that doesn't necessarily mean that they do.

Kirkland's 1971 article may have been a one-off; at least my basic internet searches on her name reveal no other publications.

Fast-forward another decade to 1980 and another literature review, this one by researchers at the UCLA Center for the Study of Evaluation (CSE) (Lazar-Morrison, Polin, Moy, & Burry). Since the 1960s, UCLA's Graduate School of Education has hosted various manifestations of an education policy research center. (In more recent decades it has settled on the appellation Center for Research on Evaluation, Standards, and Student Testing (CRESST).)

The federal government, various foundations, and contract work have bestowed upon CRESST many millions over the years. Probably no group in the history of the world has held more resources or a greater mandate to conduct high quality and thorough reviews of the research literature on educational test use.

The Center's 1980 "Review of the Literature on Test Use," however, cited only fifty-five sources, of which six were their own. Moreover, only ten

⁴I also found thirteen qualitative studies that Kirkland failed to mention.

⁵I was looking only for studies of the effect of testing on student achievement, whereas Kirkland was, allegedly, looking for studies of all manner of effects on students and schools.

predated 1970. Apparently, the Center authors felt no need to review the pre-1970 literature because someone else had already done so—that someone being Marjorie Kirkland—and her effort allegedly "revealed only a few small-scale and somewhat peripheral empirical studies."

Center authors even ignored the little empirical research that Kirkland had included. Instead, they showcased at length Kirkland's assertions of a lack of research, and added dozens of their own, for example:

"a dearth of empirical support on actual test use practices." (3)

"then as now, little empirical research has accumulated on [instructional issues and testing]." (3)

"the bulk of the testing literature being a series of position papers citing little empirical data." (4)

"There is little empirical research available that can answer the questions that have arisen." (5)

"Virtually nothing is known about the amount of testing taking place using other types of assessments." (7)

"The literature on curriculum-embedded tests is equally scant." (8)

"The kinds of contextual factors which influence testing and the use of test results are just beginning to be appreciated." (9)

"The literature does not appear to reflect any great follow-up [regarding teacher competence with testing]" (9)

"As of yet, there is no evidence about how teacher attitudes toward other types of tests affect the use of those assessments." (19)

"The effect of the actual testing environment on test use is only beginning to emerge." (19)

"The investigation of these variables as factors affecting teachers' use of tests and test data is minimal." (20)

"In the community, parent involvement, accountability pressures, and news media coverage of test scores . . . have yet to be researched." (20)

"We know very little about the costs of testing." (20)

"There are a number of areas concerning teachers and testing for which there is no information." (24)

"The settings and factors which affect the use of tests and their results is yet another uninformed area." (25)

Raw declarations all: the authors provide no description of where or how (or even if) they looked for source material. The reader is expected to assume that they did.

The search period for my meta-analysis encompasses the Center's—from 1910 to 1980. Yet, I found ninety-two empirical studies—mostly randomized experiments—on test use in the schools that they failed to mention.⁶ And, again, I was only looking within a subset of a larger research literature that the authors claimed to know intimately.⁷ Moreover, I found fifty empirical studies from the 1971–1980 decade—the period of time in between Kirkland's review and theirs. Of the 600+ authors mentioned in C.C. Ross's 1942 book, the UCLA authors cited none.

After this inauspicious beginning, the UCLA center would enjoy decades of generous federal government and foundation funding ostensibly designated for the purpose of filling the many gaps in the research literature they had supposedly identified. Despite all the funding, however, few gaps seem to have been filled.

Dismissive statements similar to those above, alleging a lack of and need for research, would stream continuously from the Center's talks and publications from then on.

With one more decade came one more literature review from the same UCLA-based research center, now called the CRESST.⁸ Its newer effort comprises eight pages (4–12) within a larger report “Effects of Standardized Testing on Teachers and Learning—Another Look.”

The 1991 review contains thirty-nine citations. Twenty-nine lead to CRESST work and another seven to the work of close allies and frequent collaborators. The authors allotted but ten citations for the rest of the world.

For the same search period up to 1991, I found 148 empirical studies on test use in the schools that the CRESST authors failed to mention.⁹ And, again, I was only looking within a subset of a larger research literature that the authors claimed to know in depth.¹⁰ Moreover, I found fifty-six empirical studies from the 1981–1990 decade—the period of time in between the Center's earlier review and this one. Of the 600+ authors mentioned in C.C. Ross's 1942 book, the CRESST 1991 literature review cited none.

Typically, CRESST publications reference their own work, that of sympathetic allies and frequent collaborators, along with some recent work by others in the profession so well known that they cannot be ignored. Virtually all

⁶I also found twenty-eight qualitative studies that the authors failed to mention.

⁷I was looking only for studies of the effect of testing on *student achievement*, whereas the Center authors were, allegedly, looking for studies of all manner of effects on student and schools.

⁸With partners at the Universities of Colorado, Southern California, Pittsburgh, California at Santa Barbara, Arizona State and Stanford Universities, and the Rand Corporation.

⁹I also found fifty-seven qualitative studies that the authors failed to mention.

¹⁰I was looking only for studies of the effect of testing on *student achievement*, whereas the Center authors were, allegedly, looking for studies of all manner of effects on student and schools.

research conducted by the vast population of lesser knowns and dear departed is ignored, declared nonexistent, or denigrated as not worth mentioning.

Over its half-century life, the various iterations of CRESST have produced some good work with their taxpayer funding, albeit with a profound anti-testing bias. More importantly, however, they have suppressed, buried, dismissed, denigrated, and misrepresented a substantially greater quantity of relevant and useful evidence about educational test use and its effects.

The larger U.S. education research community would appreciate their effort. In due course, CRESST principals and trusted collaborators would comprise the dominant majority or plurality on the Board of Testing and Assessment at the National Research Council, and for the work on testing, standards, evaluation, or education accountability at the National Academy of Education and the International Academy of Education. They would capture the education policy committee of the educational testing and measurement profession's flagship professional organization, the National Council on Measurement in Education, and write the education policy section for its periodically updated encyclopedia, *Educational Measurement*. CRESST insiders would be elected president of the even larger and broader-in-membership American Educational Research Association in 1975, 1985, 1987, 1995, 1999, 2002, 2003, 2006, 2008, and 2015.

Jump ahead another decade to the earliest years of the new millennium. Former Texas governor George W. Bush was the Republican candidate in the presidential election of 2000 and, for the first time in American history, standardized testing in the schools emerged as a major national campaign issue. Education professors gang-tackled it with hundreds of anti-testing books, op-eds, panel discussions, and interviews.

In response to this mugging, Republican Party education policy wonks had hardly anything to say. For whatever reason, the GOP had long relied on economists for its more academic education policy information, and economists had paid little attention to education program evaluation, academic standards, or testing and measurement. These GOP-policy-advisor economists—along with a few political scientists—knew little of the rich research literature on educational test use and its effects in schools. Psychologists had conducted most of that research.

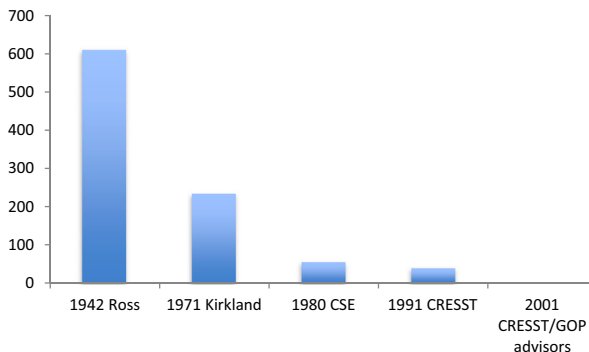
Yet, a national testing program was coming and the new Republican administration needed policy advice. If the advisors admitted to how little they knew, they risked forfeiting their places in the power elite just at the moment their side had taken over and was setting national policy. After all, genuine experts know the research literature in their field.

CRESST presented an attractive alternative: assert that no previous research exists and one could claim expert status with just a single new study.

To really know a research literature typically requires years of patient study. By pretending there to be no previous work to study, one can get right to work. Moreover, any work one does in a barren, blank-slate research field is "new," "first," and "pioneering." "First" research work is more prestigious, more likely to attract the public's attention, and more likely to be considered newsworthy by journalists.

In early 2003, shortly after the passage of the No Child Left Behind Act, with its educational testing mandate, the U.S. House Education and Workforce Committee would publish the following from a Brookings Institution-based GOP advisor in a press release: "It is important to keep in mind the limited body of data on the subject . . . We are just getting started in terms of solid research on standards, testing and accountability." With that, the acknowledged quantity of a century's worth of research on educational testing declined to zero.

Chart 1
Number of Studies on Testing in Schools
Reported by Researcher or Research Organization



For the same search period up to 2001, I found 188 empirical studies on test use in the schools that CRESST and the GOP policy advisors failed to mention.¹¹ Moreover, I found forty-one empirical studies from the 1991–2001 decade—the period of time between CRESST's earlier review and the passage of the No Child Left Behind (NCLB) Act, which would require annual school testing nationwide.

Mind you, my continuing work summarizing the research literature on the effect of testing on achievement is done on the cheap. I do not pay for most

¹¹I also found 124 qualitative studies and fifty-two survey studies that they neglected to mention.

studies hidden behind paywalls or found in distant libraries that require compensation for photocopying and mailing. I have not attempted to contact the many school districts and states that evaluated their system wide testing programs throughout the twentieth century. Yet, still I found hundreds of studies—most of them reporting on multiple experiments—which our country's most influential education policy scholars continue to declare nonexistent.

The Republican policy advisor alliance of convenience with CRESST continues today. Their information suppression method is so simple: reference only that work done by others within one's group and dismiss the rest, either by declaring it nonexistent or so inferior in quality as to not be worth mentioning. Dismissive reviews carry several advantages over engaging the wider research literature. A scholar, 1. saves time and avoids the tedium of reading the research literature; 2. adds to the in-group's citation totals while not adding to rivals'; 3. gives readers no help in finding rival evidence (by not even citing it); 4. attracts more attention by allegedly being "first," "original," "a pioneer"; 5. increases the likelihood of press coverage for the same reason; and, 6. increases the likelihood of research grant funding to "fill knowledge gaps."

Moreover, if one's dismissive review is popular in the profession because it hides unpopular research evidence, one may attract many sympathetic citations. It may be more common than not in education policy to reference only that research that supports one's preferences, and ignore all the rest. Thus, just a few erroneous but popular articles may achieve widespread dissemination even while many other accurate but unpopular articles are ignored. Moreover, with vigilance, this dynamic may continue forever.

Consider the case of one very popular, often interviewed, and widely cited researcher, a longtime CRESST affiliate who now teaches at Harvard's education school. He has been claiming for thirty years that little to no research exists on a topic of particular interest to him—test coaching (also known as test prep). He argues that the alleged lack of research is due to the difficulty of studying the topic and unwillingness of any school district administrator—among the tens of thousands of them—to cooperate in conducting such a study.

Although he has received generous funding over the past few decades to work toward filling the alleged gap in research, the gap apparently remains just as wide today as it was three decades ago. He would like more funding for himself and his colleagues so they might continue working on this profoundly relevant topic.

It doesn't require much searching, however, to find a cornucopia of test coaching studies, many of them randomized experiments. They date at least as

far back as 1953. I managed to find seventy-six relevant studies published between 1955 and 2019. Meta-analyses or research summaries of test coaching studies were published in 1981, 1983(3), 1984, 1990, 1993, 2005(2), 2006, and 2017. There's even a U.S. Education Department summary of some college admission test prep studies available on the internet, though it only dates back to the mid-90s.

To be sure, some of the studies were conducted by or for test development organizations that, some might say, had a vested interest in a particular result. But most were not.

Confronted by a research volume of this extraordinary size, the advantage of declaring it nonexistent looms apparent. Were the Harvard education professor to acknowledge the full extent of the research literature, he would not only be found wrong in his assertions and risk his status as the topic's foremost expert, he could be forever mired in time-consuming debates. By simply denying the existence of the contrary body of research, all those studies disappear as if by magic. Poof, they're gone. Furthermore, he can persist in effectively dismissing the research literature so long as the most important information gatekeepers in education research persist in consulting, referencing, and directing all attention toward his work and that of cooperative allies, and not the others.

Scholars worldwide continue researching educational test use and its effects, as they have for over a century, and CRESST affiliates and allies continue to claim that most such work does not exist, or couldn't be any good if it did exist. They are working on those issues themselves, they say, with more "rigorous" and "sophisticated" methods. Moreover, to fill the lingering research gaps, they need more money from taxpayers and foundations to continue their work.

Like Peanuts' Charley Brown trying to kick Lucy's football, we keep giving them more money.

The end result is that only a fraction of relevant research and experience has been incorporated into current U.S. educational testing policy. Policymakers have not consulted the most knowledgeable experts. And, current U.S. educational testing programs are structured unlike those anywhere else in the developed world—burdensome, unpopular, inefficient, ineffective, and uninformative.