Correcting Fallacies About Educational and Psychological Testing

Edited by

Richard P. Phelps

American Psychological Association • Washington, DC

Published by
American Psychological Association
750 First Street, NE
Washington, DC 20002
www.apa.org

To order
APA Order Department
P.O. Box 92984
Washington, DC 20090-2984
Tel: (800) 374-2721; Direct: (202) 336-5510
Fax: (202) 336-5502; TDD/TTY: (202) 336-6123
Online: www.apa.org/books/
E-mail: order@apa.org

In the U.K., Europe, Africa, and the Middle East, copies may be ordered from
American Psychological Association
3 Henrietta Street
Covent Garden, London
WC2E 8LU England

Typeset in Goudy by Stephen McDougal, Mechanicsville, MD

Printer: Maple-Vail Book Manufacturing Group, York, PA
Cover Designer: Mercury Publishing Services, Rockville, MD
Technical/Production Editor: Harriet Kaplan

Library of Congress Cataloging-in-Publication Data
Correcting fallacies about educational and psychological testing / edited by Richard P. Phelps.
p. cm.
Includes bibliographical references and index.
ISBN-13: 978-1-4338-0392-5
ISBN-10: 1-4338-0392-5
1. Educational tests and measurements—Standards. 2. Psychological tests—Standards. 3. Employment tests—Standards. 4. Professional education—Standards. I. Phelps, Richard P.

LB3051.C6386 2009
371.26'2—dc22 2008026233

British Library Cataloguing-in-Publication Data
A CIP record is available from the British Library.

Printed in the United States of America
First Edition

EXHIBIT 2
Professional Standards and Guidelines for Test Use: Selected Citations

American Counseling Association & Association for Assessment in Counseling. (2003). *Responsibilities of users of standardized tests* (3rd ed.). (RUST). Alexandria, VA: Authors.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing.* Washington, DC: Authors.

American Federation of Teachers, National Council on Measurement in Education, & National Education Association. (1990). *Standards for teacher competence in educational assessment of students.* Washington, DC: Authors.

American Psychological Association. (1992). *Ethical principles of psychologists and code of conduct.* Washington, DC: Author.

American Psychological Association, Practice and Science Directorates. (2000). *Report of the task force on test user qualifications.* Washington, DC: Author.

American Psychological Association. (2001, May). *Appropriate use of high-stakes testing in our nation's schools.* Washington, DC: Author.

Association for Assessment in Counseling and Education. (2002, November). *Standards for educational and psychological testing—What counselors need to know.* Alexandria, VA: Author.

Association for Assessment in Counseling. (2003). *Standards for multicultural assessment.* Alexandria, VA: Author.

Association of Test Publishers. (2002). *Guidelines for computer-based testing.* Washington, DC: Author.

International Personnel Management Association—Assessment Council. (2004, June). *Policies and procedures manual.* Alexandria, VA: Author.

International Test Commission. (2001). International guidelines for test use. *International Journal of Testing, 1,* 93–114.

International Test Commission. (2005). *International guidelines on computer-based and Internet-delivered testing.* Retrieved July 19, 2008, from http://www.intestcom.org/guidelins/index.html

Joint Advisory Committee. (1993). *Principles for fair assessment practices for education in Canada.* Edmonton, Alberta, Canada: Author.

Joint Committee on Standards for Educational Evaluation. (2007). *The personnel evaluation standards.* Thousand Oaks, CA: Sage.

Joint Committee on Standards for Educational Evaluation. (2007). *The program evaluation standards* (2nd ed.). Thousand Oaks, CA: Sage.

Joint Committee on Standards for Educational Evaluation. (2007). *The student evaluation standards.* Thousand Oaks, CA: Sage.

Joint Committee on Testing Practices. (2005). *Code of fair testing practices in education.* Washington, DC: Author.

National Association of College Admission Counselors. (1988). *Statement of principles of good practice.* Alexandria, VA: Author.

National Commission for Certifying Agencies. (2004). *Standards for the accreditation of certification programs.* Washington, DC: National Organization for Competency Assurance.

National Council on Measurement in Education Ad Hoc Committee on the Development of a Code of Ethics. (1995). *Code of professional responsibilities in educational measurement.* Washington, DC: Author.

Nester, M. A., Bruyere, S., & Wall, J. (2003). *Pre-employment testing and the ADA.* Alexandria, VA: Association for Assessment in Counseling and Education; Cornell, NY: American Rehabilitation Counseling Association.

Test Taker Rights and Responsibilities Working Group of the Joint Committee on Testing Practices. (1998, August). *Rights and responsibilities of test takers: Guidelines and expectations.* Rockville, MD: American Speech–Language–Hearing Association. Retrieved July 19, 2008, from http://www.asha.org/docs/html/RP2002-00198.html

Society of Industrial and Organizational Psychology. (2003). *Principles for the validation and use of personnel selection procedures* (4th ed.). Bowling Green, OH: Author.

U.S. Department of Labor & U.S. Department of Justice, Civil Service Commission, Equal Employment Opportunity Commission. (1978). Uniform guidelines on employee selection procedures. *Federal Register, 43*(166), 38290–39315.

3

EDUCATIONAL ACHIEVEMENT TESTING: CRITIQUES AND REBUTTALS

RICHARD P. PHELPS

Fallacies in critiques of educational achievement testing are so numerous that an entire book could be written about them alone. Indeed, an entire book already has been written about them: *Defending Standardized Testing* (Phelps, 2005a). It includes separate chapters from four of the current volume's contributors—Kurt Geisinger, Ron Hambleton, Stephen Sireci, and me—as well as other testing experts. *Defending* is an excellent resource, and there is no cause to replicate it here. Instead, in this chapter, I focus on several fallacies that were not included in the earlier text. For interested readers, Table 3.1 lists many of the educational achievement testing fallacies dissected in *Defending Standardized Testing* and elsewhere, along with citations to expert rebuttals to those fallacies. Still more point–counterpoint can be found in chapter 2 of another book, *Kill the Messenger: The War on Standardized Testing* (Phelps, 2003). Citations to some of the many sources of the fallacies are included in these other two books.

The views expressed here are the author's own and not necessarily those of ACT, Inc.

TABLE 3.1
Selected Testing Opponent Fallacies and Rebuttal Sources

Fallacy	Rebuttal source
Teacher grading and testing is more valid and more reliable than standardized testing.	Brookhart, 1993; McMillan, 2001; Stiggins and Conklin, 1992
Tests are developed secretively and obscurely.	Sireci, 2005
The public opposes the use of high-stakes tests.	Phelps, 1998, 2005b
Tests promote improper test preparation and teaching to the test.	Phelps, 2003; Crocker, 2005; Roediger and Karpicke, 2006a, 2006b
The best way to prepare students for tests is to substitute focused test preparation for regular subject matter instruction.	Camara (chap. 4, this volume); Crocker, 2005; Moore, 1991; Palmer, 2002; Tuckman, 1994; Tuckman and Trimble, 1997
One can perform well on multiple-choice tests without knowing the subject matter simply by learning tricks.	Becker, 1990; Briggs, 2001; Powers and Rock, 1999
Multiple-choice penalizes deep thought and creativity.	Powers and Kaufman, 2002; Roediger and Marsh, 2005
Constructed-response test items are superior to multiple-choice test items.	Bridgeman, 1991; Feinberg, 1990; Rudman, 1992; Traub, 1993
Standardized tests reduce educational achievement.	Phelps, 2005c
Tests inevitably narrow the curriculum.	Crocker, 2005; Phelps, 2003; Roediger and Marsh, 2005
Tests are too costly in money, time, and lost opportunity.	Goodman and Hambleton, 2005; Phelps, 1994, 2000a, 2003;
Standardized tests do not measure what is important.	Camara (chap. 4, this volume); Goodman and Hambleton, 2005; Phelps, 2003
A test can only validly be used for a single purpose.	Eckstein and Noah, 1993
There is more testing in the United States than in their countries.	Phelps, 1996, 1997, 2000b
Large-scale assessments are full of biased test items.	Camara (chap. 4, this volume); Goodman and Hambleton, 2005; Phelps, 2007b
Too much emphasis is placed on a single test score.	Camara (chap. 4, this volume); Goodman and Hambleton, 2005; Phelps, 2007b
Passing scores are set arbitrarily.	Plake, 2005; Sireci, 2005
Passing scores are set too high.	Goodman and Hambleton, 2005; Plake, 2005
School score trends are too "volatile" to be useful.	Bourque, 2005; Rogosa, 2005
Education accountability systems rely solely on tests.	Goodman and Hambleton, 2005; Phelps, 2003
Ability and achievement are completely unrelated.	Gottfredson (chap. 1, this volume); Lohman, 2006
Standardized tests are unfair to women and minorities.	Cole and Willingham, 1997; Farkus, Johnson, Immerwahr, and McHugh, 1998; Sandham, 1998
Standardized tests are unfair to students with disabilities.	Geisinger, 2005

In this chapter, I focus on fallacies that have less to do with the character of educational achievement testing and more to do with the dissemination (or lack thereof) of information about such testing. I describe several fallacies that are the direct result of either wholesale censorship and information suppression or naive beliefs about how information, particularly education research, is disseminated in the United States. I also briefly describe the policy implications of widespread belief in each of the fallacies. Public policies based on fallacies are not likely to be optimal. First, however, I introduce some terminology.

Educational achievement (or *proficiency*) *tests* are designed to measure what has been learned. In program evaluation terminology, achievement tests are generally *summative*, particularly when they have *stakes* (i.e., consequences). When achievement tests are used, instead, to monitor progress, or set benchmarks, they can be identified as *formative*.

Achievement tests are meant to measure the level of knowledge or skill attained within a *content domain*, or subject matter area. *Standards-based* (or *criterion-referenced*) achievement tests are designed to cover a predetermined, and sometimes legally mandated, body of subject matter content, usually identified by *content standards*. *Norm-referenced* achievement tests are designed to measure a student's level of knowledge and skill relative to a norm group—typically a representative sample of students from a large population of interest (e.g., all U.S. fourth-graders)—and cover a content domain that is determined by the test developer and not through a state or local political process.

By definition, most teacher-made classroom tests are achievement tests. This chapter, however, focuses on large-scale, systemwide, standardized achievement tests and some of the attendant fallacies proffered by their critics.

This chapter is organized according to popular fallacies about education achievement testing pertaining to its cost, to score "inflation," to the research literature on the effects of testing on achievement, and to the cause of widespread misunderstanding about such testing. However, it could just as validly have been organized around a single overarching fallacy: You can't stop progress. Many believe the other fallacies simply because they are all they ever hear and, in some quarters, all they are allowed to hear. Counterpoints are censored, suppressed, or obfuscated, and those who dare to speak them may be demonized, ostracized, threatened, or otherwise silenced. A century's worth of research on educational achievement testing has been so successfully removed from the collective working memory that it was not even considered by policymakers in the design of the most far-reaching federal testing mandate in our country's history. Instead, the research was declared nonexistent. Critics of educational achievement testing have not only stopped research progress, they seem to be well along the way to reversing it.

Those outside the field may assume that education research is like any other type of research, to wit: The most prominent research has been fully

vetted and so can be trusted; the most celebrated researchers are most likely the best; a wide range of evidence and points of view are fairly considered; and there is progress. Unfortunately, none of this validly characterizes contemporary research on educational achievement testing. In reality, educational achievement testing represents a threat to the status quo (because it monitors its productivity) and is treated as such.

Within the overarching fallacy, this chapter is organized by four other fallacies, each of which is illustrated by case studies:

- Fallacy 1: Tests cost too much.
- Fallacy 2: High stakes induce artificial test score increases.
- Fallacy 3: There is little evidence of the effects of testing and no evidence of its benefits.
- Fallacy 4: Testing is mischaracterized because it is difficult to understand.

Placed at the end of the chapter is a longer section on the policy implications of the overarching fallacy. The proved ability of vested interests to stop (and reverse) research progress on educational achievement testing portends the elimination of much of the past century's accumulated wisdom on the topic—an extinction vortex.

OVERARCHING FALLACY: YOU CAN'T STOP PROGRESS

A key component of our faith in progress is the corollary belief that our knowledge base continually expands—that is, we know what we already know, and we are always learning more. The continual expansion of knowledge requires both that the historical accumulation of knowledge be preserved and that new knowledge be welcomed. Moreover, modern society is so free and open, and the means of communication are now so varied, effortless, and cheap, that information suppression might be considered impossible. Ironically, the widespread belief that a continuous expansion of knowledge is inevitable and unstoppable helps to make its contraction possible because people do not consider the powerful constraints on research dissemination. To be sure, opponents of educational achievement testing neither burn books nor attempt to ban them—but then, they do not need to. Other, more subtle methods work well enough.

The simplest means of suppressing unwanted information is to ignore it—to pretend or even to declare that it does not exist. Most of the efforts to suppress thousands of scholarly studies of the effects of testing conducted over the past century have been of this type. There are several advantages to this method. First, it seems benign and not antagonistic because it is nonconfrontational (by contrast, asserting that someone has declared an extant research literature nonexistent can seem personal and antagonistic).

Second, if claimants are caught erroneously affirming the nonexistence of information, they can simply fall back on the excuse of innocent ignorance (i.e., there is plausible deniability). Third, declaring information nonexistent discourages efforts to look for it, thus helping to make such declarations self-fulfilling prophecies. Fourth, declaring nonexistent any research that competes with one's own helps to eliminate competition in the marketing of one's work.

Regression of research supports both the dominant ideologies in education, in which testing is considered bad for a variety of reasons, and the self-interest of the dominant groups. For with no valid, reliable, externally administered measure of their performance, these groups are free to do as they please.

High-Stakes Summative Testing: An Example of Poor Information Dissemination

The Association for Supervision and Curriculum Development (ASCD) is a professional organization with more than 160,000 members who

> span the entire profession of educators—superintendents, supervisors, principals, teachers, professors of education, and school board members. ASCD was initially envisioned to represent curriculum and supervision issues. Over the years, its focus has changed, and it now addresses all aspects of effective teaching and learning, such as professional development, educational leadership, and capacity building. (ASCD, 2005)

Because the ASCD is so large and offers its publications as part of membership, its books on testing (i.e., on authentic assessment or portfolios) can often be found at the top of best-selling rankings in the testing, assessment, standards, or accountability categories. It represents a powerful voice on these topics.

Each day I receive the ASCD's *SmartBrief*, an e-mail newsletter that contains a list of hyperlinked news stories that the organization considers interesting or important as well as job and sponsor advertisements and announcements of ASCD's own services and publications (ASCD, 2008). Also included are daily reminders of the positive value and importance of "authentic," formative, and performance-based testing, along with references and links to relevant books, professional development workshops, Internet instructional guides, and more. For any member interested in knowing more about authentic assessment, formative assessment, or performance testing, the ASCD offers a cornucopia of information and instruction.

However, what of the ASCD member interested in knowing more about the far more prevalent and consequential summative, selected-response standardized test? I have received ASCD *SmartBriefs* for more than a year and have yet to witness anything but an occasional blanket condemnation of

such testing. For ASCD members curious to know more about the accountability testing that represents such a central part of their working lives—and how to cope with that testing—the ASCD offers no help whatsoever. In view of its failure to report research on tests with selected-response formats, one could conclude that ASCD advocates authentic assessment and promotes it as the only legitimate form of assessment.

Like many professional educator associations these days, the ASCD is an advocacy organization run by "soft despotism" or a "tyranny of the majority"(de Tocqueville, 1835 and 1840/2003). The leadership apparently has decided that authentic assessment is better than testing with selected-response formats, so it promotes only authentic assessment. Indeed, it acknowledges only authentic assessment as legitimate.

Consider one day's *SmartBrief* (ASCD, 2008). One link takes the reader to a local newspaper's editorial titled "Florida Needs to Move Beyond Testing." A link to an article in another ASCD publication is introduced thusly: "It's old news that high-stakes, summative assessment practices don't help students learn, although word hasn't yet trickled up to politicians."

During the 2000 presidential campaign, the ASCD commissioned two public opinion polls on standardized testing, which was at the time a prominent campaign issue. The unpublished results showed the public strongly supportive of high-stakes standards-based testing. The ASCD then shelved one of the two polls and, with the other, added the percentage of neutral responses (e.g., "don't know," "no opinion") to that of the negative responses. This creative arithmetic spawned statements such as, "Approximately half [of the poll respondents] disagree or are undecided about whether these tests should determine graduation." The actual results for that poll item were as follows: "agree," 51%; "disagree," 38%; and "neither," 11% (Phelps, 2005b, pp. 18–20).

The ASCD is not exceptional in practicing a soft form of censorship and information suppression. During the 2000 presidential campaign, I visited the Web sites of dozens of national organizations of professional educators and observed a propensity to present to their members evidence and points of view from only one side of the debate—the antitesting side. I wrote each of them with suggestions for links to Web sites (e.g., Mass Insight, the Southern Regional Education Board, Educational Testing Service) that offered alternative perspectives. None of the organizations added any of the suggested references.

Research Progress: 1895 to 1985

The other day, I was reading C. C. Ross's (1941) authoritative text *Measurement in Today's Schools* and was struck by its policy relevance. Ross summarized the findings of hundreds of research studies across more than 4 decades of scholarly effort and included one long chapter on the motiva-

tional effects of testing and other chapters on the optimal design of testing programs. Although relevant and informative, one sees scant reference to Ross's work these days. Moreover, to my knowledge, neither his insights nor those of any of the many research psychologists he cited have been considered in any of the recent testing program or accountability system design discussions among national policymakers or their advisors. Here are just some of the research findings included in *Measurement in Today's Schools* that could have helped policy planners recently—for example, in the design of the federal No Child Left Behind (NCLB) Act:

- Standardized tests with consequences influence teachers' degree of effort as well as their allocation of time and emphasis toward topics known or assumed to be covered on the tests (pp. 333–334).
- When past examination content is assumed to predict future examination content, or future examination content is otherwise known in advance, schools can become cramming schools, defeating the purpose of examinations (p. 334).
- Students learn more when there are known consequences to that learning.[1]
- Tests with consequences influence students' allocation of study time toward topics known or assumed to be covered on the tests; tests without consequences do not (pp. 360–361).
- The act of testing alone, irrespective of other factors, tends to improve achievement (p. 342).

Some other research findings covered in Ross's work relevant to contemporary testing policy discussions include the following:

- Using a single threshold, or cut score, for passing a test tends to motivate most those students whose academic performance would place them just below that threshold, but not the students whose academic performance places them comfortably above the threshold or some substantial distance below (pp. 334–335).
- Students learn more when they are made aware of the quality of their academic performance and allowed an opportunity to adjust (p. 336).
- Frequent testing helps low-achieving students more than it helps high-achieving students (pp. 340, 342–343, 362).
- Examination strengthens memory and the sooner an exam is given after exposure to the information the stronger will be the memory of that information (p. 341).

[1]Note that the NCLB Act applied consequences for performance to schools but not to students.

- Testing with review or feedback improves achievement even more (pp. 336, 342–343).
- The simple awareness that there will be an examination later on substantially improves achievement (if the examination has consequences)—both on expected subject matter and even on unexpected subject matter (pp. 343–346).

Indeed, Ross in his 1941 book summarized research topics that some current researchers have claimed to be the first to study. If Ross (who unfortunately passed away years ago) could have been involved in the crafting of the NCLB Act, its design and execution might have been better informed.

As one would expect, however, research on the effects of testing and the optimal design of testing and accountability programs did not stop in 1941. Instead, the large body of research that had accumulated to that point stimulated additional work that was more detailed, more sophisticated, and more varied. Research pertinent to testing and accountability policy blossomed over the next half century, with thousands of studies conducted by research psychologists, educational practitioners, and program evaluators.

Research on the effects of testing has been conducted in many fields (and subfields), of inquiry, including education research (language learning, mastery learning, remedial–developmental, gifted and talented, assessment, promotion and retention, selection, higher education assessment, admission, diagnosis, motivation, certification and licensure, school effectiveness, adult education), psychology (memory and cognition, industrial–organizational, selection, diagnosis and counseling, clinical, personnel, educational psychology, allocation, motivation), program evaluation, and sociology. Furthermore, research on the effects of testing has been sponsored by many types of organizations, such as educational institutions (at all levels), governments (at all levels), international organizations, the military, and most of the world's developed countries. Of all these categories, I have thus far conducted a thorough search through the literature of only the first: education research. There I found more than 1,000 studies on the effects of testing and test-based accountability. Some of the major concentrations of evidence surround particular research themes, such as those listed in Exhibit 3.1.

The research literature is deep and wide, painstakingly constructed over the course of a century by hundreds of hardworking, earnest scholars, including some of the giants in the history of psychology. Recent claims of a barren research literature and "first-ever" studies on the aforementioned topics should be placed for comparison alongside this mountain of knowledge and experience, then judged accordingly.

Research Regress: 1985 to Present

Research progress on educational achievement testing began to stall in the 1980s when the U.S. federal government started funding specialized re-

search centers run by regressive groups, and these, in turn, co-opted other well-funded federal institutions—most notably the Board on Testing and Assessment at the National Research Council (NRC).

A *regressive* research group is one that denies or dismisses information and evidence, such as that embodied in a research literature. Typically, the dismissed research is replaced in their words and publications with their own and that of those who share their point of view. To be sure, more than one group in the field has denied or dismissed information and evidence related to educational achievement testing. However, the Center for Research on Evaluation, Standards, and Student Testing (CRESST) in particular has been the most successful at it. First funded by the federal government in the 1980s—the era of *A Nation at Risk* (National Commission on Excellence in Education, 1983)—as the Center for the Study of Evaluation, CRESST's direct funding from U.S. taxpayers for the period 1996–2006 alone totaled more than $35 million. Headquartered in the education schools at University of California, Los Angeles (UCLA) and the University of Colorado, CRESST has joined with various partners over the decades, most notably the Rand Corporation and the University of Pittsburgh.

CRESST's primary advantages in selling its ideas (and suppressing others) are money and the aura of intellectual authority. It has vastly greater financial resources at its disposal than do any of the individual scholars who might wish to contest its claims. Moreover, it has leveraged those resources productively to expand its reach even further. For example, CRESST can easily make deals (i.e., partnerships) with other researchers and research centers because it can offer pay, in-kind services, publicity, dissemination of publications, and more. CRESST appears authoritative for the simple reason that it has been the only federally funded research center uniquely devoted to the topics of standards and testing. It is the logical place for journalists to call for expertise about educational achievement testing. As I mentioned previously, CRESST is not nor has it ever been the only organization promoting regression in research on educational achievement testing; however, I argue that it has been the most successful and the most important. The following case studies illustrate how and why.

FALLACY 1: TESTS COST TOO MUCH

To people outside the field of education, the cost of standardized student testing would likely seem a rather straightforward topic. Within the field, however, it is an anxiety-producing subject that spawns tense arguments. These arguments tend to turn on the worth or intrinsic educational value of the tests themselves, the amount of time taken up by test taking and test preparation, and the assignment or lack thereof of particular cost components as attributable to standardized testing.

EXHIBIT 3.1
Selected Listing of Effects-of-Testing Research and Researchers

The mastery learning and mastery testing experiments conducted from the 1960s until the present vary incentives, frequency of tests, types of tests, and many other factors to determine the optimal structure of testing programs. Researchers have included such notables as L. W. Anderson, J. H. Block, B. Bloom, R. B. Burns, J. B. Carroll, K. P. Cross, S. L. Gates, T. R. Guskey, G. M. Hymel, E. H. Jones, F. Keller, J. Kulik, C.-L. Kulik, J. R. Okey, M. Tierney, and T. L. Wentling.

Psychologists' experimental work on memory retention and loss dates back more than a century and studies the optimal frequency of testing and other factors, again to determine the optimal structure of testing programs. Researchers have included H. A. Greene, E. H. Jones, A. N. Jorgensen, and C. C. Ross, who have been mentioned in the text of this chapter, as well as L. W. Anderson, R. L. Bangert-Drowns, G. Hanna, L. K. Henry, N. Keys, A. Khalaf, J. E. Kirkpatrick, J. Kulik, C.-L. Kulik, and B. F. Skinner.

Language acquisition researchers attempt to optimize the use of testing in language instruction and so are keen students of the "washback" (or backwash) effect of testing. Researchers have included J. C. Alderson, K. M. Bailey, J. B. Carroll, A. Hughes, and D. Wall.

Developmental (i.e., remedial) education researchers have conducted many studies to determine what works best to keep students from failing in their "courses of last resort," after which there are no alternatives. Researchers have included L. Bliss, B. Bonham, H. Boylan, D. Chang, S. Chen, C. Claxton, R. Kirk, J. Kulik, C.-L. Kulik, R. McCabe, J. Roueche, C. Schonecker, and C. Wheeler.

The vast literature on effective schools dates back a half century and arrives at remarkably uniform conclusions about what works to make schools effective—goal setting, high standards, and frequent testing. Researchers have included D. A. Astuto, C. R. Clark, K. Cotton, T. L. Good, D. A. Grouws, R. Kiemig, M. Jones, D. U. Levine, L. W. Lezotte, L. S. Lotto, S. C. Purkey, M. Rutter, M. S. Smith, A. Taylor, B. Valentine, and B. M. Wildemuth.

The many studies of district and state minimum competency or diploma testing programs popular from the 1960s through the 1980s found positive effects for students just below the cut score and mixed effects for students far below and anywhere above it. Researchers have included A. L. Abrams, J. L. Anderson, D. J. Bateson, B. Battiste, K. Bembry, D. Blackmore, H. Boylan, S. M. Brookhart, T. B. Corcoran, C. Fincher, W. D. Hawley, M. L. Herrick, F. H. Hultgren, M. Jackson, J. Jacobsen, S. Mazzoni, R. L. Mendro, W. Muir, T. Orsack, W. T. Rogers, D. P. Saxon, W. D. Schafer, C. C. Seubert, D. E. Tanner, W. J. Webster, D. Weerasinghe, and M. A. Zigarelli.

Many researchers have studied the role of testing in motivation. They have included R. Bootzin, S. M. Brown, C. Chen, M. V. Covington, T. J. Crooks, R. Drabman, A. Kazdin, K. D. O'Leary, J. W. Olmsted, S. L. Pressey, L. B. Resnick, D. P. Resnick, A. Staats, H. W. Stevenson, R. W. Tyler, H. J. Walberg, and R. G. Wood.

Others have considered the role of tests in incentive programs. These researchers have included R. Bootzin, J. Cameron, T. B. Corcoran, A. P. Csanyi, R. Drabman, M. A. Gonzales, L. Homme, A. Kaszdin, J. McMillan, K. D. O'Leary, W. D. Pierce, M. R. Rechs, A. Staats, and B. L. Wilson.

International organizations, such as the World Bank or the Asian Development Bank, have studied the effects of testing on education programs they sponsor. Researchers have included N. Brooke, U. Bude, D. W. Chapman, S. P. Heynemann, J. Oxenham, B. Pronaratna, G. Psacharopoulis, A. Ransom, A. Somerset, C. W. Snyder, and E. Velez.

Another major area of inquiry dating back many decades has been the effect of goal setting, standards, and alignment on teachers, instruction, and student learning. The researchers involved have included M. Csikszentmihalyi, J. Fontana, F. B. Knight,

M. A. Lowther, I. Panlasigui, C. Pine, M. Pomplun, L. B. Resnick, B. H. Robinson, K. M. Shaw, J. S. Stark, T. D. Thomas, and R. W. Tyler.

Finally, there has been considerable research on the learning effect of test taking, conducted by F. N. Dempster, S. M. Luipersbeck, H. L. Roediger, and T. C. Toppino, among others.

Note. For more detailed information, see Phelps (2005c).

If one chooses to believe, for example, that standardized test-taking and test-preparation time have no intrinsic instructional value and, further, that standardized tests are separate from and contribute nothing to the instructional plan of a school, then one might well consider standardized tests to be costly because they take up time that might otherwise be devoted to instruction. To such critics, the problematic costs associated with standardized tests are not represented by the purchase price paid to the commercial vendors but, rather, by the lost opportunity for learning that could have taken place in the time devoted to standardized tests.

Case Study: The Texas Teacher Test

Local son and corporate leader H. Ross Perot headed a blue ribbon commission in the early 1980s that studied the Texas school system, long considered one of the country's poorest performers. Two findings of the commission were that some Texas teachers were illiterate and that there were no high-stakes requirements for new teachers. The commission recommended the development of a basic literacy test, the Texas Examination of Current Administrators and Teachers (TECAT) and a requirement that all teachers pass it. By all accounts, the test was extremely easy, but nonetheless some teachers failed it—some after multiple attempts.

CRESST conducted a cost–benefit analysis of the test and declared its net benefits to be negative—by about $70 million (Shepard, Kreitzer, & Grau, 1987). Indeed, CRESST was extremely critical of every aspect of the test. CRESST recommended that the test not be high stakes and that if the test were used at all, failure should at most mean a teacher would be required to take a literacy course. CRESST's calculations, and my recalculations, can be found elsewhere (Phelps, 2003, pp. 105–115); I only summarize them here. CRESST attained its negative net-benefit figure through the following items:

- arbitrary exclusions of benefits (e.g., salary savings for more than half the teachers dismissed for failing the test—those in vocational education, industrial arts, special education, business education, and kindergarten—did not "count" in Shepard et al.'s [1987] benefit calculations because, the authors argued, literacy is not important in their work);
- arbitrary inclusions of costs (e.g., teacher time spent taking the test during one of their prescribed-topic in-service days is

counted as a pure cost, implying that tests are not acceptable vehicles for teaching subject matter; by contrast, passive listening to a lecture on literacy instead would not have been considered a cost);

- miscalculations of the value of time (they valued teachers' after hours at their full salary rate and ignored the future [discounted] value of recurring benefits); and
- counting certain costs as gross that should have been counted as net (i.e., to include the value of countervailing benefits).

Correcting only for the more obvious of Shepard et al.'s (1987) mistakes and using their own base assumptions and estimates pushes the TECAT program's net benefits into the black—and by a wide margin—to $330 million. Other fixes to their calculations, methods, and assumptions push the net-benefit figure still higher.

Policy Implications

CRESST's preference was to preserve the status quo, eschew accountability requirements, and continue citizens' sole reliance on input measures and trust in the schools' own quality control to provide the teachers who taught their children at their expense. Shepard et al. (1987) also criticized the TECAT as simplistic, too narrow in format, and too general in content, but they did not advocate a "better" testing program. They favored eliminating teacher tests altogether.

Shepard et al. (1987) repeatedly attacked the test for its alleged low-level nature. Yet ultimately that is beside the point, because the authors were opposed to any type of teacher test. It is beside the point, too, because it is clear that the citizens of Texas wanted some accountability in their teacher certification system and would not have been content with the minor modifications of the status quo—consisting of more input requirements—that the authors recommended. Even the authors admitted that half the teachers interviewed thought the test accomplished its purpose: "to weed out incompetent teachers and reassure the public" (p. iv).

An alternative that CRESST did not consider was to move the TECAT to an earlier point in the teacher training process, say, at the end of, or even at the beginning of, graduate school. This would have met the concerns of the citizens of Texas; it would have achieved all the same benefits. However, most of the costs that Shepard et al. (1987) enumerated would have evaporated. There would have been no loss of teacher time. The responsibility for preparing the teachers for the test would have been placed on the teacher training schools or, better, on the potential education students themselves. Best of all, the time of unqualified would-be teachers (and their students) would not have been wasted. A reasonable alternative to the authors' complaints about the alleged simplistic nature of the TECAT would have been

to initiate a required "higher level" exam for teachers, in addition to the TECAT.

As it turns out, the citizens of Texas did not follow CRESST's advice. Rather, they followed the path just drawn, making the basic literacy exam an entrance exam for education school and requiring new teachers to pass another, newly created exam that focused on each teacher's content area and on pedagogy and professional development. They increased the benefits and reduced the costs, even according to CRESST's creative cost–benefit accounting criteria. Finally, they ended up with more tests, not fewer.

Case Study: The General Accounting Office Report on the Extent and Cost of Testing

In the early 1990s, the U.S. Congress asked its research agency, known then as the General Accounting Office (GAO), to estimate the extent and cost of systemwide standardized testing in the country and the potential overlap of President George H. W. Bush's proposed American Achievement Tests on that amount and cost. To complete its study, the U.S. GAO (1993) developed and administered surveys of local district and state testing directors and achieved a high rate of response from a nationally representative population. A who's who of notables in the evaluation, statistical, and psychometric worlds reviewed various aspects of the study. Nothing like it in quality or scale had ever been done.

During each of the next 3 years, CRESST invited papers and hosted panel discussions on the cost of testing at its annual conferences. The panels were populated by authors of other, competing studies of testing costs, including one sponsored by CRESST. The GAO report was lambasted as simplistic and poorly done. The primary accusation was that it did not consider personnel costs (e.g., the cost of teacher time spent proctoring exams). In fact it had, with personnel costs accounting for more than half of its cost estimates.

Having been involved in the GAO study as project director, I protested to the CRESST directors for the misrepresentation and for their refusal to allow me to join any of the panels. A vague promise of a correction in some future CRESST newsletter was hinted at but never fulfilled. Protests made to the researchers directly responsible for the false accusations were similarly ignored.

The characterization of the GAO report as "flawed" spread unimpeded. In its place, other reports were promoted and published purporting to show that standardized tests are enormously costly and overwhelm school schedules in their volume. The studies were based on (a) a single field trial in a few schools, (b) three telephone calls, and (c) one state (the CRESST report on testing costs was limited to Kentucky), or (d) the facts were just made up. The studies that used some data for evidence heaped all sorts of nontest ac-

tivities into the basket and called them costs of tests. In the case of CRESST's Kentucky report, some teacher-respondents counted their entire school year as "test preparation time," and that time was then multiplied by classroom teachers' wage rates and counted as a cost of testing (Picus & Tralli, 1998).

It was only after several years and after the original directors of CRESST relinquished some of their directorial duties that anything was done about their continuing misrepresentation of the GAO report. A new director consented to correct one paragraph in CRESST's Kentucky report that had contained the most blatant mischaracterization.

Policy Implications

The GAO study produced the most reliable and complete estimates ever of the costs of testing. Moreover, it generated the most reliable and detailed database of state and school district testing programs developed to date. Yet to my knowledge, no scholar other than myself has ever used that database, which was meticulously built at taxpayer expense. The GAO study was extraordinarily well done and produced uniquely useful and trustworthy information; unfortunately, it was hounded into obscurity.

FALLACY 2: HIGH STAKES INDUCE ARTIFICIAL TEST SCORE INCREASES (TEST SCORE INFLATION)

In 1987, a West Virginia physician, John Jacob Cannell, published the results of a study in *Nationally Normed Elementary Achievement Testing in America's Public Schools*. He had been surprised that West Virginia students kept scoring "above the national average" on a national norm-referenced standardized test (NRT), given the state's low relative standing on other measures of academic performance. He surveyed the situation in other states and with other NRTs and discovered that the students in every state in the nation were "above the national average." The phenomenon was dubbed the "Lake Wobegon effect," in tribute to the mythical community of Lake Wobegon, where "all the children are above average." The Cannell report implied that half the school superintendents in the country were lying about their schools' academic achievement. It further implied that with poorer results, the other half might lie, too.

School districts could purchase NRTs off the shelf from commercial test publishers and administer them on their own. With no external proctors watching, school and district administrators were free to manipulate any and all aspects of the tests. They could look at the test items beforehand and let their teachers look at them as well. They could give the students as much time to finish as they desired. They could keep using the same form of the test year after year. They could even score the tests themselves. The results

from these internally administered tests primed many a press release (see Cannell, 1989, chap. 3).

Cannell followed up with a second report (1989), *How Public Educators Cheat on Standardized Achievement Tests*, in which he added similar state-by-state information for the secondary grades. He also provided detailed results of a survey of test security practices in the 50 states (Cannell, 1989, pp. 50–102) and printed some of the feedback he received from teachers in response to an advertisement his organization had placed in *Education Week* in spring 1989 (Cannell, 1989, chap. 3).

Case Study: The Lake Wobegon Effect

The Cannell (1987, 1989) reports attracted a flurry of research papers (and no group took to the task more vigorously than CRESST). Most researchers concurred that the Lake Wobegon effect was real—across most states, many districts, and most grade levels, more aggregate average test scores were above average than would have been expected by chance, many more.

The CRESST researchers, however, asserted that deliberate educator cheating had nothing to do with the Lake Wobegon effect. Theirs are among the most widely cited and celebrated articles in the education policy research literature. For 2 decades, CRESST members have asserted that high stakes caused the "artificial" test score gains reported by Cannell (1987, 1989) and found elsewhere. They identified "teaching to the test" (i.e., test preparation or coaching) as the direct mechanism that produces this "test score inflation."

The empirical evidence cited by CRESST researchers to support their high-stakes-cause-test-score-inflation claim is less than abundant, however, and consists of the following:

- the Lake Wobegon reports of John Jacob Cannell (1987, 1989), as they interpret them;
- certain patterns in the pre- and posttest scores from the 1st decade or so of the Title I Evaluation and Reporting System (Linn, 2000, pp. 5, 6); and
- the "preliminary findings" from an unreplicable experiment that CRESST conducted in the early 1990s in an unidentified school district, with two unidentified tests, one of which was "perceived to be high stakes" (Koretz, Linn, Dunbar, & Shepard, 1991).

Furthermore, some strikingly subjective (nonempirical) observational studies have sometimes been cited as evidence as well (see, e.g., McNeil, 2000; McNeil & Valenzuela, 2000; Smith 1991a, 1991b, 1991c; Smith & Rottenberg, 1991). How good is this evidence?

Many educators and testing opponents consider the Cannell (1987, 1989) reports alone ample proof of the "score-inflationary" effects of high-

stakes testing and propose banning such testing entirely, arguing that results from accountability tests cannot be trusted. Indeed, Cannell's data provide convincing evidence of artificial test score inflation. However, with the exception of one Texas test, none of those that Cannell analyzed had any stakes. Rather, all but one of his Lake Wobegon tests were used for system monitoring and diagnosis and carried no consequences for students or teachers.

Cannell's (1987, 1989) reports provide brief mentions of some state standards-based tests that had high stakes. Cannell contrasted their tight test security with the lax test security typical for the no-stakes NRTs he analyzed. He did not analyze the scores or trend in scores on the high-stakes standards-based tests. The Lake Wobegon tests—the tests with scores that were inflated artificially over time—were the no-stakes tests (Phelps, 2005d).

Being mostly or entirely under the control of education administrators, the NRTs could be manipulated and their resulting scores published, making the administrators look good. Cannell's (1987, 1989) data show that generally low-performing states were more prone to NRT score inflation, perhaps because administrators felt embarrassed by their states' showing on other measures and strove to compensate (Phelps, 2005d).

Because the score-inflated tests themselves had no stakes, how could states have inflated their scores? This would be possible only if the stakes attached to other tests somehow affected the administration of the NRTs. The states of Mississippi, North Carolina, and Arkansas, for example, exhibited strong score inflation with their NRTs in Cannell's (1987, 1989) data, and all three states had other testing programs with high stakes (with high levels of test security for those programs). However, Cannell's own state of West Virginia also had terribly inflated NRT scores and no high-stakes testing program. The same was true for the neighboring state of Kentucky (Phelps, 2005d).

Nonetheless, I decided to look further into the CRESST hypothesis. I surmised that if high stakes cause test score inflation, one should find the following:

- direct evidence that test coaching (i.e., teaching to the test), when isolated from other factors, increases test scores and
- grade levels that are closer to a high-stakes event (e.g., a high school graduation test) showing more test score inflation than grade levels that are further away.

This research is described in Appendix A (see http://www.apa.org/books/resources/Phelps/). In summary, the appendix indicates that the high-stakes-cause-test-score-inflation hypothesis is not supported by empirical evidence.

Why Low Stakes Are Associated With Test Score Inflation

Given current law and practice, the typical high-stakes test is virtually certain to be accompanied by item rotation, sealed packets, monitoring by external proctors, and the other test security measures itemized as necessary

by Cannell (1987, 1989) in his late-1980s appeal to clean up the rampant corruption in educational testing and reporting.

Two decades ago, Cannell (1987, 1989) suspected a combination of educator dishonesty and lax test security to be causing test score inflation. However, educators are human, and educator dishonesty (in at least some proportion of the educator population) is not going away any time soon. So if Cannell's suspicions were correct, the only sure way to prevent test score inflation would be with tight test security. In Cannell's review of 50 states and even more tests, testing programs with tight security had no apparent problems with test score inflation. High stakes are associated with reliable test results, then, because high-stakes tests are administered under conditions of tight test security. That security may not always be as tight as it could be and should be, but it is virtually certain to be much tighter than the test security that accompanies low- or no-stakes tests (i.e., when the low- or no-stakes tests impose any test security at all).

In addition to current law and professional practice, other factors that can enhance test security and that also tend to accompany high-stakes tests are high public profile, media attention, and voluntary insider (be it student, parent, or educator) surveillance and reporting of cheating. Do a Web search for stories of test cheating, and you will find that in many cases, cheating teachers were turned in by colleagues, students, or parents (see, e.g., the link to "Cheating in the News" at http://www.caveon.com).

Public attention does not induce otherwise honest educators to cheat, as CRESST claims. The public attention enables otherwise successful cheaters to be caught. In contrast to CRESST's assertions, under current law and practice, it is typically high-stakes tests that are public, transparent, and explicit in their test attributes and public objectives, and it is typically low-stakes tests that are not.

The most certain cure for test-score inflation is tight test security and ample item rotation, which are common with externally administered, high-stakes testing. An agency external to the local school district must be responsible for administering the tests under standardized, monitored, secure conditions, just as is done in hundreds of other countries (see, e.g., American Federation of Teachers, 1995; Britton, Hawkins, & Gandal, 1996; Eckstein & Noah, 1993; Phelps, 1996, 2000b, 2001). If the tests have stakes, then students, parents, teachers, and policymakers alike tend to take them seriously, and adequate resources are more likely to be invested toward ensuring test quality and security.

Any test can be made a Lake Wobegon test. All that is needed is an absence of test security and item rotation and the slightest temptation for (some) education administrators to cheat. How a test is administered determines whether it becomes a Lake Wobegon test—one with artificial score gains over time. Ultimately, the other characteristics of the test—the name, the purpose, the content, the format—are irrelevant.

In addition to good test security and ample item rotation, both of which are more common with high-stakes tests, a second, quite different type of test administration can prevent artificial test score gains (i.e., score inflation). This type produces scores that are untraceable to schools or districts. Some system-monitoring and diagnostic tests bear this characteristic. Any test producing scores that are traceable to particular schools, districts, or states can also be used to make the administrators of those institutions look good. Cannell's (1987, 1989) studies demonstrate that little incentive is required to tempt at least some education administrators to cheat on standardized tests. Successful cheating, however, requires means, motive, and opportunity. When external agencies administer a test under tight security (with ample item rotation), motivated school administrators are denied the means and opportunity to cheat, and there is no test score inflation. There were no stakes for anyone, including teachers, with (all but one of) Cannell's Lake Wobegon tests—no external evaluation or oversight. Researchers who insisted after the fact that stakes were involved simply fabricated this excuse. Education administrators cheated, or set things up so that teachers could not help but passively cheat (e.g., by giving them the same test form to use year after year), reported the fake results, and then boasted. They did so because they wanted to and, more important, because they could. The motivation was not pressure but self-aggrandizement. Indeed, the cheating was made possible by an absence of pressure.

Policy Implications

For its part, CRESST took the clear evidence of widespread educator cheating and misrepresentation of test results and managed to convince most interested parties that those educators were not responsible for their actions. Rather, the pressure of high-stakes testing was to blame, regardless of the fact that the tests in question had no stakes.

Cannell's (1987, 1989) studies showed that artificial test score gains were the result of educators' opportunistic exploitation of lax security, which happens to be more common with no-stakes testing. A reasonable policy solution would have been to legislate high levels of security for all systemwide testing programs, regardless of the stakes. CRESST suggested in the 1980s, and continues to recommend to this day, a policy solution that is the opposite of what the evidence suggests is needed. Claiming that "teaching to the test" and high-stakes testing cause score inflation, CRESST labeled these bad practice. Yet in standards-based systems, teaching to the test is exactly what teachers are supposed to do.

Further, CRESST's questionable method of verifying the validity of test score trends—comparing score trends on a no-stakes test that is not based on a particular state's curriculum to those on a high-stakes test that is—was incorporated into the NCLB Act in early 2002. The National Assessment of Education Progress (NAEP) was to be used to shadow state standards-based

tests, regardless of the fact that state standards varied widely, including in their degree of similarity to NAEP content.

FALLACY 3: THERE IS LITTLE EVIDENCE REGARDING THE EFFECTS OF TESTING AND NONE REGARDING ITS BENEFITS

The following passage from Greene and Jorgensen's (1929) authoritative handbook *The Use and Interpretation of Education Tests* struck me as representative for our times:

> Within the past score of years tests and measuring devices in nearly all subject matter fields have been developed. What the future of this movement will be no one can predict. In many ways the rapid development has been unfortunate for it has resulted in confusion on the part of the classroom teacher, the one who should profit most from the program. (p. 334)

Being thorough scholars, Greene and Jorgensen provided "a representative list" of 18 test distributors and publishers (their Appendix A), along with a 16-page "list of [several hundred] educational tests" that "is in no sense a complete catalogue of standard tests" (p. 337).

E. H. Jones (1923–1924) offered a masterful review of the research, "The Effects of Examination on the Permanence of Learning." Jones reviewed much of the experimental research literature on the optimal timing, spacing, and duration of testing for memory (i.e., permanent learning) that was available at the time. He sketched "frequency surfaces," illustrating memory "decay functions" under varying experimental conditions. Experiments included in Jones's review were conducted by many top research psychologists, including A. I. Gates, C. H. Kent, A. J. Rosanoff, Lanfan Lee Ang, and B. R. Simpson.

Harry Greene, Albert Jorgensen, and E. H. Jones are but a few of the thousands of scholars who some of today's most celebrated education researchers effectively claim never existed. Likewise for the experimental studies they cited—they simply never happened, according to some contemporary researchers. These researchers might well be offended by this casual dismissal of their life's work, but we will hear complaints from none of them—they have long since passed on.

Case Study: The National Research Council Study of Test Utility

The NRC appointed a Committee on the General Aptitude Test Battery, which wrote *Fairness in Employment Testing* (Hartigan & Wigdor, 1989), a report extraordinary in several aspects, including (a) the odd composition

of the committee; (b) the repeated insistence of the committee that there was only meager evidence for the benefits of testing, in the face of thousands of studies in personnel psychology research demonstrating those benefits; (c) the theory of the zero-sum labor market; and (d) the logical contradiction in the report's primary assertions that all jobs are unique, so general ability tests will be invalid for each, but there is no benefit from selection because any worker's abilities will be equally useful anywhere they work, no matter what their training and no matter what the field of work. This research is described in Appendix D (see http://www.apa.org/books/resources/Phelps/).

Policy Implications

It would appear that those at the NRC responsible for the evaluation of testing issues were biased and, further, that the NRC Board on Testing and Assessment had been "captured" by education interests opposed to the use of high-stakes testing. Moreover, those interests extended their control of information outside of their own field of education and over a scholarly domain of psychologists, in this case personnel (i.e., industrial–organizational) psychologists.

As for the well-validated General Aptitude Test Battery (GATB), the U.S. Department of Labor heeded the NRC's advice and chose not to allow its use in employment centers throughout the United States. In the aftermath of the NRC report, the Canadian government weighed the evidence, perceived an opportunity, and purchased the GATB for use in its employment centers nationwide—to good effect, apparently, because the GATB is used in Canada to this day for the purpose for which it was originally intended in the United States.

Case Study: The National Research Council on High-Stakes Testing

In the late 1990s, an NRC study of high-stakes testing provided a similar example of antitesting bias: *High Stakes: Testing for Tracking, Promotion, and Graduation* (Heubert & Hauser, 1999). The most revealing aspect of the NRC's 1999 report is its choice of source material. Sources that buttressed the views of the Board on Testing and Assessment were included, and hundreds of sources that did not were ignored. The majority of citations went to CRESST research and researchers.

With large resources at its disposal (a budget of more than $1 million), the NRC board minimized its research effort. On issue after issue, it threw its lot in with a single researcher or a single group of researchers. For example, the chapter on tracking is really about the work of just one person. The counterevidence and counterarguments on that issue are kept completely hidden from the reader. The early childhood, readiness testing, and promotion and retention sections of the report also feature only one person's point of view. Chapter 10 cites only three sources. Chapter 11 essentially cites only

two sources, George Madaus and Walt Haney, whose work is discussed later in the current chapter. In sum, two thirds of the citations in the report refer to fewer than a dozen research sources.

For a book on a psychometric topic, the NRC report strangely ignores psychology research. Only 10 citations of 400 come from psychology journals, and these pertain only to a discussion of assessment standards and theoretical concepts of validity. The report avoids the huge mass of accumulated empirical evidence on high-stakes selection from psychology journals. The report refers exclusively to research in education journals and reports and, even then, only to the work of a small group.

The opinions of the general public are dismissed just as effortlessly. The report acknowledges the high level of public support for high-stakes testing but discounts it thusly:

> Despite some evidence that the public would accept some of the potential tradeoffs, it seems reasonable to assume that most people are unaware of the full range of negative consequences related to . . . high-stakes test use. Moreover, it seems certain that few people are aware of limits on the information that tests provide. No survey questions, for example, have asked how much measurement error is acceptable when tests are used to make high-stakes decisions about individual students. The support for testing expressed in polls might decline if the public understood these things. (Heubert & Hauser, 1999, pp. 44–45)

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

Policy Implications

High Stakes includes more than 40 recommendations. With some exceptions, any one of them taken alone seems reasonable. Taken together, they would impose a burden on the states that none could feasibly meet. The report even floats a proposal to require that tests be pretested before they can be used for high-stakes purposes, using a new, general standard of predictive validity. Because testing proponents argue that high-stakes tests promote more learning or better employment, the NRC board argued that we should hold off certifying the use of any high-stakes test until it can be proved that over time (e.g., once a student reaches college), the test increases learning and improves employment outcomes. It would take years to conduct such an experiment, even if the experiment were feasible. Of course, it is not. One cannot test the effects of high-stakes tests when the stakes are not high as, presumably, they would not be during the life of the experiment.

The NRC's (Heubert & Hauser, 1999) *High Stakes* report was released at a propitious time: just before the debate over and design of the NCLB Act. For those who regarded the NRC's work to be objective and trustworthy, it would serve as a caution and nothing more. A century's worth of program

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

Case Study: The October Surprise of the 2000 Presidential Campaign

In 2000, a small group of CRESST researchers working at its affiliate, the Rand Corporation, decided to conduct an analysis of the Texas testing program (Klein, Hamilton, McCaffrey, & Stecher, 2000). They said it was pure coincidence that their report was released only a few weeks before the presidential election. They also said it was only coincidence that they chose to study Texas, the home state of one of the presidential candidates, rather than any number of other states with testing programs similar to the one in Texas. As Rand's James A. Thomson (2000), chief executive officer of the "scrupulously nonpartisan institution" said in a press release, "Texas was studied because the state exemplifies a national trend toward using statewide exams as a basis for high-stakes educational decisions."

The Rand report condemned the Texas Assessment of Academic Skills (TAAS) program, asserting there was no evidence of the improvements in student academic achievement the program administrators had claimed and that in fact there was considerable evidence of harm. Moreover, Rand recommended that plans for similar testing programs should be postponed until more research could be done. Rand's claims were made against the following background: Of the states that had participated in the state-level NAEP math and reading assessments in the 1990s, only one other state, North Carolina, had improved its scores more than Texas. North Carolina tested its students even more often, and for higher stakes, than Texas did.

If one simply adds up the scale-score gains (or losses) over time from the various NAEP administrations for each state, one finds the following results: North Carolina increased by 33 scale points overall, Texas by 27 points, and Connecticut by 25 points. These top three states all tested their students a lot. In the case of Connecticut, high stakes were not attached to test performance for the students, but the state education department used the test information to evaluate schools and districts in a rigorous manner (Connecticut's education department was as intrusive in local affairs as many European national education departments, its quality monitoring being as thorough and intensive). After these top three states, the cumulative scale score gains dropped to 19 in Kentucky (another state with a lot of testing) and further down to –10 in the District of Columbia, which had no testing at the time.

The Rand report's criticism of the Texas testing program rested on the following claims:

- Although there was improvement in fourth-grade NAEP mathematics scores in Texas over time, there was no improvement in eighth-grade math scores or fourth-grade reading scores.
- What improvements there were in math and reading did not last past fourth grade. Between the fourth and eighth grades, the gain in scores over time was no greater than the average for the nation.
- Because TAAS scores improved by a greater proportion than Texas' NAEP scores, the TAAS scores must have been "inflated" and not reflective of "real" gains in achievement.

Texas' net cumulative score gains on the NAEP were more than twice the national average. The Rand researchers claimed the state's gains were no different from the rest of the nation's by separating the big picture (all the grade levels tested and compared across the entire time period) into several smaller pictures (each grade level tested separately and compared only with the nearest time period) and then relying on statistical-testing artifacts within each. This was methodologically invalid, because they reported a conclusion about the big picture without actually conducting a statistical test on it. Most researchers try to increase the size of their data sets and thus the power of their statistical tests; Rand did just the opposite. Instead, it looked at a segment of gains in fourth-grade math, a segment of gains in eighth-grade math, and so on. With each segment, the researchers conducted a statistical test that relied on arguably standard, but still arbitrary, cutoff thresholds to determine "statistically significant" differences. For each separate case in isolation, there is nothing wrong with this. The Rand researchers probably noticed, however, that for the segments in which the Texas gains did not reach the cutoff points, they just barely did not make it. The Texas gains in the case of every segment were large by normal standards of "large," just not large enough in each and every segment to make the cutoff point for the statistical test Rand chose to use in each case.

If one combines the various segments (in statistical jargon, this is called *pooling*), however, one can both increase the statistical power of the test (by increasing the sample size) and conduct the correct test—for the NAEP performance of Texas as a whole rather than for separate, discrete bits. Combining separate tests, or subtests, at the same level of difficulty, even on different subject matter, is often done when identical scales are used. Witness the many studies that use SAT combined (verbal + math) scores in their analyses.

The Rand researchers also argued that because the score gains on the TAAS exceeded those on the NAEP, they could not be "real" and must have been inflated. They used the same logic they had used in the CRESST Lake Wobegon study discussed earlier. Again, scores on two tests cannot be per-

fectly correlated without their being the exact same test. The TAAS and the NAEP were not the same, nor were they supposed to be, so their scores could not be perfectly correlated. The fact that the score increases in the TAAS over time were greater than Texas students' score gains on the NAEP was to be expected; any other result would have suggested a serious problem. The TAAS contained subject matter that matched the curriculum standards of the state of Texas. The NAEP did not. Teachers were supposed to teach the subject matter covered by the TAAS, not that covered by the NAEP.

Policy Implications

When accurate information was most needed, we got this instead. In late 2001, midway between the election of George W. Bush and the U.S. Congress' passage of the NCLB Act, I read the following statements in an education journal:

> Nearly 20 years later, the debate surrounding [minimum competency tests] remains much the same, consisting primarily of opinion and speculation. . . . A lack of solid empirical research has allowed the controversy to continue unchecked by evidence or experience. . . . The lack of empirical research on the achievement effects of mandatory graduation exams is striking, particularly in light of their growing popularity across the nation. . . . The evidence on graduation exams and achievement is limited and mixed. (Jacob, 2001, p. 334)

I assumed that an opponent of President Bush's policies and of the NCLB Act had written it. I wrote to the author, then employed as an instructor at one of my alma maters, and pointed out that there was, in fact, a great deal of empirical research on minimum competency testing and on the achievement effects of mandatory graduation exams and that the empirical evidence on graduation exams and achievement was neither limited nor mixed. The author defended his statements by asserting that he had checked them with our country's foremost researchers on the effects of standardized educational achievement testing, naming four CRESST researchers, including three of the authors of the October surprise report. The joke was on the Republican Party, however. This fellow, whose carelessness undermined GOP education policy, became a trusted policy advisor.

Case Study: Co-Optation of the National Council on Measurement in Education

The National Council on Measurement in Education (NCME) had long served as a bulwark of psychometric respectability against the more regressive elements within education research that dominate the much larger American Educational Research Association (AERA). However, most NCME members are also AERA members; indeed, the two organizations hold their

annual meetings at the same time and the same place. Moreover, most NCME members work in the education business, as professors in education schools or for firms that serve the education market. NCME is not impervious to regressive influences

The psychometrician William Mehrens (1998) delivered his presidential address to the NCME in the mid-1990s on the topic of the effects of standardized testing. He talked as if he had looked at the sum total of what is known about these effects. The picture he saw did not look pretty: Standardized testing, particularly when it had high stakes, seemed to do more harm than good. Unfortunately, Mehrens had apparently conducted no literature search whatsoever. Rather, he relied on his own instincts and assumed that what he happened to read over the years comprised a representative sample of the existing literature.

Among more than 60 citations in Mehrens's (1998) speech, one can find none from psychologists, program evaluators, sociologists, economists, or researchers outside the United States and only one dated earlier than 1984. Most of his attention (and more than one third of his citations), in fact, focused on the work of a single organization, CRESST. Half of the rest of the sources were outspoken, self-acknowledged opponents of high-stakes testing. Sadly, in the decade since Mehrens' address, testing opponents have eagerly and liberally cited his presidential address as authoritative evidence that the research literature on the effects of testing is meager at best.

As if to place a capstone atop the monument to CRESST's success in information suppression, researchers Dan Koretz and Laura Hamilton (2007) penned a chapter for the most current version of the "bible of testing research," the NCME-sponsored reference book, *Educational Measurement*. Their chapter, "Testing for Accountability in K–12," is more remarkable for what is left out than for what is included. Little of the abundant counterevidence to their work is mentioned and, when offered at all, is the weakest available. Among the 253 references are 82 (32%) to works by themselves and their CRESST colleagues. Another 10% are official works—legislation, government reports, statistical compendia, and the like. This still leaves a good bit of room for references to a century's worth of research on the effects of testing. Instead, one finds other sources that claim a research dearth and the plaint that much more research will be needed before we can dare to use tests responsibly.

In an appendix to *Defending Standardized Testing* (Phelps, 2005a), I listed more than 300 studies of the effects of testing, the vast majority of which provide empirical evidence of beneficial effects. Eighty of the listed studies are meta-analyses or reviews of multiple separate studies. I compared this list with the references in the Koretz and Hamilton (2007) chapter, which purports to summarize all of the available research. I found but three sources in common.

Those who claim a paucity of research on the effects and benefits of testing or on the structure of test-based accountability systems either have not looked very hard or have not wished to find what is available in the literature. If the topic of the effects of standardized testing can be persistently exposed to the glare of journalists' floodlights yet successfully censored and suppressed, then any topic can be. Is the increasing concentration of education-research dissemination in fewer and fewer hands likely to improve education? It may not matter how one answers the question, for the forces working to dissolve and disintegrate the hard-won accumulation of education knowledge seem only to be growing stronger. The chief problem is that accurate and useful ideas and information in education—indeed, perhaps most accurate and useful ideas and information—are suppressed and ignored in policy discussions.

Case Study: Co-Optation of the Republican Policy Advisors

The following statements come from CRESST researchers:

- Despite the long history of assessment-based accountability, hard evidence about its effects is surprisingly sparse, and the little evidence that is available is not encouraging. (Koretz, 1996)
- Although much has been written on achievement motivation per se, there has been surprisingly little empirical research on the effects of different motivation conditions on test performance. (Kiplinger & Linn, 1993, p. 3)

Several years ago, I spent some time conducting computer searches and strolling library aisles for signs of the research literature on test-based accountability and the relationship between motivation and test performance that CRESST researchers have repeatedly declared either nonexistent or scarce. Lo and behold, I discovered a few hundred studies. My search was tedious, but it was not difficult. Given the height of the pile of books, articles, and bibliographies I have yet to comb through, it would appear that I will discover a few hundred more.

Studies measuring the effects of standardized testing in education date back to the early 1900s and range across virtually all relevant types of research methodologies—meta-analyses, controlled experiments, quasi-experiments, program evaluations, case studies and structured interviews, interrupted time series with shadow measures, pre–post designs, polls and surveys, cost–benefit analyses, multivariate regressions, multilevel structural equation models, and data analyses of administrative records (Phelps, 2005c).

CRESST researchers have probably been the most persistent in their paucity- and absence-of-research claims, but they have hardly been alone.

This assertion is widely advertised (see, e.g., Barth, 2006; Mitchell, 2006; Olson, 2002). Moreover, the belief now seems to transcend political and ideological boundaries. Both opponents and supporters of high-stakes standardized testing assert the claim (see, e.g., Cizek, 2001; Hanushek & Raymond, 2002, 2003; Jacob, 2001, 2002, 2003; Loveless, as cited in New Report Confirms, 2003; Roderick, Jacob, & Bryk, 2002). As the belief in the research literature's nonexistence has become more pervasive and deeply held, efforts to reference it have become less frequent, less thorough, or casually dismissed.

One might have reasonably assumed, given the thrust of U.S. education policy in the early 2000s, that this research literature would have been exposed, made widely familiar, and meticulously analyzed. Yet just the opposite happened—the bulk of an available research literature that could have helped to guide our society in the implementation of its primary, and controversial, education policy was declared nonexistent.

For example, consider the following characterizations of the research literature on the effects of test-based accountability. "Most of the evidence is unpublished at this point" (Olson, 2002, p. 13). "There is little empirical evidence on test-based accountability (also referred to as high-stakes testing)" (Jacob, 2002, p. 2). It is "a young and highly selective body of work" (Hanushek & Raymond, 2002, p. 1). "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" (Loveless, as cited in New Report Confirms, 2003, p. 1). "The evidence on this outcome is just beginning to come in" and "the evidence on positive and negative consequences is necessarily skimpy" (Cizek, 2001, p. 7). These quotes come from the period during which the NCLB Act was being considered or designed. All come from Republican policy advisors.

The presidential election campaign of 2000 was the first in U.S. history in which standardized testing was a central campaign issue; testing opponents were prominent and vocal throughout. What did the Republican policy advisors have to say? They declared there to be no evidence that high-stakes standardized testing did any good and that on balance, it seemed to be harmful. More research was needed on the topic (and these advisors were willing to do it). NCLB opponents could not have written the script any better themselves.

Policy Implications

With the election of George W. Bush, GOP policy advisors faced a historic opportunity, with enormous implications, to benefit U.S. education. They had the resources to blast open the seal of censorship covering a huge research literature on standardized testing's achievement effects. Instead, whether by mistake or design, they chose to reinforce the seal, despite the critical need of their Republican politician clients for exactly the opposite behavior.

After seconding testing opponents' claim that little to no research existed on the effects of standardized testing, some of the Republican think tankers declared themselves to be pioneers in conducting such research. Apparently, one assertion used to persuade Republican policymakers of the paucity-of-research hypothesis was that all education research was poorly done, so any research they (mostly economists) conducted would be the first high-quality research on the topic (see, e.g., Hanushek, 2006).

Think of the assumptions necessary for economists to adopt this line of thinking. Educational standards and standardized tests have existed for millennia. Psychologists first developed the "scientific" standardized test more than a century ago, and they, along with program evaluators and education practitioners, have conducted hundreds of thousands of studies with or about them since that time. Nonetheless, since the mid-1990s, a number of economists have proposed that none among these scores of psychologists and practitioners ever thought to study the various effects of educational testing (and that these professionals could not even conduct such studies responsibly). In their book chapter titled "Economics Wins, Psychology Loses, and Society Pays," Bazerman and Malhotra (2006) described several miserably failed public policies, crafted by economists, on topics for which psychologists had long developed expertise—expertise ignored by the economists in policy positions. With federal standardized testing programs in the first decade of the 21st century, economists may have added yet another such policy.

Being the first to conduct research on a topic can of course enhance one's career prospects fabulously, whereas citing and summarizing work already done by others can make one look like a slacker. To my knowledge, none of the cases cited earlier with erroneous "firstness" claims ultimately attracted any negative consequence toward the claimant. The act of dismissing a research literature, no matter how large, appears to be risk free. What consequence might that lack of consequence portend for the preservation of the education research literature?

Besides, who's going to complain about the regression? Most of the scientists and evaluators responsible for the past century's worth of research on the effects of testing are deceased. Many of those still alive work in the field of psychology and, even if they were to become aware of others' grandiose claims to pioneering the field, they have no standing in the other professions (e.g., economics) from which to lodge a complaint. Those working within education who might object have long since been professionally marginalized by the regressors. That leaves few to carry on as advocates for the preservation of an increasingly endangered research literature.

I have personally challenged some of the research regressors directly on their firstness and paucity-of-research claims. The typical reply insinuates something like, "You're just jealous because your work is not getting attention." Incidentally, I hear this retort whether my own work represents less than 1% of the relevant dismissed research literature or none of it.

One slightly more thoughtful reply suggests that declaring a century-old research literature nonexistent is of no practical consequence because people are conducting these studies now and surely will discover whatever previous researchers discovered anyway. Perhaps, but it might take them another hundred years and another thousand studies to accumulate as much knowledge. Besides, the knowledge was needed during the period 2000–2002, while the NCLB Act was being debated and designed. The few economists' studies conducted after 2003 were too late to be useful.

The co-optation of Republican policy advisors was hugely important. Not only did testing critics, including CRESST researchers, gain policy leverage, they were appointed to influential executive branch committees, commissions, and projects. Not bad for folk who tried to sabotage the Bush campaign just 3 weeks before the 2000 presidential election. Moreover, with the Republican think tankers on board, the most influential potential advocates for educational achievement testing programs were neutralized. The enormous resources of the federal government, business groups, and foundations that could have been organized to halt the research regression instead supported it. Finally, because most education journalists grant the small cadre of Republican think tankers an effective monopoly to represent the "other side" of most education issues, Republican acquiescence in snuffing a vibrant research literature represented for journalists unanimous confirmation of a fact: The research literature did not exist.

FALLACY 4: TESTING IS MISCHARACTERIZED BECAUSE IT IS DIFFICULT TO UNDERSTAND

Testing and measurement experts—*psychometricians*—tend to be smart people. Most have doctorates, which they earned after completing a dozen or more courses on abstruse topics in statistics, computer programming, scaling, equating, item response theory, and other technical exotica.

Some have assumed that the recondite nature of the subject matter is to blame both for the animosity felt by many education advocates toward testing (i.e., one fears what one does not understand) as well as the misrepresentation of the topic (i.e., it is simply too technical a topic for the average educator, or education journalist, to understand). I believe that fear of the unknown may well fuel the animosity, but I do not believe that the pervasive misrepresentation of the topic among education advocates can be explained by misunderstanding. By ideology, perhaps. By professional self-interest, perhaps. But not by ignorance.

The purpose of this section is to demonstrate how easy it is for advocates to repress the simplest, most easily verified facts and convince the public of falsehoods. This ease is illustrated with arguably the easiest to understand bit of information in the testing and measurement field.

Of the many testing controversies, counting the number of tests given may be the most trivial. Yet that is the point. Of all the claims about educational achievement testing, this is the easiest to verify, and the argument over this issue should be the easiest to settle. Instead, most citizens and policymakers remain misinformed on this presumably simple topic.

Case Study: Facts Unprecedented, Unparalleled, Indisputable

In an *Education Week* editorial titled "Standardized Testing and Its Victims," Alfie Kohn (2000) wrote:

> Standardized testing has swelled and mutated, like a creature in one of those old horror movies, to the point that it now threatens to swallow our schools whole. But let's put aside metaphors and even opinions for a moment so that we can review some indisputable facts on the subject.
>
> Fact 1: Our children are tested to an extent that is unprecedented in our history and unparalleled anywhere else in the world. . . . Few countries use standardized tests for children below high school age—or multiple-choice tests for students of any age.
>
> Fact 2: Our children are tested to an extent that is unprecedented in our history and unparalleled anywhere else in the world. (p. 60)

Similarly, in a September 2001 *Frontline* interview on the Public Broadcasting System, National Public Radio journalist John Merrow interviewed the U.S. Secretary of Education, Rod Paige (Interview: Rod Paige, 2001). In reference to the new NCLB program, he asked, "It raises the question of too much testing. American kids are already tested more often than kids in any other industrialized country. Are we testing our kids too much?" (¶ 22).

The original source of the "indisputable fact" that U.S. schools test more than do other countries' schools was a contractor report for the now-defunct U.S. Office of Technology Assessment (OTA) written by education professors George Madaus and Thomas Kellaghan (1991). Their report became a chapter in a longer OTA report (OTA, 1992, pp. 135–164) and later an article in a professional journal (Feuer & Fulton, 1994). Their claims have permeated the media and the education research literature (e.g., Kellaghan & Madaus, 1995; Kellaghan, Madaus, & Raczek, 1996; Madaus, 1991a, 1991b; Medina & Neill, 1990; National Commission on Testing and Public Policy, 1990; Neill, 1992; Rothman, 1995; Sacks, 1999; Viadero, 1994).

The aforementioned authors claimed all of the following:

- "American students are already the most heavily tested in the world." (Madaus, 1991a, 2)
- The trend in other developed countries was toward less standardized testing. (OTA, 1992, p. 144)

- The trend in other developed countries encompassed all levels of education "even at the postsecondary level." (OTA, 1992, p. 143)
- The trend in other developed countries was unidirectional—large-scale, external tests were being "abolished" (OTA, 1992, p. 143).
- External examinations in other developed countries were "no longer used to make decisions about students' educational paths during the period of compulsory education" (Kellaghan et al., 1996, p. 59).
- "Standardized national examinations before age 16 have all but disappeared from Europe and Asia" (OTA, 1992, pp. 135, 144).
- "The United States is unique in the extensive use of standardized tests for young children" (OTA, 1992, p. 135).
- "None of the countries studied by OTA has a single, centrally prescribed examination that is used for all purposes—classroom diagnosis, selection, and school accountability. Most examinations overseas are . . . not used for school or system accountability" (OTA, 1992, p. 135).

All these claims were false and unsupported by empirical data, but they nonetheless infused policy debates surrounding three presidential testing proposals from the early 1990s to the early 2000s—George H. W. Bush's American Achievement Tests, Bill Clinton's Voluntary National Tests, and George W. Bush's NCLB Act. The report's persuasive force relied on a single, plausible rationale: The authors argued that other countries were dropping large-scale external tests because they no longer needed them as selection devices. European and Asian countries had expanded the number of places available in secondary schools, polytechnics, and universities, so access had been made available to all, or at least most, who desired it. This proposed rationale for other countries' allegedly dropping educational selection tests served to distract from the weakness of the study's claims. One of the coauthors would go on to manage more than a decade's worth of research on standardized testing at the NRC; another would end up conducting research on standardized testing for the World Bank.

Around the same period of time, four other, more rigorous test-counting studies were conducted that went far beyond rationale and anecdote and were generally ignored. First, the Organisation for Economic Co-operation and Development (OECD) conducted a survey of its member countries in 1990–1991 addressing the number and duration of their systemwide tests. This survey revealed that U.S. students faced fewer hours and fewer numbers of high-stakes standardized tests than their counterparts in every one of the 13 other countries studied. Further, U.S. students underwent fewer hours of

mandated tests than their counterparts in 12 of the 13 other countries (Phelps, 1996, p. 25).

Second, Eckstein and Noah's (1993, pp. 149, 167) classic eight-country set of case studies ranked the United States lowest both in "examination burden" and "examination difficulty." The authors concluded,

> In addition to certification and selection, other countries use their end-of-secondary-school examinations for a variety of other functions: for example, to define what knowledge and skills are of most worth, to set performance expectations of students, teachers, and schools, and to provide yardsticks against which individual schools and the school system as a whole can be assessed.
>
> The United States . . . lacks any systematic and general way of certifying completion of a specified course of secondary school study and, unlike other countries, has no consistent national criteria or means for selection beyond that stage, whether for employment or for particular types of postsecondary education or training. (pp. 238–239)

Third, in a seven-country survey of secondary school math and science examinations, Britton, Hawkins, and Gandal (1996) asserted,

> While only 6.6 percent of US students take Advanced Placement (AP) examinations, roughly a quarter to a half of all students in other nations take and pass advanced subject-specific examinations.
>
> In each country except the United States, college-bound students seeking to study in a university must pass demanding, subject-specific examinations. In France, Germany, and Israel, even many students who do not go on to college take these examinations because they are a prestigious credential in their societies. (pp. 202–203)

Finally, in a review of widely available documentary source material, I found that over the period 1974–1999 in 31 countries and provinces, 59 large-scale external testing programs were added, and only 5 were dropped. The OTA (1992) had implied that large-scale tests were used only for selection to or exclusion from the next level of education. During 1974–1999, however, other countries and states added 22 monitoring exams, 6 subject-area end-of-course testing programs, 2 primary-to-secondary-level achievement tests, and 2 diagnostic exams. Thirty tests with medium or high stakes were added, and only four were dropped (Phelps, 2000b, pp. 17, 18).

All the countries mentioned in the anecdotal Madaus–Kellaghan–Feuer study (i.e., Feuer & Fulton, 1994; Madaus & Kellaghan, 1991; OTA, 1992) were covered by these four more rigorous studies that presented empirical and case study evidence that U.S. students faced fewer and easier tests than did their counterparts in other nations. None of the latter studies received much attention in the education media or research literature, however, whereas the Madaus–Kellaghan–Feuer study received a lot.

Policy Implications

Why did a study void of evidence receive more attention than studies that incorporated data and first-person case studies? In my judgment, it is because education advocates can promote studies that reach conclusions they like and suppress studies reaching conclusions they do not like, and this can be based on ideology or self-interest.

In such an atmosphere of information dissemination, any topic needs only one study to promote a favored conclusion, and that one study only needs some trappings of legitimacy. Whether the study was done well or the conclusions were warranted may be irrelevant. In the case of the Madaus–Kellaghan–Feuer study, the OTA imprimatur was sufficient. The four other studies reaching contrary conclusions were simply ignored and disappeared from view.

In this manner, public policies "based on research" can amount to nothing more than a powerful interest group's preferences—the factual basis being that which the interest group wishes to be fact. Public policies formed this way will always serve the needs of the vested interests.

For their part, education journalists had plenty of opportunity to study the issue and apply a standard amount of journalistic skepticism to the more extravagant claims about the amount of testing in the United States alone or in comparison with other countries. However, I have yet to witness a journalist referring to the grandiose test-counting claims as anything but reliable and objective fact (see, e.g., Chandler, 1999; Merrow, 2002; Sacks, 1999; Strauss, 2006; Teel, 2001; Wolk, 2002). More on education journalists' profound lack of skepticism regarding claims made by testing opponents can be found in another source (see Phelps, 2003, chap. 6).

Case Study: The Amount of Testing Is Unknowable, and It Is a Huge Number

Madaus would serve as coauthor in another set of claims about the amount of testing that occurs, this time for the United States alone (Haney, Madaus, & Lyons, 1993):

> We contacted the College Board and ACT directly and were informed that 1,980,000 SATs and 1,000,000 ACTs were given in 1986–87. We thus have relatively firm figures on the number of such college admissions tests given. But there are several ways of counting the number of separately scorable subtests in these testing programs. The SAT has two subtests, the SAT—Verbal and the SAT—Math. Moreover, two subscores are reported for the SAT—Verbal, namely reading and vocabulary. Also, almost all students who take the SAT take the Test of Standard Written English. . . . Similarly, the ACT assessment has four subtests, but since a composite score is also calculated, we have used 4 and 5 as bases for high and low estimates. The results . . . indicate that

> between nearly 4 million and 10 million SAT subtests and 4 million to 5 million ACT subtests are administered annually. . . . Altogether then we estimate that in 1986–87, 13 million to 22 million college admissions "tests" were administered. (pp. 65–66)

The passage continues to sum the total of all U.S. standardized student tests—and not just college entrance exams, but the quotation marks around the word "tests" disappears, *et voilà*, all parts of tests become whole tests.

> In sum then . . . we estimate that between 143 million and 395 million tests are administered annually to the nation's population of roughly 44 million elementary and secondary school students, equivalent to between 3 and 9 standardized tests for each student enrolled in elementary and secondary schools. (Haney et al., 1993, p. 66)

At the beginning of the passage, a test is called a *test*. In the middle, the reader is told that tests have parts. Those separate parts are counted up, and in the next paragraph, the parts are called *tests*. After this semantic magic is complete, the authors assert that there are from 3 to 9 times as many standardized student tests administered annually as there actually are.

Another oddity of the passage is its use of the word *estimates*. Two telephone calls to the SAT and ACT offices provided exact counts of the numbers of their tests administered. Further, at the beginning of the passage, an ACT test is referred to in the singular and the total annual number of ACTs administered is declared to be 1 million. After the authors do their parts-as-wholes counting, they end up with an "estimate" for the annual number of ACTs of from 4 to 5 million. Four million is their "lower bound estimate" for a number of tests they had just claimed for a fact to be only 1 million.

One education professor finessed the issue this way:

> No one knows for certain how many students are tested in a given year or how many tests the typical student takes because comprehensive and unambiguous data are not available. Richard Phelps (1997) estimates that 36 million district-wide and statewide tests are given each year in the United States. Peter Sacks (1999) cites estimates of 127 million standardized tests of all types being given in a year. Walter Haney, George Madaus, and Robert Lyons (1993) provide low and high estimates. . . . On the low end, they estimate that slightly more than 143 million students a year were tested and that the average child took 2.7 tests per year. On the high end, they estimate that just over 395 million students a year were tested and that the average child took 5.4 tests per year. (Snowman, 2002, p. 490)

The 36 million mentioned in the foregoing passage was an accurate estimate of the number of systemwide tests administered in U.S. public schools in the mid-1990s. This translates to less than 1 test and less than half a day of testing per year per student. All the other numbers mentioned in the passage

were derived from the work of Haney, Madaus, and Lyons (1993), count subtests rather than tests, and double count some tests (Phelps, 1997):

Counting the number of tests, or even the number of test items, administered in any given year in the United States may be tedious, but it is easy. It is almost as easy to count the amount of time devoted to test administration (Phelps, 1997). Even if one chooses to count tests by more than one method—for example, by the number of test forms or the number of subtests—it is a remarkably easy task. Verification of the numbers requires no complex theory or special analytical skills, just some persistence and a mastery of simple arithmetic.

These days, most educational achievement tests are administered either nationwide or statewide. Those administered nationwide include the NAEP, the ACT, and the SAT. Their annual numbers of individual test (and subtest) administrations, as well as their durations and the fees paid, can be found in the respective organizations' publications or by telephoning them. Likewise with each of the 50 states, tests may be developed by private firms, but they are sponsored and typically administered by state education agencies.

Policy Implications

If even this, the most resolvable of standardized testing disputes, cannot be resolved or, rather, clarified with accurate information, how can any of the other, more complex issues be resolved? If even these, the most easily obtained facts about standardized testing, can be so effortlessly muddled and manipulated, how can any facts about educational testing be clearly communicated to and understood by the public?

Perhaps they simply cannot be. The fog of ambiguity and misinformation does not dissipate as one climbs the ladder of topical complexity. Other controversial issues, both simple and complex, are just as muddled, and the debate just as one sided. In the case of at least several issues, the factual information commonly accepted by most educators and journalists is not only erroneous but demonstrably the opposite of reality.

OVERARCHING POLICY IMPLICATIONS

To my observation, there is a clear quality difference between research conducted on educational achievement testing policy-related topics before the mid-1980s and the more prominent such work conducted after the mid-1980s. Most of the earlier work can be found in the scholarly psychology literature, although a good number of studies were conducted in the field by technically trained educational practitioners as well. Generally, the work is typical of open-minded scientific inquiry. All sides of issues and most previous relevant work seems to have been considered. Over the years, more knowledge accu-

mulated and was built on top of what had previously been learned. There was progress in humanity's understanding of educational achievement testing.

Objective, open-minded, scholarly study of policy-related educational achievement testing topics continued beyond the mid-1980s but has largely been marginalized from public discussion. The work that seems to get the attention of both policymakers and the press is that conducted by prominent education professors, researchers in federally funded research centers, and several think tank–based economists or political scientists. They seldom have acknowledged the earlier work on educational achievement testing or any contemporary work conducted outside their own small circles of colleagues. Often they have claimed that there has been no study of particular effects of achievement testing before the start of their own work.

Some widely cited educational achievement testing studies have been rigged so that only negative results were possible; for example, see the survey instruments used by West and Viator for the National Science Foundation (Phelps, 2005b, pp. 19–20), the survey result interpretations of the ASCD (Phelps, 2005b, pp. 18–19), or the range of sources cited by the aforementioned Mehrens (1998) and NRC (Heubert & Hauser, 1999) studies. Still other widely cited studies produced results revealing positive effects from the use of educational achievement tests that were nonetheless presented by their authors as negative results (e.g., see Corbett & Wilson, 1991; B. D. Jones & Egley, 2003; B. D. Jones & Johnston, 2002; Kellaghan, Madaus, & Airasian, 1982). These efforts at information suppression, irrespective of their motivation, have been largely successful. Prominent education researchers have managed to delete large segments of the education research literature from the collective working memory and hide large amounts of information that could have informed U.S. education policy (Phelps, 2007a).

As if the suppression of information within the field of education were not detrimental enough, some of the same researchers have managed to expand their information removal activities into other fields, such as personnel psychology.

ONE HUNDRED YEARS OF RESEARCH AND EXPERIENCE LEFT BEHIND

Imagine this scenario. After a new executive administration is elected and installed in Washington, a deadly epidemic spreads across the country. The administration's policy advisors declare that medications and supplies that could save lives do not exist, but, for a fee, they will build up a new stock. None of the policy advisors are epidemiologists, however; rather, they are cardiologists and orthopedic surgeons. The medical crisis runs its course, and many die because the new stock of supplies only becomes available after the crisis has passed. Much later, it is revealed that there had been a large

stock of medical supplies available at the time of the outbreak, but the policy advisors had simply accepted someone else's word that it did not exist and had not looked for it.

An overly dramatic story, perhaps, but subtract the dead bodies, and it is essentially what happened during the period 2000–2002 while the NCLB Act was being designed and debated. A century's worth of useful experience and research was declared nonexistent and not even considered by the Bush administration's policy advisors. Instead, they offered to start a research literature from scratch, and their offer was accepted.

Few economists and political scientists had demonstrated an interest in educational achievement testing during the 20th century; the standardized test is the psychologist's invention. Yet virtually all the Bush administration's education policy advisors are economists and political scientists. Their small collection of research studies on the effects of testing has dribbled out since 2002, with some of those responsible declaring themselves research pioneers (see, e.g., Hanushek, 2006).

The jurist Richard A. Posner (2001) warned of the societal dangers of celebrity researchers, whom he labeled "public intellectuals," making claims outside their field where those with the requisite expertise can hold them accountable. Unfortunately, when the Bush administration's education policy advisors and think tank consultants have published naive research in economics journals, they have not been reviewed by scholars familiar with the relevant literature. Their work might never have passed muster with psychology and measurement journals. Examples of such naiveté include beliefs and assumptions that all tests are pretty much the same and validly comparable whether they have stakes, are administered securely, apply stakes to the teachers or the students, or are norm referenced or standards based.

Ironically, these faux pioneers were paid out of taxpayer funds—first, to replicate studies for which taxpayers had already paid, and second, to declare the earlier public investments nonexistent. The NCLB Act could have been informed by a cornucopia of research and experience. Instead, it was informed by virtually none. Prior research and experience would have told policymakers that most of the motivational benefits of standardized tests required consequences for the students, not just for the schools. Those stakes need not be high to be effective, but there must be some. Because NCLB imposes stakes on schools but not on students, who knows whether the students even try to perform well.

Prior research and experience would have informed policymakers that educators are intelligent people who respond to incentives and who will game a system if they are given an opportunity to do so (see, e.g., Cannell, 1987, 1989). The NCLB Act left many aspects of the test administration process that profoundly affect scores (e.g., incentives and motivation, cut scores, degree of curricular alignment) up for grabs and open to manipulation by local and state officials.

Prior research and experience would have informed policymakers that different tests get different results, and one should not expect average scores from different tests to rise and fall in unison over time (as some interpreters of the NCLB Act seem to expect with the NAEP benchmark). Prior research and experience would have informed policymakers that the public was not in favor of punishing poorly performing schools (as NCLB does), but was in favor of applying consequences to poorly performing students and teachers (which NCLB does not; see, e.g., Phelps 2005b).

What is the effect of test-based accountability? Appendix 3.1 lists a small sample of useful, insightful, relevant, well-done studies that effectively answer this question, could have informed the design of NCLB, and have been declared by prominent researchers not to exist.

One could not find these studies mentioned in the "what the research says" education policy advice pages of major, policy-influential organizations (e.g., Education Commission of the States, *Education Week*, the Education Writers Association, the National School Board Association's Center for Public Education) in the early 2000s while the NCLB Act was being considered and then implemented. In their place, one found mention of a paltry number of works from contemporary "Influentials" (to borrow a term popular with *Education Week*). Had the policymakers and planners involved in designing the NCLB Act simply read the freely available research literature instead of funding expensive new studies and waiting for their few results, they would have received more value for their dollars as well as more and better information, and they would have had this information earlier, when they actually needed it.

One person directly involved in research regression activities told me that there had been no research before the early 2000s that had been conducted under the exact conditions specified by the NCLB Act. However, I am unaware of research studies that focused exclusively on the effects of minimum-competency tests on left-handed students, on Tuesdays, in the month of February, and on rainy days. Does this mean, then, that we cannot make any assumptions about the effects of any tests if they are minimum-competency tests, if left-handed students are involved, and if the tests are administered on rainy February Tuesdays? To the contrary, we can, because we know which factors matter and which do not, and all of the factors that matter have been studied many times.

With the single exception of the federal mandate, there was no aspect of the NCLB accountability initiative that had not been tried and studied before. Every one of the NCLB Act's failings was perfectly predictable on the basis of decades of prior experience and research. Moreover, there were better alternatives for every characteristic of the program that had also been tried and studied thoroughly by researchers in psychology, education, and program evaluation. Yet policymakers were made aware of none of then.

CONCLUSION

Understanding of educational achievement testing may be shrinking among the public at large. The technical psychometric research literature would seem to be safe from the censorship and suppression of vested interests, but the research literature related to testing policy (i.e., its administration, program structure, use, extent, effects, cost, benefits, public opinion, and research dissemination) is diminishing. There are simply too few who cite the research literature in any substantial depth or breadth and too many who are eager to declare it barren. At the same time, there seems to be little hesitation on the part of many researchers to skip lightly past the annoying obligation of a search of the literature yet nonetheless to claim a mastery of it. A thorough literature review requires a great deal of time and patience, virtues often lacking among the most ambitious and narcissistic. In one effort of mine—accumulating studies on the impact of standardized testing—I started out thinking that there were probably a dozen or so such studies. A few years ago, I was aware that there were hundreds. Now I know that their number exceeds a thousand, and, despite the rhetoric of some critics, only a tiny proportion of them were conducted after the year 2000.

In the end, however, it will not matter for society's sake if we find 10,000 studies. There will remain other education researchers, prominent and with abundant resources at their disposal—researchers whose work is frequently covered by education journalists—who will continue to insist that no such studies ever existed. It is education research's dirty secret: Unpopular research and research that generates unpopular results can be successfully—and easily—censored and suppressed (see, e.g., Phelps, 1999, 2000a, 2003 [preface, chap. 7], 2005c [chap. 3]).

A biological species cannot survive when mating individuals cannot find each other. When numbers decline to such an extent that predators (or hunters) can more easily find members of a species than can potential mates, the species crosses a demographic threshold and heads toward its inevitable extinction. Those who work with endangered species call this the *extinction vortex*.

Similarly, the censorship and suppression of the research literature on the effects of educational achievement testing have become so successful that it has become difficult to find the literature's progenitors. For example, I may have spent more time than anyone else on Earth combing the research literature. Nonetheless, I was a few years into my effort before I discovered the work of Frank Dempster (1991, 1997), one of the world's foremost authorities, or that of Jim Haynie (e.g., 1994, 2007) in career and technical education. Why did it take me so long to find their work? It is not popular among the vested interests in education—they find the benefits of testing to be strong and persistent—and thus it is not widely advertised. Indeed, work

like theirs is far more often declared nonexistent than recognized. Few miners persist in tunneling deep shafts for unfashionable gems. Likewise, few researchers pursue unfashionable topics in the face of persistent discouragement and very real professional disincentives (see, e.g., Phelps, n.d.).

U.S. education's single greatest need may be for an independent education press. Unfortunately, what we have now is anything but independent. Vested interest organizations are guaranteed seats on the board of the Education Writers Association, the only large professional organization of education journalists (Lieberman, 2007, chap. 11). The Goliath of education news publications, *Education Week*, is arguably the least independent of any education organization inside the (Washington, DC) Beltway. *Education Week* advertises its willingness to partner with other organizations on research and news projects. Its partners must bring resources to the table, however, and only those with power and money can do that. (Imagine the *New York Times* or the *Washington Post* entering into working research, news, and dissemination partnerships with think tanks, federally funded research centers, and professional advocacy groups.) *Education Week* editors serve on the boards of partisan organizations for which they provide headlines, thus freely participating in education's "interlocking directorates" of vested interests (see Domhoff, 2006).

Not surprisingly, *Education Week*'s pages often read like a wintry mix of PR Newswire and *Variety*, more focused on celebrity and influence than substance and accuracy. It spends ample resources ranking public intellectuals in its periodic popularity contest of the "Influentials" while it ignores an abundance of important information and evidence that could be provided by the many it deems not influential. Its house blogger has openly solicited votes ranking those same celebrities on their physical appearance. Alexander Russo's "Hot for Education" series clarified the purpose of contemporary American education journalism: It has nothing to do with truth, justice, or progress and everything to do with seeing and being seen, rubbing elbows with glitterati, and maintaining one's status on their invitation lists (see, e.g., Russo, 2005). The ironic end result is that the most "influential" newspaper in education helps the vested interests to suppress information.

The dissolution of education knowledge is unfortunate for our society, but it is no small task to convince those outside the field that the problem even exists. Some skeptics simply refuse to believe that censorship and suppression on such a scale is possible in the Internet age, inadvertently reinforcing it. Others inside the education business benefit profoundly, personally, and professionally and would not be keen to relinquish their advantage.

REFERENCES

American Federation of Teachers. (1995). *Defining world class standards*. Washington, DC: Author.

Association for Supervision and Curriculum Development. (2005). *ASCD 1984–2004: Defining moments, future prospects*. Alexandria, VA: Author.

Association for Supervision and Curriculum Development. (2008, January 22). *ASCD SmartBrief*. Alexandria, VA: Author.

Barth, P. (2006, April 9). *High stakes testing and instruction: What the research says*. Washington, DC: National School Boards Association, Center for Public Education. Retrieved January 15, 2007, from http://www.centerforpubliceducation.org/atf/cf/%7B13A13846-1CA6-4F8A-B52E-2A88576B84EF%7D/HIGH-STAKES_TESTING_04092006.pdf

Bazerman, M. H., & Malhotra, D. (2006). Economics wins, psychology loses, and society pays. In D. de Cremer, M. Zeelenberg, & J. K. Murnighan (Eds.), *Social psychology and economics* (pp. 263–280). Mahwah, NJ: Erlbaum.

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

Bourque, M. L. (2005). Leave no standardized test behind. In R. P. Phelps (Ed.), *Defending standardized testing* (pp. 227–253). Mahwah, NJ: Erlbaum.

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

Briggs, D. C. (2001, Winter). The effect of admissions test preparation. *Chance, 14*(1), 10–18.

Britton, E. D., Hawkins, S., & Gandal, M. (1996). Comparing examinations systems. In E. D. Britton & S. A. Raizen (Eds.), *Examining the examinations* (pp. 201–218). Boston: Kluwer Academic.

Brookhart, S. M. (1993, Summer). Teachers' grading practices: Meaning and values. *Journal of Educational Measurement, 30*(2), 123–142.

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

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

Chandler, M. (Producer). (1999, Fall). Secrets of the SAT [Television series episode]. *Frontline*. Washington, DC: Public Broadcasting System.

Cizek, G. J. (2001). More unintended consequences of high-stakes testing? *Educational Measurement: Issues and Practice, 20*(4), 19–27.

Cole, N., & Willingham, W. (1997). *Gender and fair assessment*. Princeton, NJ: Educational Testing Service.

Corbett, H. D., & Wilson, B. (1991). *Testing, reform, and rebellion*. Norwood, NJ: Ablex.

Crocker, L. (2005). Teaching *for* the test. In R. P. Phelps (Ed.), *Defending standardized testing* (pp. 159–174). Mahwah, NJ: Erlbaum.

Dempster, F. N. (1991, April). Synthesis of research on reviews and tests. *Educational Leadership*, 71–76.

Dempster, F. N. (1997). Using tests to promote classroom learning. In R. F. Dillon (Ed.), *Handbook on testing* (pp. 332–346). Westport, CT: Greenwood Press.

de Tocqueville, A. (2003). *Democracy in America.* New York: Penguin Classics. (Original work published 1835 and 1840)

Domhoff, G. W. (2006). *Who rules America? Power, politics, and social change* (5th ed.). New York: McGraw-Hill.

Eckstein, M. A., & Noah, H. J. (1993). *Secondary school examinations: International perspectives on policies and practice.* New Haven, CT: Yale University Press.

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

Farkus, S., Johnson, J., Immerwahr, J., & McHugh, J. (1998). *Time to move on.* New York: Public Agenda.

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

Feuer, M. J., & Fulton, K. (1994). Educational testing abroad and lessons for the United States. *Educational Measurement: Issues and Practice, 13*(2), 31–39.

Geisinger, K. F. (2005). The testing industry, ethnic minorities, and individuals with disabilities. In R. P. Phelps (Ed.), *Defending standardized testing* (pp. 187–204). Mahwah, NJ: Erlbaum.

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

Greene, H. A., & Jorgensen, A. N. (1929). *The use and interpretation of educational tests.* New York: Longmans, Green.

Guskey, T. R., & Gates, S. L. (1986). Synthesis of research on the effects of mastery learning in elementary and secondary classrooms. *Educational Leadership, 43,* 73–80.

Haney, W. M., Madaus, G. F., & Lyons, R. (1993). *The fractured marketplace for standardized testing.* Boston: Kluwer Academic.

Hanushek, E. A. (2006). *Policy analysis: Is it, or could it be, the Fifth Estate?* (2005 Spencer Foundation Distinguished Lecture in Education Policy and Management). Washington, DC: Association for Public Policy Analysis and Management.

Hanushek, E. A., & Raymond, M. E. (2002, June 9–11). *Lessons about the design of state accountability systems.* Paper presented at the conference "Taking Account of Accountability: Assessing Policy and Politics," Harvard University, Cambridge, MA.

Hanushek, E. A., & Raymond, M. E. (2003). Lessons about the design of state accountability systems. In P. E. Peterson & M. R. West (Eds.), *No child left behind? The politics and practice of accountability* (pp. 126–151). Washington, DC: Brookings Institution.

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

Haynie, W. J., III. (1994). Effects of multiple-choice and short-answer tests on delayed retention learning. *Journal of Technology Education, 6*(1), 32–44.

Haynie, W. J., III. (2007). Effects of test taking on retention learning in technology education: A meta-analysis. *Journal of Technology Education, 18*(2), 24–36.

Heubert, J. P., & Hauser, R. P. (Eds.). (1999). *High-stakes: Testing for tracking, promotion, and graduation*. Washington, DC: National Research Council.

Interview: Rod Paige. (2001, September). Boston: WGBH. Retrieved November 3, 2006, from http://www.pbs.org/wgbh/pages/frontline/shows/schools/interviews/paige.html

Jacob, B. A. (2001). Getting tough? *Educational Evaluation and Policy Analysis, 23*, 99–121.

Jacob, B. A. (2002). *Accountability, incentives and behavior: The impact of high-stakes testing in the Chicago Public Schools*. Unpublished manuscript.

Jacob, B. A. (2003). High stakes in Chicago. *Education Next, 1*, 66.

Jones, B. D., & Egley, R. J. (2003, April). *The carrot and the stick*. Paper presented at the Annual Meeting of the Eastern Educational Research Association, New Orleans, LA.

Jones, B. D., & Johnston, A. F. (2002, April). *The effects of high-stakes testing on instructional practices*. Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans, LA.

Jones, E. H. (1923–1924). The effects of examination on the permanence of learning. *Archives of Psychology, 10*, 36–54.

Kellaghan, T., & Madaus, G. F. (1995). National curricula in European countries. In E. W. Eisner (Ed.), *The hidden consequences of a national curriculum* (pp. 79–118). Washington, DC: American Educational Research Association.

Kellaghan, T., Madaus, G. F., & Airasian, P. W. (1982). *The effects of standardized testing*. Boston: Kluwer-Nijhoff.

Kellaghan, T., Madaus, G. F., & Raczek, A. (1996). *The use of external examinations to improve student motivation*. Washington, DC: American Educational Research Association.

Kiplinger, V. L., & Linn, R. L. (1993). *Raising the stakes of test administration: The impact on student performance on NAEP* (CSE Technical Report 360). Los Angeles, CA: Center for Research on Education Standards and Student Testing.

Klein, S. P., Hamilton, L. S., McCaffrey, D. F., & Stecher, B. M. (2000). *What do test scores in Texas tell us?* Santa Monica, CA: RAND Education.

Kohn, A. (2000). Standardized testing and its victims. *Education Week, 20*(4), 60, 46–47.

Koretz, D. M. (1996). Using student assessments for educational accountability. In E. A. Hanushek & D. W. Jorgenson (Eds.), *Improving America's schools: The role of incentives* (pp. 171–195). Washington, DC: National Academies Press.

Koretz, D. M., & Hamilton, L. S. (2007). Testing for accountability in K–12. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 531–578). Washington, DC: American Council on Education.

Koretz, D. M., Linn, R. L., Dunbar, S. B., & Shepard, L. A. (1991, April 5). *The effects of high-stakes testing on achievement: Preliminary findings about generalization across tests*. Paper presented at the Annual Meeting of the American Educational Research Association, Chicago.

Lieberman, M. (2007). *The educational morass: Overcoming the stalemate in American education*. Lanham, MD: Rowman & Littlefield.

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

Lohman, D. F. (2006). Beliefs about differences between ability and accomplishment: From folk theories to cognitive science. *Roeper Review, 29*, 32–40.

Madaus, G. F. (1991a, June). *The effects of important tests on students*. Paper presented at the American Educational Research Association Conference on Accountability as a State Reform Instrument, Washington, DC.

Madaus, G. F. (1991b, November). The effects of important tests on students: Implications for a national examination system. *Phi Delta Kappan*, 226–231.

Madaus, G. F., & Kellaghan, T. (1991). *Student examination systems in the European Community: Lessons for the United States* (OTA Contractor PB92-127570). Washington, DC: Office of Technology Assessment, U.S. Congress. (ERIC Document Reproduction Service No. ED340781)

McMillan, J. H. (2001). Secondary teachers' classroom assessment and grading practices. *Educational Measurement: Issues and Practice, 20*(1), 20–32.

McNeil, L. M. (2000). *Contradictions of school reform*. New York: Routledge.

McNeil, L. M., & Valenzuela, A. (2000). *The harmful impact of the TAAS system of testing in Texas*. Cambridge, MA: Harvard University, Civil Rights Project.

Medina, N., & Neill, M. (1990). *Fallout from the testing explosion*. Cambridge, MA: FairTest.

Mehrens, W. A. (1998). Consequences of assessment: What is the evidence? *Education Policy Analysis Archives, 6*(13). Retrieved January 19, 2008, from http://epaa.asu.edu/epaa/v6n13.html

Merrow, J. (2002, March 29). *Testing our schools* [Transcript, Live Online discussion]. Retrieved September 26, 2008, from http://discuss.washingtonpost.com/zforum/02/tv_frontline032902.htm

Mitchell, R. (2006, March 30). *Key lessons: High-stakes testing and effects on instruction*. Washington, DC: National School Boards Association, Center for Public Education. Retrieved January 14, 2008, from http://www.centerforpubliceducation.org/site/c.kjJXJ5MP IwE/b. 1533781/k.B7BE/Key_lessons_Highstakes_testing_and_effects_on_instruction.htm

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

National Commission on Excellence in Education. (1983, April). *A nation at risk: The imperative for educational reform*. Washington, DC: U.S. Department of Education. Available at http://www.ed.gov/pubs/NatAtRisk/index.html

National Commission on Testing and Public Policy. (1990). *From gatekeeper to gateway: Transforming testing in America*. Chestnut Hill, MA: Author.

Neill, M. (1992). Correcting business leaders' assumptions about testing [Letter]. *Education Week, 11*(27), 46.

New Report Confirms Accountability Tests Are Powerful Tool in Ensuring Students Are Not Left Behind [Press release]. (2003, February 11). U.S. Congress, Committee on Education and the Workforce.

Office of Technology Assessment. (1992). *Testing in American schools: Asking the right questions* (OTA-SET-519). Washington, DC: U.S. Government Printing Office.

Olson, L. (2002, June 19). Accountability studies find mixed impact on achievement. *Education Week, 21*(41), 13.

Palmer, J. S. (2002). *Performance incentives, teachers, and students: Estimating the effects of rewards policies on classroom practices and student performance.* Unpublished doctoral dissertation, The Ohio State University, Columbus.

Phelps, R. P. (n.d.). Censorship has many fathers. Admonitions on tone, rigor, polemic, and other excuses for censoring information that one dislikes. *Third Education Group Review.* Retrieved July 7, 2008, from http://www.thirdeducationgroup.org/Foundation/CensorshipHasManyFathers

Phelps, R. P. (1994). The fractured marketplace for standardized testing [Book review]. *Economics of Education Review, 13*, 367–370.

Phelps, R. P. (1996, Fall). Are U.S. students the most heavily tested on Earth? *Educational Measurement: Issues and Practice, 15*(3), 19–27.

Phelps, R. P. (1997). The extent and character of system-wide student testing in the United States. *Educational Assessment, 4*, 89–121.

Phelps, R. P. (1998, Fall). The demand for standardized student testing. *Educational Measurement: Issues and Practice, 17*(3), 5–23.

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

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

Phelps, R. P. (2000b). Trends in large-scale, external testing outside the United States. *Educational Measurement: Issues and Practice, 19*(1), 11–21.

Phelps, R. P. (2001, August). Benchmarking to the world's best in mathematics. *Evaluation Review, 25*, 391–439.

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

Phelps, R. P. (Ed.). (2005a). *Defending standardized testing.* Mahwah, NJ: Erlbaum.

Phelps, R. P. (2005b). Persistently positive. In R. P. Phelps (Ed.), *Defending standardized testing* (pp. 1–22). Mahwah, NJ: Erlbaum.

Phelps, R. P. (2005c). The rich, robust research literature on testing's achievement benefits. In R. P. Phelps (Ed.), *Defending standardized testing* (pp. 55–90). Mahwah, NJ: Erlbaum.

Phelps, R. P. (2005d). The source of Lake Wobegon. *The Third Education Group Review, 1*(2). Retrieved July 7, 2008, from http://www.thirdeducationgroup.org/Review/Articles/v1n2.pdf

Phelps, R. P. (2007a, Summer). The dissolution of education knowledge. *Educational Horizons, 85*, 232–247.

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

Picus, L. O., & Tralli, A. (1998). *Alternative assessment programs: What are the true costs?* (CSE Technical Report 441). Los Angeles, CA: Center for Research on Education Standards and Student Testing.

Plake, B. S. (2005). Doesn't everyone know that 70% is passing? In R. P. Phelps (Ed.). *Defending standardized testing* (pp. 175–186). Mahwah, NJ: Erlbaum.

Posner, R. A. (2001). *Public intellectuals: A study of decline*. Cambridge, MA: Harvard University Press.

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

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

Roderick, M., Jacob, B., & Bryk, A. (2002). The impact of high-stakes testing in Chicago on student achievement in the promotional gate grades. *Educational Evaluation and Policy Analysis, 24*, 333–357.

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

Roediger, H. L., & Karpicke, J. D. (2006b). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science, 17*, 249–255.

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

Rogosa, D. (2005). A school accountability case study. In R. P. Phelps (Ed.), *Defending standardized testing* (pp. 205–226). Mahwah, NJ: Erlbaum.

Ross, C. C. (1941). *Measurement in today's schools*. New York: Prentice-Hall.

Rothman, R. (1995). *Measuring up*. New York: Wiley.

Rudman, H. C. (1992). Testing for learning [Book review]. *Educational Measurement: Issues and Practice, 11*(3), 31–32.

Russo, A. (2005, August 12–13). Hot for education: The top five best-looking school reformers in the nation. *This Week in Education*. Retrieved July 19, 2008, from http://thisweekineducation.blogspot.com/2005_08_01_archive.html

Sacks, P. (1999). *Standardized minds*. Cambridge, MA: Perseus.

Sandham, J. L. (1998). Ending SAT may hurt minorities. *Education Week, 17*, 5.

Shepard, L. A., Kreitzer, A. E., & Grau, M. E. (1987). *A case study of the Texas teacher test* (CSE Report No. 276). Los Angeles, CA: Center for Research on Education Standards and Student Testing.

Sireci, S. G. (2005). The most frequently unasked questions about testing. In R. P. Phelps (Ed.), *Defending standardized testing* (pp. 111–121). Mahwah, NJ: Erlbaum.

Smith, M. L. (1991a). *The role of testing in elementary schools* (CSE Technical Report 321). Los Angeles: Center for Research on Education Standards and Student Testing.

Smith, M. L. (1991b, June). Put to the test. *Educational Researcher, 20,* 8–11.

Smith, M. L. (1991c, Fall). Meanings of test preparation. *American Educational Research Journal, 28,* 521–542.

Smith, M. L., & Rottenberg, C. (1991). Unintended consequences of external testing in elementary schools. *Educational Measurement: Issues and Practice, 10,* 10–11.

Snowman, J. (2002). *Psychology applied to teaching.* Boston: Houghton-Mifflin.

Stiggins, R. J., & Conklin, N. F. (1992). *In teachers' hands: Investigating the practices of classroom assessment.* New York: SUNY Press.

Strauss, V. (2006, October 10). The rise of the testing culture. *Washington Post,* p. A09.

Teel, M. L. (Producer). (2001, April 1). *CBS Sunday Morning* [Television broadcast]. New York: CBS News.

Thomson, J. A. (2000, October 26). *Statement of Rand President and CEO James A. Thomson.* [Press release]. Santa Monica, CA: Rand.

Thompson, T. D. (1990). *When mastery testing pays off.* Unpublished doctoral dissertation, University of Oklahoma, Norman.

Traub, R. E. (1993). On the equivalence of the traits assessed by multiple choice and constructed-response tests. In R. E. Bennett & W. C. Ward (Eds.), *Construction versus choice in cognitive measurement* (pp. 29–44). Hillsdale, NJ: 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, LA. (ERIC Document Reproduction Service No. ED368790).

Tuckman, B. W., & Trimble, S. (1997, August). *Using tests as a performance incentive to motivate eighth-graders to study.* Paper presented at the 105th Annual Convention of the American Psychological Association, Chicago. (ERIC Document Reproduction Service No. ED418785)

U.S. General Accounting Office. (1993, January). *Student testing: Current extent and expenditure* (GAO/PEMD-93-8). Washington, DC: Author.

Viadero, D. (1994). National tests in other countries not as prevalent as thought. *Education Week, 13*(37), 10.

Wolk, R. A. (2002, April). Multiple measures. *Teacher Magazine, 13*(7), 3.

APPENDIX 3.1
SELECTED CITATIONS FROM 20TH-CENTURY RESEARCH RELEVANT TO THE DESIGN OF THE NO CHILD LEFT BEHIND ACT

Studies marked with an asterisk (*) are research reviews, research syntheses, or meta-analyses. Studies that are listed more than once are given in shortened form after the first listing.

EFFECTS OF STANDARDS, ALIGNMENT, GOAL SETTING, SETTING REACHABLE GOALS

Aguilera, R. V., & Hendricks, J. M. (1996, September). Increasing standardized achievement scores in a high risk school district. *Curriculum Report, 26*(1), 1–6.

Anderson, J. O., Muir, W., Bateson, D. J., Blackmore, D., & Rogers, W. T. (1990, March 30). *The impact of provincial examinations on education in British Columbia: General report.* Victoria, British Columbia, Canada: British Columbia Ministry of Education.

*Bamburg, J., & Medina, E. (1993). Analyzing student achievement: Using standardized tests as the first step. In J. Bamburg (Ed.), *Assessment: How do we know what they know?* (pp. 35–40). Dubuque, IA: Kendall-Hunt.

Banta, T. W., Lund, J. P., Black, K. E., & Oblander, F. W. (1996). *Assessment in practice: Putting principles to work on college campuses.* San Francisco: Jossey-Bass.

Bishop, J. H. (1993, December). *Impact of curriculum-based examinations on learning in Canadian secondary schools* (Working Paper 94-30). Ithaca, NY: Center for Advanced Human Resource Studies, New York State School of Industrial and Labor Relations, Cornell University.

Bottoms, G., & Mikos, P. (1995). *Seven most-improved "High Schools That Work" sites raise achievement in reading, mathematics, and science.* Atlanta, GA: Southern Regional Education Board.

Brooke, N., & Oxenham, J. (1984). The influence of certification and selection on teaching and learning. In J. Oxenham (Ed.), *Education versus qualifications? A study of relationships between education, selection for employment and the productivity of labor* (pp. 147–175). London: George Allen & Unwin.

Brown, D. F. (1992, April). *Altering curricula through state testing: Perceptions of teachers and principals.* Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco.

Czikszentmihalyi, M. (1990). *Flow: The psychology of optimal experience.* New York: Harper Perennial.

Eckstein, M. A., & Noah, H. J. (1993). *Secondary school examinations: International perspectives on policies and practice.* New Haven, CT: Yale University Press.

Estes, G. D., Colvin, L. W., & Goodwin, C. (1976, April). *A criterion-referenced basic skills assessment program in a large city school system.* Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco.

Heyneman, S. P., & Ransom, A. W. (1992). Using examinations and testing to improve educational quality. In M. A. Eckstein & H. J. Noah (Eds.), *Examinations: Comparative and international studies* (pp. 105–121). Oxford, England: Pergamon Press.

Hilloks, G., Jr. (1987). Synthesis of research on teaching writing. *Educational Leadership, 44*(8), 71–82.

LaRoque, L., & Coleman, P. (1989). Quality control: School accountability and district ethos. In M. Holmes, K. A. Leithwood, & D. F. Musella, (Eds.), *Educational policy for effective schools* (pp. 168–191). Toronto, Ontario, Canada: Ontario Institute for Studies in Education.

*Levine, D. U., & Lezotte, L. W. (1990). *Unusually effective schools: A review and analysis of research and practice.* Madison, WI: The National Center for Effective Schools Research and Development.

Mattsson, H. (1993, May–June). *Impact of assessment on educational practice and student behavior in the Swedish schools system, School-based and external assessments: Uses and issues.* Paper presented at the 19th Annual Conference of the International Association for Educational Assessment, Mauritius.

Miles, W., Bishop, J., Collins, J., Fink, J., Gardner, M., Grant, J., et al. (1997). *Ten case studies of "All Regents" high schools: Final report to the State Education Department.* Albany, NY: State Education Department.

Mitchell, F. M. (1999, April). *All students can learn: Effects of curriculum alignment on the mathematics achievement of third-grade students.* Paper presented at the Annual Meeting of the American Educational Research Association, Montréal, Québec, Canada.

Morgan, R., & Ramist, L. (1998, February). *Advanced placement students in college: An investigation of course grades at 21 colleges* (Rep. No. SR-98-13). Princeton, NJ: Educational Testing Service.

*Natriello, G., & Dornbusch, S. M. (1984). *Teacher evaluative standards and student effort.* New York: Longman.

Office of Program Policy Analysis and Government Accountability, the Florida Legislature. (1997, June). *Improving student performance in high-poverty schools.* Tallahassee, FL: Author.

Ogle, D., & Fritts, J. B. (1981). Criterion-referenced reading assessment valuable for process as well as for data. *Phi Delta Kappan, 62*(9), 640–641.

Panlasigui, I., & Knight, F. B. (1930). The effect of awareness of success or failure. In F. B. Knight (Ed.), *Twenty-ninth yearbook of the National Society for the Study of Education: Report of the society's committee on arithmetic* (pp. 611–619). Chicago: University of Chicago Press.

Pomplun, M. (1997). State assessment and instructional change: A path model analysis. *Applied Measurement in Education, 10*(3), 217–234.

Rentz, R. R. (1979). Testing and the college degree. In W. B. Schrader (Ed.), *Measurement and educational policy: New directions for testing and measurement* (pp. 71–78). San Francisco: Jossey-Bass.

Resnick, D. P., & Resnick, L. B. (1985). Standards, curriculum, and performance: A historical and comparative perspective. *Educational Researcher, 14*(4), 5–20.

*Rosswork, S. G. (1977). Goal setting: The effects of an academic task with varying magnitudes of incentive. *Journal of Educational Psychology, 69*, 710–715.

Schmoker, M. (1996). *Results: The key to continuous school improvement.* Alexandria, VA: Association for Supervision and Curriculum Development.

*Southern Regional Education Board. (1998). *High schools that work: Case studies.* Available at http://www.sreb.org

U.S. General Accounting Office. (1993, April). *Educational testing: The Canadian experience with standards, examinations, and assessments* (GAO/PEMD-93-11). Washington, DC: Author.

Wellisch, J. B., MacQueen, A. H., Carriere, R. A., & Duck, G. A. (1978). School management and organization in successful schools. *Sociology of Education, 51*, 211–226.

Whetton, C. (1992, April). *Advice to US systems contemplating performance assessment.* Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco.

Willingham, W. W., & Morris, M. (1986). *Four years later: A longitudinal study of advanced placement students in college* (College Board Rep. 86-2). New York: College Entrance Examination Board.

EFFECTS OF TESTS, ACCOUNTABILITY PROGRAMS, OR BOTH ON MOTIVATION AND INSTRUCTIONAL PRACTICE

Brown, S. M., & Walberg, H. J. (1993). Motivational effects on test scores of elementary students. *Journal of Educational Research, 86*(3), 133–136.

Brunton, M. L. (1982, March). *Is competency testing accomplishing any breakthrough in achievement?* Paper presented at the Annual Meeting of the Association for Supervision and Curriculum Development, Anaheim, CA.

Chao-Qun, W., & Hui, Z. (1993). Educational assessment in mathematics teaching: Applied research in China. In M. Niss (Ed.), *Cases of assessment in mathematics education: An ICMI study* (pp. 183–192). Boston: Kluwer Academic.

Clarke, D., & Stephens, M. (1996). The ripple effect: The instructional impact of the systemic introduction of performance assessment in math-

ematics. In M. Birenbaum & F. J. R. C. Dochy (Eds.), *Alternatives in assessment of achievements, learning processes and prior knowledge* (pp. 63–92). Boston: Kluwer Academic.

Eckstein and Noah (1993)

Egeland, P. C. (1995). *The effect of authentic assessments on fifth-grade student science achievement and attitudes*. Unpublished doctoral dissertation, Northern Illinois University, Dekalb.

Foss, O. (1977, May). A new approach: Vocational foundation courses and examinations. In F. M. Ottobre (Ed.), *Criteria for awarding school leaving certificates: An international discussion* (pp. 191–209). Based on the Proceedings of the 1977 Conference of the International Association for Educational Assessment held at the Kenyatta Conference Center, Nairobi, Kenya. Cambridge, England: International Association for Educational Assessment.

Johnson, J. F., Jr. (1998). The influence of a state accountability system on student achievement in Texas. *Virginia Journal of Social Policy & the Law*, 6(1), 155–178.

Keys, N. (1934). The influence on learning and retention of weekly tests as opposed to monthly tests. *Journal of Educational Psychology*, *25*, 427–436.

*Kirkland, M. C. (1971). The effects of tests on students and schools. *Review of Educational Research*, *41*, 303–350.

*Levine and Lezotte (1990)

Marsh, R. (1984, November/December). A comparison of take-home versus in-class exams. *Journal of Educational Research*, 78(2), 111–113.

Miles et al. (1997)

O'Sullivan, R. G. (1989, February). *Teacher perceptions of the effects of testing on students*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, San Francisco.

Pennycuick, D., & Murphy, R. (1988). *The impact of graded tests*. London: Falmer Press.

Plazak, T., & Mazur, Z. (1992). University entrance in Poland. In P. Black (Ed.), *Physics examinations for university entrance: An international study* (Science and Technology Education Document Series No. 45, pp. 135–149). Paris: UNESCO.

Prais, S. (1995). *Productivity, education and training* (Vol. 2). London: National Institute for Economic and Social Research.

Ritchie, D., & Thorkildsen, R. (1994). Effects of accountability on students' achievement in mastery learning. *Journal of Educational Research*, 88(2), 86–90.

Schafer, W. D., Hultgren, F. H., Hawley, W. D., Abrams, A. L., Seubert, C. C., & Mazzoni, S. (1997). *Study of higher-success and lower-success elementary schools*. College Park, MD: University of Maryland, School Improvement Program.

Singh, J. S., Marimutha, T., & Mukjerjee, H. (1990). Learning motivation and work: A Malaysian perspective. In P. Broadfoot, R. Murphy, & H. Torrance (Eds.), *Changing educational assessment: International perspectives and trends* (pp. 177–198). London: Routledge.

Solberg, W. (1977). School leaving examinations: Why or why not?: The case for school leaving examinations: The Netherlands. In F. M. Ottobre (Ed.), *Criteria for awarding school leaving certificates: An international discussion* (pp. 37–46). Based on the Proceedings of the 1977 Conference of the International Association for Educational Assessment held at the Kenyatta Conference Center, Nairobi, Kenya. Cambridge, England: International Association for Educational Assessment.

Somerset, A. (1968). *Examination reform: The Kenya experience* (Rep. No. EDT64). Washington, DC: The World Bank.

*Southern Regional Education Board (1998)

Steedman, H. (1992). *Mathematics in vocational youth training for the building trades in Britain, France and Germany* (Discussion Paper No. 9). London: National Institute for Economic and Social Research.

Stevens, F. I. (1984, December). *The effects of testing on teaching and curriculum in a large urban school district* (ERIC/TM Rep. 86). Washington, DC: ERIC Clearinghouse on Tests, Measurement, and Evaluation.

Stevenson, H. W., & Lee, S.-L. (1997). *International comparisons of entrance and exit examinations: Japan, United Kingdom, France, and Germany*. Washington, DC: U.S. Department of Education, Office of Educational Research and Improvement.

Stuit, D. B. (Ed.). (1947). *Personnel research and test development in the Bureau of Naval Personnel*. Princeton, NJ: Princeton University Press.

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, LA.

Tuckman, B. W., & Trimble, S. (1997, August). *Using tests as a performance incentive to motivate eighth-graders to study*. Paper presented at the 105th Annual Convention of the American Psychological Association, Chicago, IL.

Wall, D., & Alderson, J. C. (1993). Examining washback: The Sri Lankan Impact Study. *Language Testing, 10*, 41–69.

Waters, T., Burger, D., & Burger, S. (1995, March). Moving up before moving on. *Educational Leadership, 52*(6), 35–40.

Wolf, A., & Rapiau, M. T. (1993). The academic achievement of craft apprentices in France and England: Contrasting systems and common dilemmas. *Comparative Education, 29*(1), 29–43.

Zigarelli, M. A. (1996). An empirical test of conclusions from effective schools research. *Journal of Educational Research, 90*(2), 103–110.

THE LEARNING EFFECTS OF TESTS THEMSELVES

Banta et al. (1996)

Dempster, F. N. (1991, April). Synthesis of research on reviews and tests. *Educational Leadership, 48,* 71–76.

Dempster, F. N. (1997). Using tests to promote classroom learning. In R. F. Dillon (Ed.), *Handbook on testing* (pp. 332–346). Westport, CT: Greenwood Press.

Kirkpatrick, J. E. (1934). The motivating effect of a specific type of testing program. *University of Iowa Studies in Education, 9,* 41–68.

Ross, C. C., & Henry, L. K. (1939). The relation between frequency of testing and learning in psychology. *Journal of Educational Psychology, 69,* 710–715.

Toppino, T. C., & Luipersbeck, S. M. (1993). Generality of the negative suggestion effect in objective tests. *Journal of Educational Research,* 86(6), 357–362.

EFFECTS OF VARYING TYPES OF INCENTIVES OR THE OPTIMAL STRUCTURE OF INCENTIVES

Abbott, R. D., & Falstrom, P. (1977). Frequent testing and personalized systems of instruction. *Contemporary Educational Psychology, 2,* 251–257.

Banta et al. (1996)

Brooke and Oxenham (1984)

Brookover, W. B., & Lezotte, L. W. (1979, May). *Changes in school characteristics coincident with changes in student achievement* (Occasional Paper No. 17). East Lansing, MI: Michigan State University, Institute for Research on Teaching.

Brooks-Cooper, C. (1993, August). *The effect of financial incentives on the standardized test performance of high school students.* Unpublished master's thesis, Cornell University, Ithaca, NY.

Corcoran, T. B., & Wilson, B. L. (1986, October). *The search for successful secondary schools: The first three years of the secondary school recognition program.* Philadelphia: Research for Better Schools.

Cronbach, L. J. (1960). *Essentials of psychological testing.* New York: Harper & Row.

*Crooks, T. J. (1988). The impact of classroom evaluation practices on students. *Review of Educational Research,* 58(4), 438–481.

Duran, R. P. (1989). Assessment and instruction of at-risk Hispanic students. *Exceptional Children,* 56(2), 154–158.

Eckstein and Noah (1993)

*Guskey, T. R., & Gates, S. L. (1986). Synthesis of research on the effects of mastery learning in elementary and secondary classrooms. *Educational Leadership, 43*(8), 73–80.

Heneman, H. G., III. (1998). Assessment of the motivational reactions of teachers to a school-based performance award program. *Journal of Personnel Evaluation in Education, 12*, 143–159.

Heyneman and Ransom (1992)

Hurlock, E. B. (1925, September). The effects of incentives on the constancy of the I.Q. *Pedagogical Seminary, 32*, 422–434.

Jacobson, J. E. (1992, October 29). *Mandatory testing requirements and pupil achievement.* Unpublished doctoral dissertation, Massachusetts Institute of Technology, Boston.

*Kazdin, A., & Bootzin, R. (1972). The token economy: An evaluative review. *Journal of Applied Behavior Analysis, 5*, 343–372.

Kelley, C. (1999). The motivational impact of school-based performance awards. *Journal of Personnel Evaluation in Education, 12*(4), 309–326.

*Kulik, C.-L., & Kulik, J. A. (1987). Mastery testing and student learning: A meta-analysis. *Journal of Educational Technology Systems, 15*, 325–345.

*Levine and Lezotte (1990)

McMillan, J. H. (1977). The effect of effort and feedback on the formation of student attitudes. *American Educational Research Journal, 14*(3), 317–330.

*O'Leary, K. D., & Drabman, R. (1971). Token reinforcement programs in the classroom: A review. *Psychological Bulletin, 75*, 379–398.

Oxenham, J. (1984). *Education versus qualifications?* London: Unwin Education.

Richards, C. E., & Shen, T. M. (1992, March). The South Carolina school incentive reward program: A policy analysis. *Economics of Education Review, 11*(1), 71–86.

*Southern Regional Education Board (1998)

*Staats, A. (1973). Behavior analysis and token reinforcement in educational behavior modification and curriculum research. In C. E. Thoreson (Ed.), *72nd yearbook of the NSSE: Behavior modification in education* (Part 1, pp. 195–229). Chicago: University of Chicago Press.

Trelfa, D. (1998, June). The development and implementation of education standards in Japan. *The educational system in Japan: Case study findings* (pp. 59–106). Washington, DC: U.S. Department of Education, Office of Educational Research and Improvement, National Institute on Student Achievement, Curriculum, and Assessment.

Venesky, R. L., & Winfield, L. F. (1979, August). *Schools that succeed beyond expectations in teaching* (University of Delaware Studies on Education Tech. Rep. No. 1). Newark: University of Delaware.

EFFECTS OF VARYING PATTERNS AND FREQUENCIES OF TESTING

*Bangert-Drowns, R. L., Kulik, J. A., & Kulik, C.-L. C. (1991). Effects of frequent classroom testing. *Journal of Educational Research, 85*(2), 89–99.

Beardon, D. (1997). *An overview of the elementary mathematics program 1996–97* (Research Rep. REIS97-116-3). Dallas, TX: Dallas Public Schools.

Gaynor, J., & Millham, J. (1976). Student performance and evaluation under variant teaching and testing methods in a large college course. *Journal of Education Psychology, 68*, 312–317.

Khalaf, A. S. S., & Hanna, G. S. (1992). The impact of classroom testing frequency on high school students' achievement. *Contemporary Educational Psychology, 17*(1), 71–77.

Kika, F. M., McLaughlin, T. F., & Dixon, J. (1992). Effects of frequent testing of secondary algebra students. *Journal of Educational Research, 85*(3), 159–162.

*Kulik, J. A., & Kulik, C.-L. C. (1989). The concept of meta-analysis. *International Journal of Education Research, 13*(3), 227–340.

Mullis, I. V. S. (1997, April). *Benchmarking toward world-class standards: Some characteristics of the highest performing school systems in the TIMSS*. Paper presented at the Annual Meeting of the American Educational Research Association, Chicago.

Rohm, R. A., Sparzo, F. J., & Bennett, C. M. (1986). College student performance under repeated testing and cumulative testing conditions: Report on five studies. *Journal of Educational Research, 80*(2), 99–104.

Taylor, B., Pearson, P. D., Clark, K., & Walpole, S. (2000). Effective schools and accomplished teachers: Lessons about primary grade reading instruction in low-income schools. *The Elementary School Journal, 10*, 121–165.

*Thompson, T. D. (1990). *When mastery testing pays off: The cost benefits and psychometric properties of mastery tests as determined from item response theory*. Unpublished doctoral dissertation, University of Oklahoma, Norman.

EFFECTS OF MONITORING, FEEDBACK, AND EVALUATION OF PERFORMANCE

Anderson et al. (1990)

Burrows, C. K., & Okey, J. R. (1975, April). *The effects of a mastery learning strategy*. Paper presented at the Annual Meeting of the American Educational Research Association, Washington, DC.

Corbett, H. D., & Wilson, B. (1989). *Raising the stakes in statewide mandatory minimum competency testing*. Philadelphia: Research for Better Schools.

*Crooks (1988)
Engel, G. S. (1977, March–April). One way it can be. *Today's Education, 66*, 50–52.
Friedman, H. (1987). Repeat examinations in introductory statistics. *Teaching of Psychology, 14*, 20–23.
Fuller, B. (1987). What school factors raise achievement in the Third World? *Review of Educational Research, 57*(3), 255–292.
Goodson, M. L., & Okey, J. R. (1978, November). The effects of diagnostic tests and help sessions on college science achievement. *Journal of College Science Teaching, 8*, 89–90.
Hess, A. C., & Lockwood, R. E. (1986, April). *The relationship of a basic competency education program to overall student achievement: A state perspective*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, San Francisco.
*Heyneman, S. P. (1987). *Uses of examinations in developing countries: Selection, research, and education sector management* (Seminar Paper No. 36). Washington, DC: Economic Development Institute, The World Bank.
Heyneman and Ransom (1992)
Lerner, B. (1990, March). Good news about American education. *Commentary, 91*(3), 19–25.
Magruder, J., McManis, M. A., & Young, C. C. (1997). The right idea at the right time: Development of a transformational assessment culture. *New Directions for Higher Education, 100*, 17–29.
Moss, H. A., & Kagan, J. (1961). Stability of achievement in recognition setting behaviors from early childhood through adulthood. *Journal of Abnormal and Social Psychology, 62*, 504–513.
Rodgers, N., Paredes, V., & Mangino, E. (1991, April). *High stakes minimum skills tests: Is their use increasing achievement?* Paper presented at the Annual Meeting of the American Educational Research Association, Chicago.
Ross, J. A., Rolheiser, C., & Hoaboam-Gray, A. (1998). *Impact of self-evaluation training on mathematics achievement in a cooperative learning environment*. Ottawa, Ontario, Canada: Social Science and Humanities Research Council.
Swanson, D. H., & Denton, J. J. (1976). *A comparison of remediation systems affecting achievement and retention in mastery learning*. (ERIC Document Reproduction Service No. ED131037)

EFFECTS OF TESTING ON AT-RISK STUDENTS, COMPLETION, DROPPING OUT, AND CURRICULAR OFFERINGS

Boylan, H., Bonham, B., Abraham, A., Anderson, J., Morante, E., Ramirez, G., & Bliss, L. (1996). *An evaluation of the Texas Academic Skills Program*. Austin, TX: Texas Higher Education Coordinating Board.

Enochs, J. C. (1978, May). Modesto, California: A return to the four Rs. *Phi Delta Kappan, 59*(9), 609–610.

Grisay, A. (1991). Improving assessment in primary schools: "APER" research reduces educational failure rates. In P. Weston (Ed.), *Assessment of pupil achievement: Motivation and school success* (Report of the Educational Research Workshop held in Liège [Belgium] 12–15 September, Council of Europe; pp. 103–118). Amsterdam, the Netherlands: Swets & Zeitlinger.

Jacobson (1992)

Johnstone, W. (1990, January). *Local school district perspectives.* Paper presented at the Annual Meeting of the Southwest Educational Research Association, Austin, Texas.

Jones, J. B. (1993). *Effects of the use of an altered testing/grading method on the retention and success of students enrolled in college mathematics.* Unpublished doctoral dissertation, East Texas State University, Tyler.

Jones, J. (1996). Offer them a carrot: Linking assessment and motivation in developmental Mathematics. *Research and Teaching in Developmental Education, 13*(1), 85–91.

McWilliams, J. M., & Thomas, A. C. (1976). The measurement of students' learning: An approach to accountability. *Journal of Educational Research, 70,* 50–52.

Pronaratna, B. (1976). *Examination reforms in Sri Lanka* (Experiments and Innovations in Education, No. 24). Paris, France: UNESCO.

Schleisman, J. (1999, October). *An in-depth investigation of one school district's responses to an externally-mandated, high-stakes testing program in Minnesota.* Paper presented at the Annual Meeting of the University Council for Educational Administration, Minneapolis, MN. (ERIC Document Reproduction Service No. ED440465)

*Southern Regional Education Board (1998)

Task Force on Educational Assessment Programs. (1979). *Competency testing in Florida: Report to the Florida Cabinet, Part 1.* Tallahassee, FL: Author.

Webster, W. J., Mendro, R. L., Orsack, T., Weerasinghe, D., & Bembry, K. (1997a). The Dallas Value-Added Accountability System. In J. Millman (Ed.), *Grading teachers, grading schools: Is student achievement a valid evaluation measure?* (pp. 81–99). Thousand Oaks, CA: Corwin Press.

Webster, W. J., Mendro, R. L., Orsack, T., Weerasinghe, D., & Bembry, K. (1997b). Little practical difference and pie in the sky. In J. Millman (Ed.), *Grading teachers, grading schools: Is student achievement a valid evaluation measure?* (pp. 120–131). Thousand Oaks, CA: Corwin Press.

Wellisch et al. (1978)

EFFECTS OF MINIMUM-COMPETENCY TESTING AND THE PROBLEMS WITH A SINGLE PASSING SCORE

Brunton (1982)

Findley, J. (1978, May). Westside's minimum competency graduation requirements: A program that works. *Phi Delta Kappan, 59*(9), 614–618.

Frederiksen, N. (1994). *The influence of minimum competency tests on teaching and learning.* Princeton, NJ: Educational Testing Service.

Ligon, G., Brightman, M., Davis, E., Hoover, H. D., Johnstone, W., Mangino, E., et al. (1990, January). *Statewide testing in Texas.* A symposium presented at the Annual Meeting of the Southwest Educational Research Association, Austin, TX.

Losack, J. (1987). *Mandated entry- and exit-level testing in the state of Florida: A brief history, review of current impact, and a look to the future.* Miami, FL: Miami-Dade Community College and Florida Office of Institutional Research.

Mangino, E., & Babcock, M. A. (1986, April). *Minimum competency testing: Helpful or harmful for high level skills?* Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco.

Ogden, J. (1979, April). *High school competency graduation requirements: Do they result in better graduates?* Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco.

Parramore, B. M. (1980, November). *Effects of mandated competency testing in North Carolina: The class of 1980.* Paper presented at the Annual Meeting of the Evaluation Research Society, Washington, DC.

Serow, R. C., Davies, J. J., & Parramore, B. M. (1982). Performance gains in a competency test program. *Educational Evaluation and Policy Analysis, 4*(4), 535–542.

Winfield, L. F. (1990, March). *School competency testing reforms and student achievement: Exploring a national perspective* (Research Rep.). Princeton, NJ: Educational Testing Service.

Note. For more information, see Phelps (2005c).