International Journal of Testing, 12: 21–43, 2012

ISSN: 1530-5058 print / 1532-7574 online
DOI: 10.1080/15305058.2011.602920

Routledge
Taylor & Francis Group

The Effect of Testing on Student Achievement, 1910–2010

Richard P. Phelps
Asheville, North Carolina

This article summarizes research on the effect of testing on student achievement as found in English-language sources, comprising several hundred studies conducted between 1910 and 2010. Among quantitative studies, mean effect sizes range from a moderate $d \approx 0.55$ to a fairly large $d \approx 0.88$, depending on the way effects are aggregated or effect sizes are adjusted for study artifacts. Testing with feedback produces the strongest positive effect on achievement. Adding stakes or frequency also strongly and positively affects achievement. Survey studies produce effect sizes above 1.0. Ninety-three percent of qualitative studies analyzed also reported positive effects.

Keywords: effect size, literature review, meta-analysis, research summary, student achievement, tests & testing

HOW CAN TESTING AFFECT ACHIEVEMENT?

Psychologists have been studying the effects of testing on educational achievement (and on memory) for a century. They theorize that testing affects achievement by way of certain mediating factors such as motivation, feedback, alignment, and the "pure" testing effect.

Test motivation has two forms: intrinsic and extrinsic. While intrinsically motivated, a student may work harder or better in order to perform well on a test simply for their own satisfaction, even if that test has no stakes (i.e., consequences). While extrinsically motivated, a student may work harder or better in order to perform well on a test with stakes such as course completion, certification, graduation, or grade promotion.

Test feedback ranges from simple awareness of test results (that measure progress or performance) to overt remediation. Feedback can be diagnostic.

Correspondence should be sent to Dr. Richard P. Phelps, PhD, One North Pack Square, Suite 301, Asheville, NC 28801, USA. E-mail: richardpphelps@yahoo.com

Alignment occurs when the content embedded in a test or the level of performance demanded by a test matches the content or intensity of an associated prescribed curriculum. Naturally, the more closely aligned the curriculum is to the test, the more likely students who have mastered the content will perform well on the test.[1]

The pure testing effect is an increase in achievement that occurs simply because students take a test instead of spending the same amount of time some other way, such as studying. The most prominent aspect of the pure testing effect appears to be the generation effect—students taking a test cannot passively absorb information as they might listening to lectures or reading textbooks, they must "generate" the information themselves, and that process apparently makes a more durable impression on memory (Bertsch, Pesta, Wiscott, & McDaniel, 2007).

METHOD

This article summarizes the research literature on the effect of testing on student achievement from 1910 to 2010, as found in English-language sources.

Searches

Two search methods—keyword searches and citation chains—were employed. A keyword search is standard in scholarly research. First, one identifies relevant document databases and research indexes and relevant keywords that should identify documents containing relevant content. Then, one lets the computer do the work finding and compiling a list of potential sources. In some pure research fields, where one could reasonably expect to find the largest proportion of relevant studies in a small selection of scholarly journals, a keyword search might be thorough enough to cover the topic reasonably well. For this study, more than 40 search terms were employed (e.g., test, exam, examination, assessment, accountability, competency, effect, consequences, impact, benefit, cost).

With citation chains, one uses the studies found in a keyword search to find related studies. Each study provides evidence of other, relevant studies inside its text, citations, or references. This is sometimes called the "ancestry method."

To be truly thorough, any literature search should follow citation chains, but the method is particularly important to use for a topic like this one. First, an interest in testing effects overlaps many research field boundaries, such as education, psychology, public policy, and sociology, and many more subfield boundaries. Second, tests are often developed by private companies and sponsored by governments, neither of which is commonly interested in publishing in scholarly journals. Third, some types of studies simply cannot be found through simple keyword searches. They include proprietary work, audits performed by governmental agencies and

program evaluations sponsored by them, most studies conducted during the first two thirds of the twentieth century, and master's theses. Fourth, a large proportion of studies that measured testing effects were focused primarily on other results, and it is those foci and those results that determined which keywords were chosen for the research data bases. When testing effect findings are unplanned or coincidental to a study, a computer search on typical keywords probably will not find it. Fifth, a reliance on keyword search can be inadequate to the task due to variation in keywords across fields; different research disciplines employ different vocabularies. A "net benefit" to an economist, for example, may be called "consequential validity" by psychometricians, "positive effects" by program evaluators, or "positive washback" by education planners.

Searches for this study were limited only by the English language. If one source led to another via citation or reference, that lead was followed. But, indexes were used systematically, too. Those indexes included the Education Resources Information Clearinghouse (ERIC); FirstSearch of the Online Computer Library Center (OCLC), which includes PsychLit and hundreds of other social science data bases; Dissertation Abstracts; several EBSCO data bases; Google; Google Scholar; Yahoo! Search; and several individual libraries' catalogues.

In addition to all the aforementioned sources, two lists unique to opinion polls were consulted: the Roper Center Public Opinion Online and Polling the Nations databases.

A complete list of the studies reviewed and their bibliographic references is posted at http://npe.educationnews.org/Review/Resources/**.htm, substituting either "QuantitativeList," "SurveyList," or "QualitativeList" for "**."

Geographic Coverage

No attempt was made to limit the search by geographic region. Naturally, however, limiting the search to English-language documents biases a search in favor of those where the English language is predominant—*le monde Anglo-Saxon*. Granted, English may be used as the language of scientific communication even in countries where another is spoken. But, the search for this study was not restricted to journal articles, and other types of documents, such as government reports and conference papers, are more likely to be written in a home language.

In addition, because the search included any opportunist discoveries made in library catalogues and thesis indexes and only libraries in the United States were visited, a bias toward United States sources is introduced. Finally, the public opinion poll indexes were limited exclusively to US and Canadian sources.

Overall, 81% of the studies included in the analysis had a primarily US focus, but the geographic coverage varies by methodology type. Whereas 89% of the quantitative studies had a US focus, only 65% of the qualitative studies did. These proportions for the United States might seem disproportionately high, but the

United States does, in fact, host over two-thirds of the English-speaking population of the industrialized world.[2]

Perhaps in part due to inclusion of the uniquely North American public opinion poll data, 95% of the survey studies had a US focus. Indeed, there were so few survey studies (N = 5) from outside the United States and Canada, they were dropped from the analysis, making the analysis of survey studies exclusively North American.

Over 3000 documents were found that potentially contained evidence of testing effects on achievement. Less than one third qualified for inclusion in the analysis, however. The other 2000 or so were carefully reviewed, found not to contain relevant or sufficient evidence, and set aside.[3] Titles of several hundred more potentially useful sources fill a do-list of documents waiting to be processed in a potential next stage of this research.

Study Types

Searching uncovered a century's worth of studies measuring the effects of testing-related intervention on student achievement. This broad reach captures quite a variety of studies, including many of the historically most typical—relatively small, focused, randomized experiments with undergraduate psychology classes—and the recently popular—hugely aggregated state- or nationwide multivariate data-crunchings of accountability regimes with hundreds of independent variables contributing explanation, obfuscation, or both.

Quantitative studies. One hundred seventy-seven quantitative research studies were reviewed that include 640 separate measurements of effects. Some studies report multiple relevant measures of effects (e.g., measured at different times, with different subgroups, under different conditions). All told, the 640 measures comprise information for close to 7 million separate individuals. Quantitative studies directly measure effects and employ, for example, regression analysis, structural equation modeling, pre-post comparison, experimental design, or interrupted time series design.

These quantitative studies manifest a variety of study designs, the most numerous being straightforward experiments (or quasi-experiments) comparing means (or mean gains, or proportions) of a treatment and a control group. For example, an experiment might randomly assign students to two different classrooms, administer a baseline test, and then test the two groups differently throughout a course, say, giving one group a mid-term test and the other group no mid-term test. Then, a final test could be administered and the pre-post gain scores compared to determine the effect of giving a mid-term test.[4]

Typically, the studies compared groups (or the same [or similar] group at different time points) facing different tests or test-related situations. One group might

have been tested more frequently than the other, tested with higher stakes (i.e., greater consequences), or provided one of two types of feedback—simply being made aware of their test performance or course progress, or given remediation or another type of corrective action based on their test results.

In other studies the experimental group was told that a test would be counted as part of their grade and the control group was told that it would not. In a few experiments, one group was given take-home tests during the course and the other group in-class tests. On the final exam, the in-class tested students tended to perform better, perhaps because the take-home–tested students limited their review of the material to the topics on the take-home tests, whereas the in-class–tested students prepared themselves to be tested on all topics.

A small minority of studies employed pre-post comparisons of either the same population (e.g., 4th-graders one year and 4th-graders in the same jurisdiction a later year), the same cohort (e.g., 4th-graders followed longitudinally from one year to a later year), or a synthetic cohort (e.g., a sample of 4th-graders one year compared to a sample of 8th-graders in the same jurisdiction four years later).

Historically, effect-of-testing research has accompanied interest in one or another of the specific types of interventions. Mastery testing studies, for example, were frequent from the 1960s through the 1980s. Accountability studies coincided with the popularity (in the United States, at least) of "minimum competency" testing from the 1970s on. Frequency-of-testing studies were most popular in the first half of the twentieth century. Memory retention studies were frequently conducted then, too, but have also encountered resurgence in popularity in recent years.

Table 1 describes this collection of quantitative studies by various attributes, including study design, and the source, sponsor, and scale of tests used in the studies.

Survey studies. Surveys are a type of quantitative study that measures perceptions of effects—either through public opinion polls or surveys of groups selected to complete a questionnaire as part of a program evaluation. Some survey studies contain multiple relevant items and/or posed relevant questions to multiple respondent groups (e.g., teachers, parents). Two hundred forty-seven survey studies conducted in the United States and Canada were reviewed summarizing about 700,000 separate, individual responses to questions regarding testing's effect on learning and instruction and preferences for common test types.

All told, 813 individual item-response group combinations (hereafter called "items") were identified. For example, if a polling firm posed the same question to two different respondent groups (e.g., teachers, parents), each item-response group combination was counted as a separate item. As was inevitable, some individual respondents are represented more than once, but never for the same item.

Figure 1 breaks out survey items by data source—public opinion poll or survey embedded in a program evaluation—and respondent group—education provider

TABLE 1
Source, Sponsor, Study Design, and Scale of Tests in Quantitative Studies

	Number of Studies	Percent of Studies
Source of Test		
Researcher or Teacher	87	54
Commercial	38	24
National	24	15
State or District	11	7
Total	**160**	**100**
Sponsor of Test		
Local	99	62
National	45	28
State	11	7
International	5	3
Total	**160**	**100**
Study Design		
Experiment, Quasi-experiment	107	67
Multivariate	26	16
Pre-post	12	8
Pre-post (with shadow test)	8	5
Experiment, Posttest only	7	4
Total	**160**	**100**
Scale of Test Administration		
Classroom	115	72
Large-scale	39	24
Mid-scale	6	4
Total	**160**	**100**

or education consumer. The education provider group comprises administrators, board members, teachers, counselors, and education professors. The education consumer group comprises the public, parents, students, employers, politicians, and tertiary education faculty.

Figure 1 reveals a fairly even division of sources among the 800+ survey items between public opinion polls (55%) and program evaluation surveys (45%). Likewise, an almost even division of respondent group types between education providers (48%) and education consumers (52%) was found.

Among the subtotals, however, asymmetries emerge. Almost five times as many evaluation survey items were posed to education providers than to education consumers. Conversely, almost three times as many public opinion poll items were posed to education consumers than to education providers.

Table 2 reveals that a majority of the survey items (62%) concern high-stakes tests (i.e., tests with consequences for students, teachers, or schools, such as retention in grade [for a student] or licensure denial [for a teacher]). This stands to

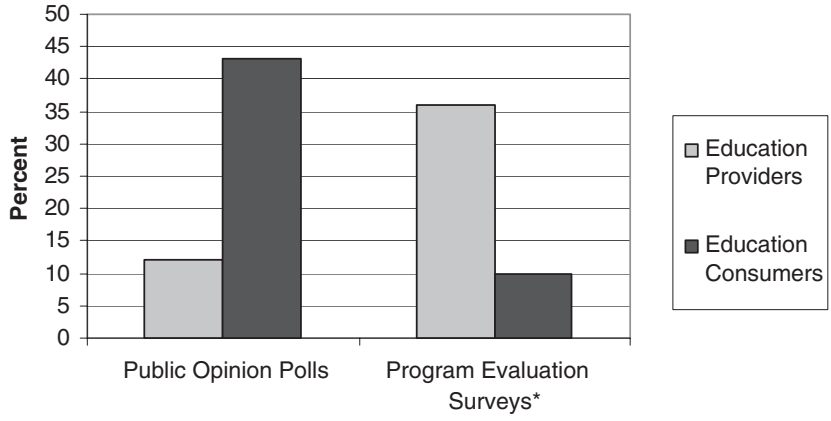

FIGURE 1
Percentage of survey items, By respondent group and type of survey.
Note. *Three program evaluation survey items were posed to a combined provider-consumer group, and for the purpose of this figure, those three items have been counted twice.

reason, as high-stakes tests are more politically controversial and more familiar to the public, so pollsters and program evaluators are more likely to ask about them.

Table 2 also classifies the survey items by the target of the stakes—the group, or groups, bearing the consequences of good or poor test performance. For a plurality of items, students bore the consequences (46%). Schools (33%) and

TABLE 2
Number and Percent of Survey Items, By Test Stakes and Their Target Group

	Number of Survey Items	Percent of Survey Items
Level of Stakes		
High	507	62
Medium	184	23
Unknown	89	11
Low	33	4
Total	**813**	**100**
Target Group		
Students	393	46
Schools	281	33
Teachers	116	14
No Stakes	64	7
Total	**854**	**100**

teachers (14%) were less often the target of test stakes. In some cases, stakes were applicable to more than one group.

Qualitative studies. Any study for which an effect size could not be computed was designated as qualitative. Typically, qualitative studies declare evidence of effects in categorical terms and incorporate more "hands on" research methods, such as direct observations, site visits, interviews, or case studies. Here, analysis of 244 qualitative studies focuses exclusively on the reported direction of the testing effect on achievement.

To determine whether a qualitative study indicated a positive, neutral, or negative effect of testing on achievement, a specific qualitative meta-analysis approach called template analysis was employed (e.g., Crabtree & Miller, 1999; King, 1998, 2006; Miles & Huberman, 1994). A hierarchical coding rubric was developed to summarize and organize the studies according to themes predetermined as important by previous research and informed by ongoing analysis of the data set. The template establishes broad themes (e.g., stakes) initially; then successively smaller themes (e.g., high, medium, or low) are generated within each broad theme. The template rubric comprised 6 themes and 27 subthemes identified in Table 3.

All studies were reviewed and coded by three reviewers: a doctoral student in meta-analysis, an editor, and me. Any disparities in coding were discussed among reviewers in an attempt to reach a consensus. In less than 10 cases in which a clear consensus was not reached, the author made the coding decisions.

Qualitative studies used various research methods, and in various combinations. Indeed, 37 of the 245 studies incorporated more than one method (Table 4).

Table 5 categorizes the qualitative studies according to the level of education and the scale, stakes, and target of the stakes of the tests involved. Most (81%) studies included in this analysis were large-scale, while 17% of the studies were at the classroom level. Two percent of the studies focused on testing for teachers.

Summary. Table 6 lists the overall numbers of studies and effects by type—quantitative, survey, and qualitative—and geographic coverage—US or non-US. Several studies fit more than one category (e.g., a program evaluation with both survey and observational case study components).

Calculating Effects

For quantitative and survey studies, effect sizes were calculated and the data summarized with weighted and unweighted means. Qualitative studies that reached a conclusion about the effect of testing on student achievement or on instruction were tallied as positive, negative, and in-between.

TABLE 3
Coding Themes and Subthemes Used in Template Analysis

Themes	Subthemes
Scale of Test	• Large-Scale • Classroom • Test for Teachers • Not Specified
Stakes of Test	• High • Medium • Low • Varies • Not Specified
Target of Stakes	• Student • Teacher • School
Research Methods	• Case Study • Interview/Focus Group • Journal/Work Log • Records/Document Review • Research Review • Experiment/Pre-Post Comparison for which No Effect Size Was Reported • Survey for which No Effect Size Was Reported
Rigor of Qualitative Study	• High • Medium • Low
Direction of Testing Effects	• Positive • Positive Inferred • Mixed • No Change • Negative

TABLE 4
Research Method Used in Qualitative Studies

Themes/Subthemes	Number of Studies	Percent of Studies
Case Study	120	43
Interview (Includes Focus Group)	75	27
Records or Document Review	33	12
Survey	22	8
Experiment or Pre-Post Comparison	21	7
Research Review	8	3
Journals or Work Logs	2	1
Total	**281**	**101**

Note. Percentage sums to greater than 100 due to rounding.

TABLE 5
Level of Education, Scale, Stakes, and Target of Stakes in Qualitative Studies

	Number of Studies	Percent of Studies
Level of Education		
Two or More Levels	115	50
Upper Secondary	55	24
Elementary	34	15
Postsecondary	18	8
Lower Secondary	7	3
Adult Education	3	1
Total	**232**	**101**
Scale of Test		
Large-scale	196	81
Classroom	42	17
Test for Teachers	5	2
Total	**243**	**100**
Stakes of Test		
High	154	63
Low	51	21
Medium	33	14
Varies/Not Specified	6	3
Total	**244**	**101**
Target of Stakes		
Student	150	62
School	80	33
Teacher	10	4
Varies	1	<1
Total	**241**	**100**

Note. Percentages may sum to greater than 100 due to rounding.

TABLE 6
Number of Studies and Effects, By Type

Type of Study	Number of Studies	Number of Effects	Population Coverage (in thousands)	Percent of Studies with US Focus
Quantitative	177	640	7000	89
Surveys & Polls	247	813	700	95
Qualitative	245	245	unknown	65
Total	**669**	**1698**		

Several formulae are employed to calculate effect sizes, each selected to align to its relevant study design. The most frequently used is some variation of the standardized mean difference (Cohen, 1988):

$$d = \frac{\bar{X}_T - \bar{X}_C}{sp} \tag{1}$$

where:

X_T = number in treatment group,
X_C = number in control group,
S_p = pooled standard deviation, and
d = effect size.

For those studies that compare proportions rather than means, an effect size comparable to the standard mean difference effect size is calculated with a log-odds ratio and a logit function transformation (Lipsey & Wilson, 2001, pp. 53–58):

$$ES = \left(\log_e \left[\frac{P_T}{1 - P_T} \right] - \log_e \left[\frac{P_C}{1 - P_C} \right] \right) \Big/ 1.83 \tag{2}$$

P_T is the proportion of persons in the treatment group with "successful" outcomes (e.g., passing a test, graduating), and P_C is the proportion of persons in control group with successful outcomes. The logit function transformation is accomplished by dividing the log-odds ratio by 1.83.

For survey studies, the response pattern for each survey is quantitatively summarized in one of two ways:

- as frequencies for each point along a Likert scale, which can then be summarized by standard measures of central tendency and dispersion; or
- as frequencies for each multiple choice response option, which can then be converted to percentages.

Just over 100 of the almost 800+ survey items reported their responses in means and standard deviations on a scale. Effects for scale items are calculated as the difference of the response mean from the scale midpoint divided by the standard deviation.

Close to 700 other survey items with multiple-choice response formats reported the frequency and percentage for each choice. The analysis collapses these multiple choices into dichotomous choices—yes and no, for and against, favorable and unfavorable, agree and disagree, support and oppose, etc. Responses that fit neither side—neutral, no opinion, don't know, no response—are not counted. An effect

size is then calculated with the aforementioned log-odds ratio and a logit function transformation.

Weighting. Effect sizes are summarized in two ways: with an unweighted mean, in which case each effect size counts the same regardless the size of the population under study, and with a weighted mean, in which case effect size measures calculated on larger groups count for more.

For between 10% and 20% of the quantitative studies (depending on how one aggregates them) that incorporated pre-post comparison designs, weights (i.e., inverse variances) are calculated thus (Lipsey & Wilson, 2001, p. 72):

$$w = \frac{2N}{4(1-r)+d^2} \tag{3}$$

where:

N = number in study population,
r = correlation coefficient, and
d = effect size.

For the remainder of the studies with standardized mean difference effect sizes, weights (i.e., inverse variances) are calculated thus (Lipsey & Wilson, 2001, p. 72):

$$w = \frac{2 \times N_T \times N_C(N_T + N_C)}{2(N_T + N_C)^2 + N_T \times N_C \times d^2} \tag{4}$$

where:

N_T = number in treatment group,
N_C = number in control group, and
d = effect size.

For the group of about 100 survey items with Likert-scale responses with standardized mean effect sizes, standard errors and weights (i.e., inverse variances) are calculated thus:

$$SE = \frac{s}{\sqrt{n}} \qquad w = \frac{n}{s^2} \tag{5, 6}$$

where:
s = standard deviation and
n = number of respondents.

For the group of about 700 items with proportional-difference responses with log-odds-ratio effect sizes, standard errors and weights are calculated thusly:

$$w = \frac{1}{SE^2} \qquad SE = \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}} \tag{7, 8}$$

where:

a = number of respondents in favor,
b = number of respondents not in group a,
c = number of respondents against, and
d = number of respondents not in group c.

Once the weights and standard errors are calculated, effect sizes are then "weighted" by the following process (Lipsey & Wilson, 2001, pp. 130–132):

1. Each effect size is multiplied by its weight: *wes*
2. The weighted mean effect size is calculated thusly:

$$\overline{ES} = \frac{\sum wes_i}{\sum w_i} \tag{9}$$

Standards for judging effect sizes. Cohen (1988) proposed that a $d < 0.20$ represented a small effect and a $d > 0.80$ represented a large effect. A value for d in between, then, represented a medium effect. In his meta-analysis of meta-analyses, however, Hattie (2009) placed the average d for all meta-analyses of student achievement effect studies near 0.40, with a pseudo-Hawthorne Effect base level around 0.10 (i.e., the level at which any achievement-related intervention not producing a clear positive or negative effect will settle).

RESULTS

Quantitative Studies

For quantitative studies, simple and weighted effect sizes are calculated for several different aggregations (e.g., same treatment group, control group, or study author). Mean effect sizes range from a moderate $d \approx 0.55$ to a fairly large $d \approx 0.88$ depending on the way effects are aggregated or effect sizes are adjusted for study artifacts. Testing with feedback produces the strongest positive effect on

achievement. Adding stakes or testing with greater frequency also strongly and positively affects achievement.

Studies are classified with dozens of moderators, including size of study population, scale of test administration, responsible jurisdiction (e.g., nation, state, classroom), level of education, and primary focus of study (e.g., memory retention, testing frequency, mastery testing, accountability program).

For the full panoply of 640 different effects, the "bare bones" mean $d \approx 0.55$. After adjustments for small sample sizes (Hedges, 1981) and measurement uncertainty in the dependent variable (Hunter & Schmidt, 2004, chapter 8), the mean $d \approx 0.71$, an apparently stronger effect.[5]

For different aggregations of studies, adjusted mean effect sizes vary. Grouping studies by the same treatment or control group or content, with differing outcome measures ($N \approx 600$) produces a mean $d \approx 0.73$. Grouping studies by the same treatment or control group, with differing content or outcome measures ($N \approx 430$) produces a mean $d \approx 0.75$. Grouping studies by the study author, such that a mean is calculated across all of a single author's studies ($N = 160$), produces a mean $d \approx 0.88$.

To illustrate the various moderators' influence on the effect sizes, the same-study-author aggregation is employed. Studies in which the treatment group tested more frequently than the control group produced a mean $d \approx 0.85$. Studies in which the treatment group was tested with higher stakes than the control group produced a mean $d \approx 0.87$. Studies in which the treatment group was made aware of their test performance or course progress (and the control group was not) produced a mean $d \approx 0.98$. Studies in which the treatment group was remediated, or received some other type of corrective action, based on their test performance produced a mean $d \approx 0.96$. (See Table 7.)

Note that these treatments, such as testing more frequently or with higher stakes, need not be mutually exclusive, can be used complementarily, and often are. This suggests that the optimal intervention would not choose among these interventions—all with strong effect sizes—but, rather use as many of them as possible at the same time.

TABLE 7
Mean Effect Sizes for Various Treatments

Treatment Group . . .	Mean Effect Size
. . . is made aware of performance and control group is not.	0.98
. . . receives targeted instruction (e.g., remediation).	0.96
. . . is tested with higher stakes than control group.	0.87
. . . is tested more frequently than control group.	0.85

TABLE 8
Source, Sponsor, Study Design, and Scale in Quantitative Studies

	Number of Studies	Mean Effect Size
Source of Test		
Researcher or Teacher	87	0.93
National	24	0.87
Commercial	38	0.82
State or District	11	0.72
Total	**160**	
Sponsor of Test		
International	5	1.02
Local	99	0.93
National	45	0.81
State	11	0.64
Total	**160**	
Study Design		
Pre-post	12	0.97
Experiment, Quasi-experiment	107	0.94
Multivariate	26	0.80
Experiment, Posttest only	7	0.60
Pre-post (with shadow test)	8	0.58
Total	**160**	
Scale of Analysis		
Aggregated	9	1.60
Small-scale	118	0.91
Large-scale	33	0.57
Total	**160**	
Scale of Test Administration		
Classroom	115	0.95
Mid-scale	6	0.72
Large-scale	39	0.71
Total	**160**	

Mean effect sizes can be compared between other pairs of moderators, too (see Table 8). For example, small population studies, such as experiments, produce an average effect size of (0.91), whereas large population studies, such as those using national populations, produce an average effect size of (0.57). Studies of small-scale test administrations (0.92) produce a larger mean effect size than do studies of large-scale test administrations (0.78). Studies of local test administrations produce a larger mean effect size (0.93) than do studies of state test administrations (0.64).

Studies conducted at universities (which tend to be small-scale) produce a larger mean effect size (1.02) than do studies conducted at the elementary-secondary level (0.76) (which are often large-scale). Studies in which the intervention occurs

before the outcome measurement produce a larger mean effect size (0.93) than do "backwash" studies, for which the intervention occurs afterwards (0.40) (e.g., when an increase in lower-secondary test scores is observed after the introduction of a high-stakes upper-secondary examination). Studies in which the subject matter content of the intervention and outcome are aligned produce a larger mean effect size (0.91) than do studies in which they are not aligned (0.78).

The quantitative studies vary dramatically in population size (from just 16 to 1.1 million), and their effect sizes tend to correlate negatively. The mean for the smallest quintile among the full 640 effects, for example, is 1.04, whereas that for the largest quintile is 0.30. Starting with the adjusted overall mean effect size of 0.71 (for all 640 effects), the mean grows steadily larger as the studies with the largest populations are removed. For all studies with populations less than 100,000, the mean effect size is 0.73; under 10,000, 0.80; under 1,000, 0.83; under 500, .88; under 100, .92; and under 50, 1.06.

Generally, weighting reduces mean effect sizes, further suggesting that the larger the study population, the weaker the results. For example, the unweighted effect size of 0.30 for the largest study population quintile compares to its weighted counterpart of 0.28. For the smallest study population quintile, the effect size is 1.04 unweighted but only 0.79 weighted.

Survey Studies

Extracting meaning from the 800+ separate survey items required condensation. First, survey items were separated into two overarching groups in which the perception of a testing effect is either explicit or inferred. Within the explicit, or directly observed, category, survey items were further classified into two item types: those related to improving learning or to improving instruction.

Among the 800+ individual survey items analyzed, few were worded exactly like others. Often the meaning of two different questions, or prompts, is very similar, however, even though the wording varies. Consider a question posed to US teachers in the state of North Carolina in 2002: "Have students learned more because of the [testing program]?" Eighty-three percent responded favorably and 12% unfavorably, while 5% were neutral (e.g., no opinion, don't know). This survey item is classified in the "improve learning" item group.

Now consider a question posed to teachers in the Canadian Province of Alberta in 1990: "The diploma examinations have positively affected the way in which I teach." Forty-one percent responded favorably, 32% unfavorably, and the remaining were neutral (or unresponsive). This survey item is classified in the "improve instruction" item group.

Second, the many survey items for which the respondents' perception of a testing effect was inferred were separated into several item groups, including: "Favor grade promotion exams," "Favor graduation exams," "Favor teacher competency

exams," "Favor teacher recertification exams," "Favor more testing," Hold schools accountable for students' scores," and "Hold teachers accountable for students' scores."

A perception of a testing effect is inferred when an expression of support for a testing program derives from a belief that the program is beneficial. Conversely, a belief that a testing program is harmful is inferred from an expression of opposition.

For the most basic effects of testing (e.g., improves learning or instruction) and the most basic uses of tests (e.g., graduation, certification) effect sizes are large, in many cases more than 1.0 and in some cases more than 2.0. Effect sizes are weaker for situations in which one group is held accountable for the performance of another—holding either teachers or schools accountable for student scores.

The results vary at most negligibly with study level of rigor or test stakes. And, the results are overwhelmingly positive whether the survey item focused on testing's affect on improving learning or improving instruction. Some of the study results can be found in the following tables. Table 9 lists the unweighted and weighted effect sizes by item group for the total sample respondent population. With the exception of the item group "hold teachers accountable for student scores," effect sizes are positive for all item groups. The most popular item groups appear to be: "Favor teacher recertification exams;" "Favor teacher competency exams;" "Favor (student) graduation exams;" and "Favor (student) grade promotion exams."

Effect sizes—weighted or unweighted—for these four item groups are all positive and large (i.e., > 0.8) (Cohen, 1988). The effect sizes for other item groups, such as "improve learning" and "improve instruction," also are positive and large whether weighted or unweighted. The remaining two item groups—"favor more testing" and "hold schools accountable for (student) scores"—garner moderately large positive effect sizes whether weighted or unweighted.

Table 8 also breaks out the effect size results by two different groups of respondents, education providers and consumers. Education providers comprise primarily administrators and teachers whereas education consumers comprise mostly students and noneducator adults.

As groups, providers and consumers tend to respond similarly—strongly and positively—to items about test effects (tests "improve learning" and "improve instruction"). Likewise, both groups strongly favor high-stakes testing for students ("favor grade promotion exams" and "favor graduation exams").

Provider and consumer groups differ in their responses in the other item groups, however. For example, consumers are overwhelmingly in favor of teacher testing, both competency (i.e., certification) and recertification exams, with weighted effect sizes of 2.2 and 2.0, respectively. Provider responses are positive but not nearly as strong, with weighted effect sizes of 0.3 and 0.4, respectively.

Clear differences of opinion between provider and consumer groups can be seen among the remaining item groups, too. Consumers tend to favor more testing

TABLE 9
Effect Sizes of Survey Responses, By Item Group, For Total Population, Education Providers, and Education Consumers: 1958–2008[1]

	Total			Education Providers[2]			Education Consumers[3]		
Item Type	Effect Size (Unweight)	Effect Size (Weight)	Number of Items	Effect Size (Unweight)	Effect Size (Weight)	Number of Items	Effect Size (Unweight)	Effect Size (Weight)	Number of Items
Testing Improves Learning	1.3	0.5	242	1.2	0.5	144	1.3	0.6	98
Testing Improves Instruction	1.0	0.7	165	1.0	0.6	125	1.2	1.0	40
Testing Aligns Instruction				0.6	0.0	28			
Favor Grade Promotion Exams	1.3	1.1	81	1.2	1.0	26	1.3	1.1	53
Favor Graduation Exams	1.4	1.1	121	1.2	0.8	38	1.4	1.2	81
Favor More Testing	0.5	0.7	72	−0.2	0.1	19	0.7	0.8	53
Favor Teacher Competency Exams	2.1	0.8	61	1.1	0.3	17	2.5	2.2	44
Favor Teacher Recertification Exams	2.1	1.5	22				2.5	2.0	18
Hold Teachers Accountable for Scores	−0.1	0.0	31	−0.6	−0.5	18	0.6	0.5	13
Hold Schools Accountable for Scores	0.6	0.5	111	0.2	−0.3	38	0.9	0.7	73

Notes. 1. Blank cells represent numbers of items too small to report validly.

2. Education providers are education administrators, principals, teachers, education agency officials, and university faculty in education programs.

3. Education consumers are students, parents, the general public, employers, public officials not working in education agencies, and university faculty outside of education.

TABLE 10
The Effect of Testing on Student Achievement in Qualitative Studies

Direction of Effect	Number of Studies	Percent of Studies	Percent w/o Inferred
Positive	204	84	93
Positive Inferred	24	10	
No Change	8	3	4
Mixed	5	2	2
Negative	3	1	1
Total	**244**	**100**	**100**

and holding teachers and schools accountable for student test scores. Providers either are less supportive in these item groups or are opposed. For example, providers' weighted effect size for "hold teachers accountable for students scores" is a moderately large –0.5.

Qualitative Studies

Analysis of the qualitative studies focuses on the direction of the testing effect on achievement or instruction. Ninety-three percent of the qualitative studies analyzed reported positive effects of testing, whereas only 7% reported mixed effects, negative effects, or no change.

In 24 cases, a positive effect on student achievement was not declared but reasonably could be inferred from statements about behavioral changes. One study, for example, reported "[using] results from test to improve coursework." If coursework was, indeed, improved one might reasonably assume that student achievement benefited as well. Whether the "positive inferred" studies are included in the summary or not, however, the counts indicate that far more studies find positive than mixed, negative, or no effects. The main results of this research summary are displayed in Table 10.

DISCUSSION

One hundred years' evidence suggests that testing increases achievement. Effects are moderately to strongly positive for quantitative studies and are very strongly positive for survey and qualitative studies.

The overwhelmingly positive results of the qualitative research review, in particular, may surprise some readers. The results should be considered reliable because the base of evidence is so large—245 studies conducted over the course of a century in more than 30 countries.

Qualitative studies have been held up by testing opponents as a higher standard for studies of educational impact. Labeling quantitative studies as narrow, limited,

or biased, they have pointed toward qualitative studies for an allegedly broader view. But, the qualitative studies they cite have not investigated the effect of tests on student achievement. Rather, they have focused on popular complaints about tests, such as "teaching to the test," "narrowing the curriculum," and the like. Indeed, many of those studies have not considered any possible positive effects and looked only for effects that would be considered negative.

Some believers in the superiority of qualitative research have also pressed for the inclusion of "consequential validity" measures in testing and measurement standards, perhaps believing that such would reliably show testing to be on balance negative. Results here show that such beliefs are unwarranted.

Other researchers have asserted that there exists no evidence of testing benefits using any methodology, or at least no evidence for high-stakes educational testing (Phelps, 2005). The "paucity of research" belief has spread widely among researchers of all types and ideological persuasions.

Such should be difficult to believe after this review of the research, however. Not all the studies reviewed here found testing effects, and not all of the effects found have been positive. But, studies finding positive effects on achievement exist in robust number, greatly outnumber those finding negative effects, and date back a hundred years.

Picking Winners

There may be a temptation for education program planners to pick out an intervention producing the "top" effect size and focus all resources and attention on it. As noted earlier, studies in which the treatment group was made aware of their test performance or course progress (and the control group was not) produced a higher mean effect size than did those in which the treatment group was remediated (or received some other type of targeted instruction). Those studies, in turn, produced a higher mean effect size than did those in which the treatment group was tested with higher stakes which, in turn, produced a higher mean effect size than did those studies in which the treatment group was tested more frequently than was the control group.

But, more important than ranking these treatments is to notice that all of them produce robustly large effect sizes and they are not mutually exclusive. They can be used complementarily, and often are. This suggests that the optimal intervention would not choose among interventions but, rather, use as many of them as is practical simultaneously.

Varying Results by Scale

Among the quantitative studies, those conducted on a smaller-scale tend to produce stronger effects than do large-scale studies. There are several possible explanations

for this. First, larger studies often contain more "noise"—information that may not be clearly relevant to testing or achievement that muddles the picture. Second, unlike designed experiments, larger studies are usually not designed to test the study hypothesis but, rather, are happenstance accounts of events with their own schedule and purpose. Third, the input and outcome measures in larger studies often are not aligned. Take, for example, the common practice of using a single-subject monitoring test (e.g., in mathematics or English) to measure the effect of an accountability standard that employs a full battery graduation examination and course distribution and completion requirements.

Those who judge the effect of testing on achievement exclusively from large-sample multivariate studies deprive themselves of the most focused, clear, and precise evidence. Some researchers, for example, have claimed that no studies of "test-based accountability" had been conducted before theirs in the 2000s (Phelps, 2009, pp. 114–117). This summary includes 24 studies completed before 2000 whose primary focus was to measure the effect of "test-based accountability." A few dozen more pre-2000 studies also measured the effect of test-based accountability although such was not their primary focus. Include qualitative studies and program evaluation surveys of test-based accountability, and the count of pre-2000 studies rises into the hundreds.

Still, the issue of scale may present a conundrum for educational planners. By necessity, some, and probably most, educational testing must be of large scale. So, can planners incorporate the "small-scale" benefits of testing—such as mastery testing, targeted instruction, other types of feedback, and, perhaps, even targeted rewards and sanctions—into large-scale test administrations? Thankfully, some intrepid scholars are working on that very issue (see, for example, Leighton, 2008/2009).

The Importance of Surveys

Even if survey data are not the best source of evidence for actual changes in student achievement due to testing, they are certainly a good source of evidence of respondents' preferences. And, in democratic societies, preferences matter. Indeed, even if other quantitative studies reported negative effects on student achievement from testing, political leaders could choose to follow the public's preference for testing anyway. Indeed, public opinion polls sometimes represent the best available method for learning the public's wishes in democracies. Expensive, highly controlled and monitored political elections do not always attract a representative sample of the citizenry (Keeter, 2008).

The effect sizes for survey items on the most basic effects of testing (improves learning or instruction) and the most basic uses of tests (grade promotion, graduation, certification, and recertification) are very large, larger than those from other quantitative studies. A skeptic might interpret this difference in mean effect sizes

as evidence that the public exaggerates the positive effect of testing. Or, one might admit that other quantitative studies cannot possibly capture all the factors that matter and, so, might underestimate the positive effect of testing.

The "bottom line" result of the review of survey studies is very strong support, at least in the United States and Canada, for the most basic forms of testing (that hold educators and students accountable for their own performance), regardless the respondent group or stakes. Moreover, the results of this study might be considered reliable because the base of evidence is massive—almost three-quarters of a million individual responses.

NOTES

1. Testing critics often characterize alignment as deleterious—"narrowing the curriculum" and "teaching to the test" are commonly heard phrases. But, when the content domains of a test match a jurisdiction's required content standards, aligning a course of study to the test is eminently responsible behavior.
2. In 2007, the United States' population represented about 73% of the population of all OECD countries whose primary language is English (i.e., Australia, Canada [primarily English-speaking population only], Ireland, New Zealand, United Kingdom, United States). Source: *OECD Factbook 2010*.
3. Of the documents set aside, a few hundred provide evidence of achievement effects in a manner excluded from this study. This "off focus" research includes studies of non-test incentive programs (e.g., awarding prizes to students or teachers for achievement gains), "opt-out" tests (e.g., passing a test allows a student to skip a required course, thus saving time and money), benefit-cost analyses, effective schools or curricular alignment studies that did not isolate testing's effect, teacher effectiveness studies, and conceptual or mathematical models lacking empirical evidence.
4. The actual effect size was calculated, for example, comparing two groups' posttest scores (using pretests or other covariates if assignment was not random), two groups' post-pre-test gain scores, the same group's posttest or gain scores on tests of two different types, or, less frequently, by retrospective matching and pre-post comparison.
5. One benefit of adjusting effect sizes for measurement uncertainty is to, at least in relative terms, reduce any "test-retest" reliability effect on the effect size. However, only a tiny minority of the experimental studies, and a small minority of all studies, used the same test pre and post.

REFERENCES

Bertsch, S., Pesta, B. J., Wiscott, R., & McDaniel, M. A. (2007). The generation effect: A meta-analytic review. *Memory & Cognition*, *35*(2), 201–210.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (Rev. Ed.). New York, NY: Academic Press.

Crabtree, B. F., & Miller, W. L. (1999). Using codes and code manual. A template organizing style of interpretation. In B. F. Crabtree & W. L. Miller (Eds.), *Doing qualitative research* (2nd ed., pp. 163–178). Thousand Oaks, CA: Sage.

Hattie, J. A. C. (2009). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*. London, UK: Routledge.

Hedges, L. V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational Statistics*, *6*, 107–128.

Hunter, J. E., & Schmidt, F. L. (2004). *Methods of meta-analysis: Correcting error and bias in research findings*. Thousand Oaks, CA: Sage.

Keeter, S. (2008, Autumn). Poll power. *The Wilson Quarterly*, 56–62.

King, N. (1998). Template analysis. In G. Symon & C. Cassell (Eds.), *Qualitative methods and analysis in organizational research: A practical guide* (pp. 118–134). London, UK: Sage.

King, N. (2006). *What is template analysis?* University of Huddersfield School of Human and Health Sciences. Retrieved from http://www2.hud.ac.uk/hhs/research/template_analysis/index.htm

Leighton, J. P. (2008/2009). Mistaken impressions of large-scale cognitive diagnostic testing. In R. P. Phelps (Ed.), *Correcting fallacies about educational and psychological testing* (pp. 219–246). Washington, DC: American Psychological Association.

Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis*. Thousand Oaks, CA: Sage.

Miles, M. B., & Huberman, A. M. (1994). *Qualitative data analysis: An expanded sourcebook*. Beverly Hills, CA: Sage.

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

Phelps, R. P. (2009). Education achievement testing: Critiques and rebuttals. In R. P. Phelps, (Ed.), *Correcting fallacies about educational and psychological testing* (pp. 89–146). Washington, DC: American Psychological Association.