Review of Educational Research

http://rer.aera.net

Is Test-Driven External Accountability Effective? Synthesizing the Evidence From Cross-State Causal-Comparative and Correlational Studies

Jaekyung Lee

REVIEW OF EDUCATIONAL RESEARCH 2008; 78; 608
DOI: 10.3102/0034654308324427

The online version of this article can be found at:
http://rer.sagepub.com/cgi/content/abstract/78/3/608

Published on behalf of

http://www.aera.net

By

http://www.sagepublications.com

Additional services and information for *Review of Educational Research* can be found at:

Email Alerts: http://rer.aera.net/cgi/alerts

Subscriptions: http://rer.aera.net/subscriptions

Reprints: http://www.aera.net/reprints

Permissions: http://www.aera.net/permissions

Downloaded from http://rer.aera.net at SUNY AT BUFFALO on May 1, 2009

Review of Educational Research
September 2008, Vol. 78, No. 3, pp. 608–644
DOI: 10.3102/0034654308324427
© 2008 AERA. http://rer.aera.net

Is Test-Driven External Accountability Effective? Synthesizing the Evidence From Cross-State Causal-Comparative and Correlational Studies

Jaekyung Lee
State University of New York at Buffalo

In the midst of keen controversies on the impact of high-stakes testing and test-driven external accountability policy, the more balanced and careful selection, interpretation, and use of scientific research evidence are crucial. This article offers a critical synthesis of cross-state causal-comparative and correlational studies that explored the effects of test-driven external accountability policies on reading and math achievement. A meta-analysis of 76 effect-size estimates from 14 selected studies showed a modestly positive effect on average but no significant effect on the racial achievement gap. Nevertheless, this review calls for further evidence on the policy-outcome linkage, revealing limitations, uncertainties, and inconsistencies in many findings. The author explores variations among the studies in terms of independent and dependent variables, analytical samples and methods, and the reporting of statistical and practical significance. Implications for accountability policy and research under the No Child Left Behind Act are discussed.

KEYWORDS: high-stakes testing, accountability, achievement, NAEP, meta-analysis.

Although the origin of public high-stakes testing dates as far back as the third century B.C., when civil service exams were used by the Han emperors of China (206 B.C. to A.D. 220), modern external examinations have grown as instruments of control over educational systems in many countries (Eckstein & Noah, 1993). High-stakes tests in American school systems are relatively recent, and they are embedded in national educational and social contexts. Although states relied more on basic skills tests in the 1970s, the report *A Nation at Risk* (National Commission on Excellence in Education, 1983) called for an end to the minimum competency testing movement (Amrein & Berliner, 2002). As the focus of high-stakes testing policy has shifted from minimum competency to proficiency, an increasing number of states have held schools and teachers accountable for test results over the past two decades. The culmination of this policy shift is seen in the most recent federal educational policy initiative, the No Child Left Behind Act of 2001 (NCLB), which is aimed at accomplishing high academic standards for all students and closing their achievement gaps.

Downloaded from http://rer.aera.net at SUNY AT BUFFALO on May 1, 2009

There are controversies about whether external, test-driven accountability policy enhances or hinders academic achievement. The case that drew the most attention was that of Texas, where the evidence on the effects of high-stakes testing on equity was mixed and often contradictory (Carnoy, Loeb, & Smith, 2001; Grissmer & Flanagan, 1998; Grissmer, Flanagan, Kawata, & Williamson, 2000; Haney, 2000; Ladd, 1999; Skrla, Scheurich, Johnson, & Koschoreck, 2004; Valencia, Valenzuela, Sloan, & Foley, 2004). Although NCLB builds on the alleged success of first-generation accountability states such as Texas and North Carolina, which had adopted test-based accountability systems prior to NCLB, assessing its impact requires more rigorous scrutiny of new evidence from the National Assessment of Educational Progress (NAEP) and state assessment results beyond a single state. Moreover, studies that evaluated the effect of high-stakes testing on the basis of each state's own performance standards suffered from threats to generalization because of a lack of comparability of results with other states and potential risk for gain score inflation over time.

Past literature reviews of the effects of high-stakes testing and accountability have several limitations, generating more questions than answers (see Harris & Herrington, 2006; Heubert, 2000; Heubert & Hauser, 1999; Kirkland, 1971; Langenfeld, Thurlow & Scott, 1997; Phelps, 2005). First, the reviews have tended to be descriptive rather than meta-analytic. Second, the reviewers were highly inclusive in their selection of relevant studies. A test was considered high stakes if its results had perceived or real consequences for students, staff members, or schools (Madaus, 1988). By including research based on this broad definition of high-stakes testing, the reviewers raised the issue of the comparability of study findings. Third, the studies included in past reviews did not fully capture recent changes in testing requirements and accountability policies; the studies examined mostly minimum competency tests, featuring their emphasis on basic skills and using students as the primary target of accountability. Finally, the studies in past reviews were restricted mostly to samples from single states or localities. Therefore, it is necessary to review emerging research evidence on the effects of new school accountability policies across states and to better inform the current educational policy debate under NCLB.

Critical premises on which the movement of test-driven external accountability is based are weak. What are the social consequences we may face in this country if school accountability is based on false premises about students' test score gains? What are the implications of emerging research evidence for educational policy and practice? This review focuses on cross-state causal-comparative and correlational studies that used secondary analyses of national assessment data to evaluate the effects of external test-driven accountability on reading and/or mathematics achievement. This review not only synthesizes findings through a meta-analysis of the "effect-size" estimates of multiple studies but also examines differences among the studies to account for variations in their findings. Methodological limitations of the studies are discussed, and some reanalyses are conducted to gain further insights into the issues. Because the studies have produced mixed findings and tend to polarize between the extremes, it is crucial to synthesize the research findings and understand the nature and degree of their variations.

Downloaded from http://rer.aera.net at SUNY AT BUFFALO on May 1, 2009

Conceptual and Analytical Framework

Context of Accountability Policy Research

Accountability often has multiple meanings and purposes, and there are several models of educational accountability (see Adams & Kirst, 1999; Darling-Hammond, 1989; Linn, 2003). The issue of who holds whom accountable and for what purpose has been contentious in the history of educational accountability (Dorn, 1998). Despite the historical debate, an accountability model that is performance driven, test driven, measurable, and statistical in nature came to dominate current policy and practice. Although an evaluation of the policy impact calls for scientific research evidence, it is necessary to understand the social and political context of high-stakes testing and accountability policy in which research has been embedded.

Public sector reform called for greater privatization, decentralization, and accountability (Osborne & Gaebler, 1992), and some ideas of the public sector reform movement, such as performance reporting and funding, spread to both the K–12 and higher education sectors (McLendon, Hearn, & Deaton, 2006). Over the past decade, many states have joined the test-driven school accountability bandwagon in the form of the "horse trade": States would grant schools and districts more flexibility in return for more accountability for academic performance (Elmore, 2002). It was appealing in principle, because governors and state legislators could take credit for improving schools without committing themselves to serious increases in funding. Moreover, these reform policies were popular because they were designed to intensify, rather than to replace, preexisting educational efforts and held out the hope of greater cost-effectiveness (Berliner & Biddle, 1995).

The new test-driven external accountability movement has changed the nature and target of high-stakes testing. Minimum competency testing in the 1970s and early 1980s shifted the burden of attaining basic skills from the state to the individual (Cohen & Haney, 1980). In contrast, the school accountability movement in the late 1980s and 1990s raised the bar to proficiency and also shifted its target to schools. Between 1985 and 1995, the number of states that used student assessment results for school accountability (i.e., school awards and recognition, performance reporting, or accreditation) increased substantially from 26 to 39. During the same period, the number of states that used test results for student accountability (i.e., student awards and recognition, promotion, or graduation) increased only slightly from 22 to 25 (Goertz, 1986; North Central Regional Educational Laboratory, 1996).

Along with the new state accountability policy movement, research on this topic has increased as well. However, previous studies of the impact of high-stakes testing on student achievement have often had several threats to validity due to their limitations: (a) a reliance on test scores from states' own assessments, which are the basis of accountability decision making and thus could cause the possible contamination of achievement gains; (b) an examination of postintervention student achievement results only, without tracking of long-term trends before and after policy enactment; and (c) an absence of control or comparison groups because of the investigation of a single state.

Under NCLB, states set the standards, choose tests to measure student performance against those standards, and hold schools accountable for the results. Under these

Downloaded from http://rer.aera.net at SUNY AT BUFFALO on May 1, 2009

circumstances, high-stakes testing works not only as an intervention but also as an instrument to measure the outcome of the intervention. On one hand, high-stakes testing generates enormous pressure for educators to improve test scores by means of narrowing the curriculum and teaching to the test. On the other hand, any inflated test scores that can result from intensive drilling and coaching under this pressure generate an illusion of real progress and give the false impression that the intervention is working. This situation will prompt more investment in high-stakes testing and further prescribed curricula. However, there are significantly smaller achievement gains when students take independent low-stakes tests such as the NAEP.

Selection Criteria

The following criteria were used in this meta-analysis to select studies for review and to determine which studies met scientific standards for evidence. First, research needed to examine states' test-driven external accountability policies, including high-stakes testing, as an independent variable. It must have involved any sort of comparative measures or classifications of state accountability policy. Second, research must have addressed achievement in reading and/or math as a dependent variable. It must have involved measures of academic achievement by using standardized tests that give comparable results across states. Studies using states' own assessments or local assessment measures were not included in this review. Third, research must have drawn on data from nationally or statewide representative samples of students so that the results could be generalized to the national or statewide target population of students. Studies using achievement measures drawn from nonrepresentative samples (such as SAT, ACT, and Advanced Placement) for cross-state comparisons were also excluded.[1] Finally, the study must have been published in a refereed journal article and reported since 1990 (i.e., during the past 18 years or so), when cross-state research on educational accountability policy has been at its peak.

First of all, true experimental research with the randomized assignment of 50 states to high-stakes testing versus low-stakes testing conditions is not possible. Nevertheless, natural interstate variations in high-stake testing policy and academic achievement provide a laboratory for ex post facto research using causal-comparative or correlational research designs. These two designs share limitations of nonexperimental methods, but the differences lie in how they operationalize and measure the policy variable; causal-comparative studies use a dichotomous measure of high-stake testing and accountability policy (yes vs. no), whereas correlational studies use a continuous measure of accountability (e.g., the degree to which rewards and sanctions apply to test results). Yet relatively few studies investigated the policy-outcome linkage at the state level. The decisive inhibiting factor has been the lack of adequate measures of state-level educational policies and achievement outcomes.

A series of cross-state causal-comparative or correlational studies attempted to undertake empirical evaluations of high-stakes testing and accountability policy. The national data those studies used included the NAEP and the National Education Longitudinal Study (NELS). Particularly, the advent of NAEP state assessment in the 1990s facilitated interstate comparisons of student achievement and policy benchmarking efforts. With the advent of NAEP state assessment as well

Downloaded from http://rer.aera.net at SUNY AT BUFFALO on May 1, 2009

as cross-state education policy surveys, research investigating whether state education policies account for students' achievement gains on the NAEP has grown.

Unlike old micro-level studies that examined policy impacts within individual states or school districts on the basis of state or local assessment results, this new generation of macro-level studies has tended to focus on states as the primary unit of analysis and compared high-stakes versus low-stakes testing states' student achievement test score trends on the basis of independent national assessments. Previous comparisons of the NAEP and state assessment results showed significant discrepancies in the level of student achievement as well as the size of statewide achievement gains (Fuller, Gesicki, Kang, & Wright, 2006; Klein, Hamilton, McCaffrey, & Stecher, 2000; Koretz & Barron, 1998; Lee, 2006b; Linn, Baker, & Betebenner, 2002). Some of the concerns raised about the use of high-stakes test results for policy evaluation do not apply to the studies that used the NAEP or other independent, low-stakes measures of achievement. The random sampling of schools and their students for testing and the lack of consequences tied to the test results may free such studies from the potential risk for contamination.[2]

Overview of Selected Studies

The search and review process identified 14 studies that meet the aforementioned selection criteria: Fredericksen (1994); Lee (1998); Grissmer and Flanagan (1998); Bishop, Mane, Bishop, and Moriatry (2001); Jacob (2001); Amrein and Berliner (2002); Carnoy and Loeb (2002); Raymond and Hanushek (2003); Rosenshine (2003); Amrein-Beardsely & Berliner (2003); Braun (2004); Lee and Wong (2004); Hanushek and Raymond (2004); and Nichols, Glass, and Berliner (2006).

Among the 14 selected studies, one of the pioneers was Fredericksen (1994), who used long-term trend NAEP data to estimate the effect of minimum competency testing on state average math achievement gain scores. The study's finding of a significantly positive effect of minimum competency testing was challenged later by Jacob (2001), who found from an analysis of NELS data that the same policy had no significant impact on 12th grade reading and math achievement. A more mixed finding was reached by Bishop et al. (2001), who reported that the effect of minimum competency testing alone was meager, but the effect of curriculum-based end-of-course exams in combination with minimum competency testing was very strong.

Grissmer and Flanagan (1998) promoted discussion of school accountability policy effects by attributing substantial achievement gains on the NAEP from 1992 to 1996 in North Carolina and Texas to those two states' challenging student performance standards and test-driven accountability policies. This study was highly speculative and did not directly estimate policy effects. Amrein and Berliner (2002) conducted a more extensive analysis of the policy-outcome linkages by tracking the performance of 18 states with high-stakes testing systems on the NAEP, SAT, and ACT. They claimed that the phenomenon of larger achievement gains in North Carolina and Texas was an "illusion arising from exclusion" (i.e., the inflation of gain scores as a result of excluding more low-achieving students from testing) and that the impact of high-stakes testing on student achievement is indeterminate.

Downloaded from http://rer.aera.net at SUNY AT BUFFALO on May 1, 2009

Amrein and Berliner's (2002) study was challenged in subsequent reanalyses of the same data by Raymond and Hanushek (2003), Rosenshine (2003), and Braun (2004). Although all three subsequent studies were very critical of the original study on methodological grounds, Braun (2004) gave a more mixed picture of the policy effect. Raymond and Hanushek and Rosenshine produced highly positive results in favor of high-stakes testing policy. Amrein-Beardsley and Berliner (2003) conducted a further analysis in their response to Rosenshine's reanalysis to support their original finding.

Carnoy and Loeb (2002) and Hanushek and Raymond (2004) added new evidence with analyses of NAEP achievement gains that supported the effectiveness of test-driven accountability policy. However, the findings of those two studies diverged with regard to the effects of state accountability policies for different racial groups. On the other hand, Lee and Wong (2004) and Nichols et al. (2006) revealed more mixed results on state accountability policy effects. Lee and Wong found a positive policy effect on improving average achievement but no significant effect on narrowing racial achievement gaps. Nichols et al. found that accountability policy effects were limited to fourth grade math.

Methods

Despite many similarities among the aforementioned studies with the common use of national data sources for cross-state comparisons, they also varied in many significant aspects of research design, including the ways in which they classified states and measured the policy variable, their time frames, and their methods to analyze policy effects on student outcomes. Therefore, this review examines variations among the studies in terms of the following key research components:

1. Independent variables: How were state accountability policies defined and measured? How do the studies vary in the nature, type, and timing of the accountability policy variable?
2. Dependent variables: What subjects, grades, and time periods were chosen for the analysis of student achievement as outcome variables? Were those achievement measures valid and reliable to test hypothesized policy effects?
3. Samples: How were students, schools, and states selected for analysis? For cross-state comparisons, did the exclusion of students with learning disabilities and English language learners bring bias into the results? How did the policy effect vary among different racial and social subgroups of students? How well is the policy effect, if significant, generalizable to a larger population, longer time frame, and other related settings?
4. Analytical methods: What statistical methods were used for examining the policy-outcome linkages? What control variables were used to enhance internal validity? Did the studies take a quasi-longitudinal (cohort-based) or repeated cross-sectional (grade-based) approach to the analysis of achievement gains?
5. Effect sizes: How did the studies calculate and report effect sizes? How do the studies vary in the unit of analysis (students or states) and the level of school system to which effect sizes apply? What criteria, if any, were used by the studies to determine the practical significance of policy effects on student outcomes?

Downloaded from http://rer.aera.net at SUNY AT BUFFALO on May 1, 2009

For the present meta-analysis, total of 76 effect-size estimates were available from the 14 studies that investigated the effects of statewide high-stakes testing and test-driven accountability policies on reading and/or mathematics achievement during the 1990s. As shown in the Appendix, most studies reported multiple measures of policy effects because of their investigation of data from multiple grades, time periods, and/or subject areas. Although many studies examined the average policy effect for all students, only a few disaggregated the results by racial subgroups and explored potential accountability policy effects on racial achievement gaps.

First, the effect size was calculated with information available from each of the 14 studies. Although many studies used common data sources, they often differed in the ways in which they calculated and reported the statistical and practical significance of the results. Some studies, including those of Grissmer and Flanagan (1998), Amrein and Berliner (2002), and Rosenshine (2003), did not report statistical significance. Other studies reported statistical significance for their tested policy effects, but only a few of them reported corresponding effect-size estimates. Even when they reported effect sizes, their chosen indices of effect-size estimation often varied, using Cohen's *d*, *r*, or regression β weights (see Cohen, 1988; Peterson & Brown, 2005; Rosenthal, 1994). In most cases, Cohen's *d* was used as the metric of effect size to calculate standardized group mean differences between high-stakes testing states and comparison states. When studies only reported correlations between a continuous measure of high-stakes testing or test-driven accountability policy and student achievement outcomes, *r* was converted to *d*. When studies reported unstandardized regression coefficients, they were converted into standardized coefficients (β weights). The Appendix provides information on the procedures of effect-size calculation for each study.

Second, descriptive statistics were used to summarize the distributions of the 76 effect-size estimates. Furthermore, one-way analysis of variance (ANOVA) was used to examine variations among the effect-size estimates by selected variables including the aforementioned key characteristics of studies. The use of ANOVA may violate the assumption about the independence of observations (see Hedges, 1990, for a discussion of statistical dependence among effect sizes). Some of the studies reanalyzed the data of previous studies. In this case, dependency may exist among those studies as a result of using the same data, even though they involved independently chosen, different analytical methods and thus produced different results. One way to handle this dependency is to create a conservatively independent sample of estimates by grouping effect sizes that were identified as possibly correlated and using the average value of effect sizes within each group for further analyses (see Hedges, Laine, & Greenwald, 1994). Therefore, in this meta-analysis, I also calculated the average effect size by grouping effect sizes that were drawn from the same or comparable sources of data, such as samples drawn from the same grade or age population and tested in the same subject, and then combining the results across groups.

Results

Distribution of the Effects

Generally, seven studies favored states with high-stakes testing, six studies had mixed or insignificant findings, and one study favored states with low-stakes testing.

Downloaded from http://rer.aera.net at SUNY AT BUFFALO on May 1, 2009

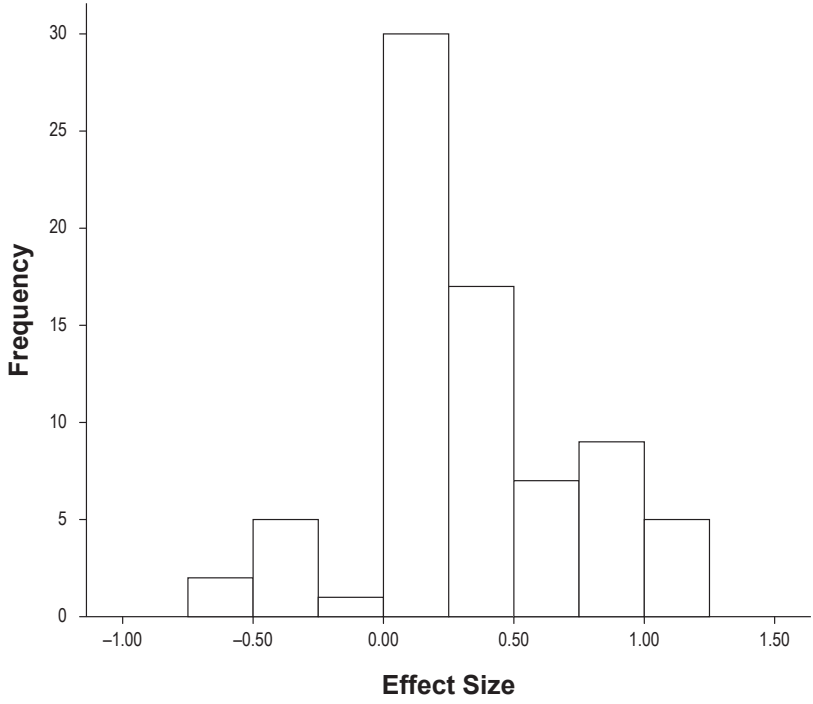

FIGURE 1. *Distribution of 76 effect-size estimates from 14 studies on the effects of high-stakes testing and accountability policy on reading and math achievement.*

The average effect size from all 76 estimates as shown in Figure 1 turned out to be modestly positive, while the effect sizes varied substantially among the measures, ranging from –0.67 to 1.24 ($M = 0.31$, $SD = 0.41$). The median value of all 76 effect-size estimates was 0.24. The 95% confidence interval for the sample estimate of its population effect size ranged from 0.22 to 0.40, and the average effect size of 0.31 was significantly greater than zero ($p < .001$).

To address statistical dependency among effect-size estimates that were drawn from the same data source, I recalculated the average effect size by grouping effect sizes into 10 distinctive groups as classified by subject and grade or age and using the number of available effect-size estimates in each group as weight. The grand mean effect size was 0.17 as obtained through an unweighted mean of 10 group means ($M = 0.21$ for 4th grade reading; $M = 0.36$ for 8th grade reading; $M = –0.01$ for 4th to 8th grade reading; $M = 0.00$ for 8th to 12th grade reading; $M = 0.51$ for 4th grade math [age 9]; $M = 0.39$ for 8th grade math [age 13]; $M = 0.04$ for math at age 17; $M = –0.09$ for 4th to 8th grade math; $M = –0.04$ for 8th to 12th grade math; $M = 0.27$ for 4th to 8th grade reading and math combined). Although this newly computed average effect size of 0.17 remained significantly greater than zero ($p < .05$), its magnitude shrank to only half of the average effect size that was obtained earlier without consideration of statistical dependence.

Downloaded from http://rer.aera.net at SUNY AT BUFFALO on May 1, 2009

Factors Accounting for Variations in the Effect-Size Estimates

The average effect size can be highly misleading, because it obscures substantial variations between and within studies. The between-study variation was significantly greater than the within-study variation, $F(13, 62) = 2.92$, $p < .01$. In the following sections, key factors that may have influenced variations in effect sizes among studies as well as within them are discussed and analyzed using ANOVA.

Independent variables. Effect sizes may vary among studies depending on the nature, types, and timing of accountability policies used as independent variables. The central question arises as to whether all of the studies actually used the same criteria for their classification of states as having high- versus low-stakes testing or strong versus weak accountability. Despite much overlap, not all studies addressed the same construct of policy treatment.

Lee (1998) used a comprehensive measure of standards-based education reform including not only high-stakes testing but also teacher certification requirements during the 1980s to explain NAEP achievement results. Fredericksen (1994), Bishop et al. (2001), and Jacob (2002) examined the effect of a single policy, high school exit exams, which was introduced first in the late 1970s and 1980s and aimed primarily at student-level accountability. Bishop et al. and Hanushek and Raymond (2004) examined school accountability policy in the 1990s, that is, the effect of giving rewards or sanctions for schools' academic performance.

Recent studies also covered a combination of student and school accountability policies. Amrein and Berliner (2002), Carnoy and Loeb (2002), and Lee and Wong (2004) considered both student and school accountability policies for their identification of high-stakes testing states or construction of accountability policy indexes. Unlike previous measures that relied on single surveys, the most comprehensive measure of test-driven external accountability policy was constructed by Lee and Wong, who combined three different policy survey data sets. It turned out that these different measures of policy index were closely related to one another, with average correlations of .7 or higher. When the list of 18 high-stakes testing states in Amrein and Berliner's study is compared with the list of 12 strong-accountability states in Lee and Wong's study, 10 states are common to both. Therefore, one may reasonably assume that the studies shared very similar policy environments capturing external test-driven accountability.

By and large, the effect sizes do not vary significantly by the primary target of accountability policy studied ($M = 0.27$ for school accountability, $M = 0.32$ for student and school accountability combined, and $M = 0.31$ for student accountability). This finding does not lend strong support to claims for school accountability (Hanushek & Raymond, 2004) or claims for student accountability (Bishop et al., 2001). The finding also raises a question about the claim that accountability policy should involve consequences for students as well as schools to effectively change student behaviors and outcomes (Peterson, 2006; Porter & Chester, 2002).

Is it accountability policy only that had an impact on the achievement gains? When examining this question, studies are vulnerable to model specification bias, that is, the omission of a confounding policy variable as a predictor of achievement gain. It is not possible to disentangle the effects of a single particular policy from other policies adopted at the same time. Although some researchers acknowledged that test-driven accountability policies are just one component of standards-based

Downloaded from http://rer.aera.net at SUNY AT BUFFALO on May 1, 2009

education reform, they did not investigate this issue. The exceptions are Braun (2004) and Lee and Wong (2004), who both attempted to control for a broader measure of standards-based education reform policy. Although both studies did not find any significant changes as a result of the control, this does not rule out any other rival explanations.

What is simply called accountability policy in some studies actually refers to test-driven external accountability policy. Although test-driven accountability policy became more popular during the 1990s, it was added on to preexisting input-based accountability policy instead of replacing it. Because many reform states were active in adopting both types of accountability policy during the 1990s, looking at only one type of accountability policy may result in an overestimation of the policy effect on student achievement.

Grissmer and Flanagan (1998) pointed out that both North Carolina and Texas had multilevel systems of accountability, with schools as the primary focus of rewards and sanctions. Challenging this argument, Darling-Hammond (2000) pointed out that student assessments were not in place and accountability policies were not in effect by the time of the 1996 NAEP assessment in those two states and suggested that the states' achievement gains may be related to higher teacher certification standards, salaries, and professional development policies. To test these competing hypotheses, a measure of teacher standards policy was added to Carnoy and Loeb's (2002) regression analysis model of eight grade mathematics achievement gains from 1996 to 2000.[3] Although this change did not influence the effect of accountability policy, it was more appropriate to look at the policy effects on achievement gains for the extended period, because those teacher certification standards were adopted before 1996. The state teacher certification policy had a significant positive effect on the NAEP eighth grade mathematics gain scores from 1990 to 2000, whereas the state accountability policy did not (see Table 1). The result of this analysis suggests that the estimation of policy effects is sensitive to model specification and that the effects of both input-oriented and performance-based educational accountability policies need to be investigated simultaneously.

Effect sizes tend to vary among studies by their time periods ($M = 0.47$ for the late 1990s, $M = –0.13$ for the early 1990s, and $M = 0.08$ for the 1980s). The average effect size from studies covering the late 1990s (1996 to 2000) was significantly larger than that from studies covering the 1980s or early 1990s (1992 to 1996) ($p = .002$). This trend may be attributable to the fact that the focus of high-stakes testing and accountability policy has shifted from ensuring minimum competency and basic skills for low-achieving students to high standards and proficiency for all students. For example, Jacob (2001) related the absence of a high school graduation exam effect on achievement to the unchallenging nature of pass–fail minimum competency testing, which produces very high passing rates. However, some states adopted more rigorous testing than others (Bishop et al., 2001; Achieve, 2004). This interstate variation in the level of performance standards and the difficulty of passing tests needs to be considered in future studies.

Does the timing of the accountability policy variable match the time frame of the achievement outcome variable to capture the hypothesized policy effect? There is a potential bias arising from not considering variation among states in the starting point and duration of their accountability policies. In evaluating the policy effect, most studies gave no explicit consideration of when the policies became

Downloaded from http://rer.aera.net at SUNY AT BUFFALO on May 1, 2009

TABLE 1
Estimated effects of accountability policy and teacher standards policy on the NAEP math state average achievement gain scores

Independent variable	1992–2000 fourth grade math gain	1992–2000 eighth grade math gain	1990–2000 eighth grade math gain
Teacher standards index	0.38 (0.82)	0.84 (1.79)	1.41* (2.25)
Accountability index	1.29* (2.19)	1.41* (2.36)	1.27 (1.67)
Baseline score	–0.11 (–1.03)	–0.03 (–0.28)	–0.14 (–1.20)
Constant	26.79 (1.13)	8.52 (0.35)	41.78 (1.32)
R^2	.37	.40	.47
N	31	30	25

Note. The teacher standards index was constructed with data from the Council of Chief State School Officers (1996) survey of state entry-level teacher certification requirements. The total number of requirements adopted by states among these five items was coded as an index of state-level teacher certification policy: (a) basic skills test, (b) professional skills test, (c) subject specialty test, (d) classroom observation, and (e) portfolios. The accountability index was drawn from Carnoy and Loeb (2002). Both the teacher standards index and the accountability index are on the same scale, ranging from 0 to 5. The accountability policy index was modestly correlated with the teacher standards policy index ($r = .35$). Regression coefficients were obtained from weighted least squares regression using the inverse of the standard error of the dependent variable as a weight. Values in parentheses are *t* statistics. NAEP = National Assessment of Educational Progress.
*$p < .05$.

effective and how long the students in the NAEP sample were exposed to the policies. Further complicating this state policy calibration is that some states may have revised their policies over time.

For example, Carnoy and Loeb's (2002) accountability policy index was constructed with data from the Consortium for Policy Research in Education's survey of policies that were in effect as of 1999–2000. It requires caution when the 1996 and 2000 NAEP measures were used to evaluate the effect of policy that had been adopted before 1996. Carnoy and Loeb argued that "since the NAEP mathematics test was given in 1996 and 2000, it provides a good measure of whether state accountability systems—many of which came into being in the mid-1990s—are having a significant effect on student learning outcomes" (pp. 308). However, this statement is not valid for some policies, particularly that of the high school exit exam, that were adopted in many states before the mid-1990s; the 1st year when eighth grade students were affected by high school exit tests was 1993–1994 for New York, 1989–1990 for North Carolina, and 1986–1987 for Texas. Consequently, not only the 2000 eighth grade cohort but also the 1996 eighth grade cohort should have been affected by high school exit tests.

Eighth graders' academic achievement is a cumulative product of schooling that they received throughout the K–8 schooling period. The schooling period for the 1996 NAEP eighth grade cohort was 1987 to 1996, and the counterpart for the 2000 NAEP eighth grade cohort was 1991 to 2000. Because a 5-year period (1991 to 1996) was common to both cohorts, the effect of a policy that

Downloaded from http://rer.aera.net at SUNY AT BUFFALO on May 1, 2009

was adopted and implemented prior to 1996 should be shared by the two cohort groups. Therefore, what is under evaluation in this case is actually not so much the effect of a full-scale policy as the effect of the varying degree of exposure of each cohort to a given policy.

Dependent variables. Effect sizes tend to vary among subjects, grades, and time periods chosen for the analysis of student achievement outcome variables. First, the accountability policy effect is greater for mathematics than for reading ($M = 0.36$ for math, $M = 0.20$ for reading). However, the mean difference of 0.16 is not statistically significant ($p = .31$). Direct comparison of the results between two subject areas needs caution, because the NAEP results for reading covers only fourth grade over a relatively short time period. Moreover, relatively few studies examined the effect of policy on reading achievement; some studies that dealt exclusively with math may have chosen to report only such highly significant results after the fact.

Second, comparison of the effect sizes by grade level shows that the effects are relatively larger at the lower grade levels ($M = 0.41$ for elementary school grades, $M = 0.27$ for middle school grades, and $M = 0.03$ for high school grades). However, the difference between grade levels is not statistically significant ($p = .14$). Studies using the NAEP (Fredericksen, 1994), NELS (Jacob, 2001), or SAT or ACT (Amrein & Berliner, 2002) did not lend support for the effect of high-stakes testing at the high school level. Because the major target of high-stakes testing policies (e.g., high school exit exams) was often the high school population, this result appears to contradict an expectation of a greater policy effect at the upper grade levels.

Third, the effect sizes do not vary systematically by the length of time period for evaluating achievement gains ($M = 0.33$ for one-shot gains, $M = 0.38$ for 4-year gains, and $M = 0.31$ for 8-year gains). The overall mean difference by the length of achievement gains is not statistically significant ($p = .95$). This result appears to be inconsistent with the expectation that longer exposure to a given policy as a treatment will generate more significant effects. There are potential threats to the validity of not only cross-sectional studies using achievement status in a particular year but also repeated cross-sectional studies that do not consider possible regression and trend artifacts in evaluating achievement gains for a limited time period.

Were the high-stakes testing states' observed achievement gains in the 1990s inflated because of "regression to the mean" or the continuation of a previous trend? As some studies pointed out, so-called high-accountability states were relatively lower performing states even before they adopted accountability policies. In their analysis of math achievement gain scores from 1996 to 2000, Carnoy and Loeb (2002) statistically controlled for the baseline measure of state performance assessed in 1996 and took into account the possibility of a regression to the mean. However, they did not consider the effect of a previous performance trend on the achievement gain scores from 1996 to 2000.

Indeed, achievement gain scores are not very stable at the state level. The correlation of the fourth grade mathematics gain score from 1996 to 2000 with that from 1992 to 1996 is .11, and the correlation of the eighth grade mathematics gain from 1996 to 2000 with that from 1992 to 1996 is .26. In other words, states that gained more from 1992 to 1996 did not necessarily continue to gain more from 1996 to 2000. The volatility of gain scores requires that one look at changes in

Downloaded from http://rer.aera.net at SUNY AT BUFFALO on May 1, 2009

performance over the long run. The studies differed in the ways in which sampling errors of gain score estimates were considered in analyzing and interpreting the effects of accountability policy. For instance, Amrein and Berliner (2002) did not consider sampling errors at all, whereas Braun (2004) conducted reanalysis with explicit consideration of sampling errors.

Finally, there are some differences among the studies in their choice of outcome variable metrics and interpretations. Most studies simply used original scale scores to compute and analyze gain scores. Some studies, such as those of Amrein and Berliner (2002) and Braun (2004), used relative gain scores by subtracting national average gain scores from state average gain scores. Such norm-referenced transformation made the state gain scores appear smaller and less significant. However, this choice did not affect the estimation of effect size in Braun's reanalysis, in which the same metric was applied to low-stakes testing states as well.

In contrast, Carnoy and Loeb (2002) chose to use a criterion-referenced measure of gain, that is, change in the percentage of students meeting at desired achievement levels, as the metric of outcome variables. The NAEP's percentage of students at or above the basic or proficient level is not a linear measure of achievement, unlike the scale score derived through scaling procedure used in item response theory. For example, a lower achieving state's gain from 0% to 10% at or above the proficient level may imply a greater amount of improvement in state average achievement than a higher achieving state's gain from 50% to 60% at or above the proficient level. Although this problem may affect the comparison of high-stakes versus low-stakes testing states' achievement gains, the choice of different metrics did not substantially alter their findings.

Analytic samples. The comparison of states in the prior studies did not include all 50 states but only states that were available in the NAEP data. Because about 30% of the states were not included in most studies' analytic samples, it is difficult to generalize the NAEP state assessment results to the entire nation. Nonparticipating states are likely to differ from participating states because they were not randomly chosen but rather self-selected to participate in the NAEP. For example, the excluded states turned out to be relatively weak accountability states on the basis of Carnoy and Loeb's (2002) state accountability policy index. The average accountability policy index for the nonparticipating states in the 1996 to 2000 NAEP ($n = 17$) is 1.7, whereas the average policy index for participating states ($n = 33$) is 2.3. If the nonparticipating states had made similar or larger achievement gains than the participating states during the same period, then the reported effect of the accountability policy on student achievement may have been slightly biased upward.

There were also inconsistencies in the number of states included in the analytic samples among the studies depending on the time period, grade, and subject. Carnoy and Loeb (2002) inflated their analytic sample size of states by using an interpolation method to estimate the gains for the states that had no reported NAEP achievement measures for a particular year. Four states—Idaho, Illinois, Ohio, and Oklahoma—did not participate in the 1996 NAEP, and thus, there could not be any measure of achievement gain from 1996 to 2000. Nonetheless, the four states were included in the authors' analytic sample, and the mathematics achievement gains from 1996 to 2000 for those four states were estimated.[4] Because the authors did

Downloaded from http://rer.aera.net at SUNY AT BUFFALO on May 1, 2009

not check and report if using the inflated list of states induced differences in their finding, the same regression analysis has been conducted without those additional states. The results of reanalyses for both the fourth and eighth grade samples showed that the effects of accountability index on gains remain significant.

Schools and student samples were randomly selected in participating NAEP states. Although the random sampling of schools and students may help ensure the representation of their target populations within each state, there are potential biases in excluding certain groups of students, including those with learning disabilities and English language learners. Because the exclusion rate of such students varied from state to state, Amrein and Berliner (2002) pointed out that the larger achievement gains in high-stakes testing states such as North Carolina and Texas are attributable partly to their relatively large increases in exclusion rates. For instance, there was a 10% increase of the exclusion rate for North Carolina and an 8% increase of the exclusion rate for Texas between the 1992 and 2000 NAEP fourth grade math assessments. However, Braun (2004) showed that those two states were outliers that deviated from the overall pattern of a relationship between change in the exclusion rate and gain scores among all participating NAEP states. In addition, in Carnoy and Loeb's (2002) and Raymond and Haushek's (2003) studies, it appears that statistically adjusting gain scores for changes in exclusion rates did not lead to significant changes in the estimation of policy effects.

Finally, studies also varied in terms of observed policy effects for racial subgroups of students. Among the 14 studies, only a few disaggregated the results by racial group to explore the policy effects on achievement gaps. Carnoy and Loeb (2002) found that the effects of accountability policy were greater for Blacks and Hispanics than for Whites, thus narrowing the racial achievement gaps. In contrast, Hanushek and Raymond (2004) found the opposite direction of policy effect: a widening Black–White gap. Finally, Lee and Wong (2004) and Nichols et al. (2006) did not find any significant policy effects on racial achievement gaps. This discrepancy may be related to their investigation of different time periods and use of different analytical methods. It turns out that the overall mean difference among racial groups (M = 0.25 for Whites, M = 0.36 for Blacks, and M = 0.32 for Hispanics), on the basis of 10 effect-size measures for each group from those 4 studies, is not statistically significant (p = .95). Among the 14 studies, only Lee and Wong examined changes in the achievement gap among socioeconomic subgroups of students (on the basis of the availability of home reading materials, eligibility for free or reduced-price lunch, and the level of parental education) under state test-driven accountability policy, showing largely insignificant policy effects on the social achievement gaps.

Analytical methods. Are the eighth grade achievement gains between two time points simply due to the fact that the eighth graders who participated in the earlier NAEP were different from the eighth graders participating in the later NAEP? Obviously, under this grade-based successive group comparison method, the two groups assessed at different years represent different cohorts of students. Although some studies, such as Carnoy and Loeb's (2002) analysis of achievement gain scores from 1996 to 2000, did consider demographic changes between the two cohort groups, there are many unknown differences in subject characteristics that might confound the estimated effects of accountability policy on student achievement.

Downloaded from http://rer.aera.net at SUNY AT BUFFALO on May 1, 2009

TABLE 2
Estimated effects of accountability policy on the NAEP math and reading state average gain scores

Independent variable	1996–2000 fourth grade math gain	1996–2000 eighth grade math gain	1996 fourth grade to 2000 eighth grade math gain	1994–1998 fourth grade reading gain	1994 fourth grade to 1998 eighth grade reading gain
Accountability index	0.49 (1.53)	1.50** (3.83)	0.17 (0.36)	0.32 (0.81)	0.46 (1.20)
Baseline score	–0.06 (–0.89)	0.03 (0.47)	0.21 (2.17)	–0.12 (–1.60)	–0.28*** (–3.99)
Constant	15.37 (1.02)	–7.77 (–0.47)	5.44 (0.25)	27.07 (1.68)	106.92 (7.04)
R^2	.12	.37	.13	.15	.47
Sample size	35	33	34	35	32

Note. The accountability index was drawn from Carnoy and Loeb (2002). Regression coefficients were obtained from weighted least squares regression using the inverse of the standard error of the dependent variable as the weight. Values in parentheses are *t* statistics. NAEP = National Assessment of Educational Progress.
$^{**}p < .01$. $^{***}p < .001$.

In contrast, the cohort-based tracking method to examine states' academic improvement on the NAEP from 1996 to 2000 was to compare NAEP scores from the 1996 fourth graders with those from the 2000 eighth graders. Because the NAEP used separate sampling procedures each year, there is no guarantee that the 1996 fourth grade sample can be well matched to the 2000 eighth grade sample. Nevertheless, the gain scores obtained through such a quasi-longitudinal tracking of the same cohort are more likely to be free from the cohort artifact.[5]

For this review, the data from Carnoy and Loeb's (2002) study were reanalyzed by substituting the mathematics gain scores from 1996 fourth graders and 2000 eighth graders as an outcome variable while keeping the same accountability index as a predictor. The effect of the accountability index on achievement gain turned out to be too small to be significant (see Table 2). The baseline performance measure (i.e., 1996 fourth grade mathematics score) has a negative effect on the amount of the gain from 1996 fourth graders to 2000 eighth graders. The accountability index remains insignificant after changes in the racial and social composition of the sample are controlled for. The policy effect is also absent for the quasi-longitudinal analysis of reading achievement gains (i.e., the gains from 1994 fourth graders to 1998 eighth graders). This implies that there is high degree of inconsistency in the estimated effect of the accountability index depending on the analytical method used and that the observed effect of accountability policy may reflect a cohort artifact.

Among the 76 effect-size measures, 14 came from cross-sectional studies and 62 from longitudinal or quasi-longitudinal studies. Sixteen of those 62 estimates used a cohort-based tracking method for the analysis of achievement gain, whereas 46 used a grade-based or age-based successive group comparison method. Figure 2 shows that the estimate of the policy effect is null for cohort-based analysis ($M = 0.03$) but moderately positive for grade-based or age-based analysis ($M = 0.40$). The mean difference of 0.37 between the two methods is

Downloaded from http://rer.aera.net at SUNY AT BUFFALO on May 1, 2009

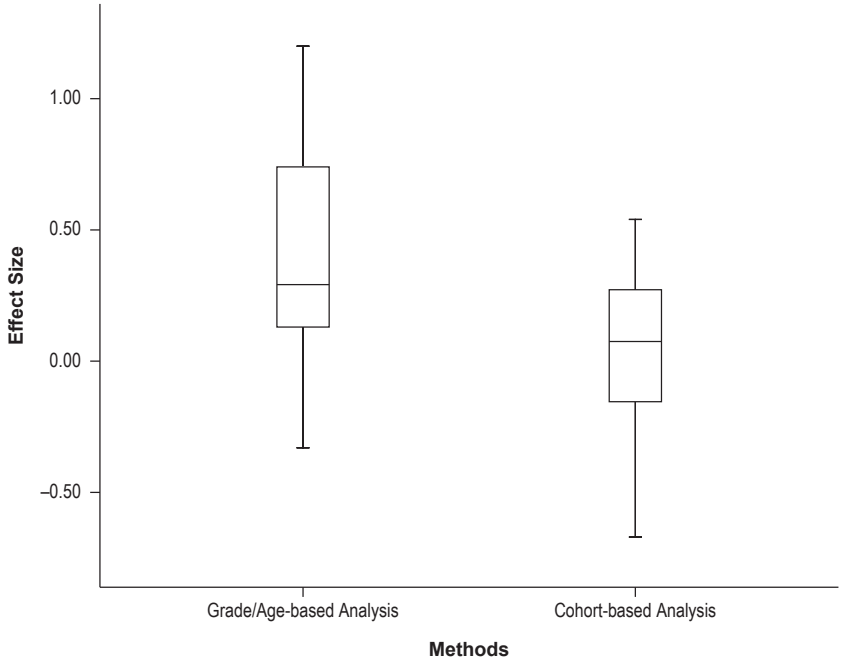

FIGURE 2. *Box plots of the distributions of accountability policy effect-size estimates by methods for analyzing gain scores (*n = *16 for cohort-based analysis,* n = *46 for grade- or age-based analysis).*

statistically significant ($p < .001$). This contrast indicates that the results are highly sensitive to the choice of analytic method. One may conjecture that this difference between the two methods, as shown by the larger effect for grade- and age-based comparison than for cohort-based comparison, is attributable to a difference in the time span of achievement gains: A cohort-based comparison method affords only a 4-year gain because of the NAEP's state assessment target grades being fourth grade and eighth grade, whereas an age- or grade-based method allows for a longer time span (e.g., 10 years from 1990 to 2000). Nevertheless, as reported in the previous section, the length of time was not significantly related to the size of achievement gains.

Another critical research design factor that may have caused differences in the findings among the studies was the point of comparison, that is, which of the states are treated as a comparison group. Raymond and Hanushek (2003) criticized Amrein and Berliner's (2002) decision to compare high-stakes testing states with the national average instead of with low-stakes testing states as the violation of a basic principle of scientific research. In fact, this choice made a difference in the average effect-size estimates between Amrein and Berliner's analysis and subsequent reanalyses of the same data ($M = 0.08$ for Amrein and Berliner and $M = 0.57$ for a combination of studies including Braun, 2004; Raymond and Hanushek, 2003; and Rosenshine, 2003).

Downloaded from http://rer.aera.net at SUNY AT BUFFALO on May 1, 2009

Effect sizes. The effect-size estimates were based on the attributes of individual studies in terms of their chosen units of analysis and types of scores used (see the Appendix). It needs to be noted that there were subtle variations among the studies in their units of analysis and types of scores used as the basis of effect-size calculation. Some studies used student-level data, whereas others used state-level aggregate data. Among the 14 studies, 3 (Fredericksen, 1994; Jacob, 2001; Lee & Wong, 2004) used individual student-level data, whereas the other 11 studies used state-level aggregate data. Therefore, yardsticks used as the denominators of the effect-size formulas varied by the level at which the standard deviation was calculated (student level vs. state level) and the types of scores used for this computation (state-level average scores vs. gain scores). The differentiation of effect-size calculations on the basis of (a) state-level standard deviation and (b) student-level standard deviation can help address policy questions at different levels of application: (a) What is the effect of adopting test-driven accountability policies on individual states' achievement status or gains relative to other states? and (b) What is the effect of test-driven accountability policies on individual students' achievement relative to the national norm? Most studies included in this review used NAEP state-level aggregate gain scores as dependent variables, and thus, the estimation of the average effect size was heavily influenced by those studies. Policy makers may be concerned about the question of how much their states would improve in terms of average achievement (compared with other states) if they adopted high-stakes testing with stronger accountability. For instance, an effect size of 1 from this type of study means that an average-improving state will advance its rank of achievement gain from the 50th percentile to the 84th percentile. Although this progress sounds enormous, it does not translate into the same level of progress for individual students, whose achievement distributions are much more heterogeneous than are those of states.

Although standard deviations of NAEP scale scores vary slightly by grade, subject, and year, the student-level standard deviation is about 4 times greater than the state-level standard deviation of average scores, which in turn exceeds the state-level standard deviation of gain scores. Clearly, estimating effect sizes on the basis of the distribution of state-level versus student-level test scores can give quite different impressions about policy effect. For instance, Hanushek and Raymond (2004) calculated effect sizes on the basis of the standard deviation of state-level average scores. Carnoy and Loeb (2002) did not directly report the effect sizes of accountability policy, but they referred to the standard deviation of state-level gain scores as their yardstick of the effect size; they claimed that

> with a mean gain of 4.8 percentage points and a standard deviation of 3.6 in average state proportions scoring at or above basic skill levels, the increase in gain from raising the external pressure on schools by the state appears to be substantial. (p. 313)

This statement can be misleading, because most states made very small and insignificant gains that appear larger than they really are when interstate variations in gain scores are used as a criterion to evaluate the effects of a policy. In contrast, Lee and Wong (2004) used the standard deviation of student-level test scores as an alternative yardstick to gauge effect size. For an increase of 2 standard deviations in accountability policy score (changing from weak to strong status), their estimated

Downloaded from http://rer.aera.net at SUNY AT BUFFALO on May 1, 2009

mathematics score gain for all eighth grade students from 1992 to 2000 was about 10 points. This 10-point gain amounts to 2 standard deviations of the state-level gain scores ($SD_{state} = 4.5$) but less than one third of the standard deviation of student-level scores ($SD_{student} = 36$).

Recalculating all 76 effect-size estimates using the student-level standard deviation as opposed to the state-level standard deviation produces a much smaller mean effect size of 0.08. This average effect size with the student-level standard deviation is only about a quarter of the average effect size that was obtained with the state-level standard deviation. This student-level effect means that individual students may improve their reading or math achievement by 8% of 1 standard deviation (e.g., from the 50th to the 53rd percentile) relative to the national population of all students across states, when their own state switches to high-stakes testing or moves from weak accountability to strong accountability. This small amount of gain may translate into the equivalent of 2 to 3 months of learning, depending on their grade level; this is based on the estimated rate of academic growth per grade in NAEP reading and math, which is about 0.4 standard deviations in the middle grades (between 4th grade and 8th grade), and a quarter of a standard deviation in high school grades (between 4th grade and 12th grade).

Discussion

Limitations of Prior Studies

Because there has been a strong call for evidence-based education policy, the validation of studies requires that their findings stand up to rigorous scientific scrutiny (Shavelson & Towne, 2002; Slavin, 2002). The aforementioned studies share several limitations. The biggest threats to the internal validity of causal-comparative and correlational studies arise from the nonrandom assignment of states to treatment (test-driven accountability policy in this case), which results in many unknown differences in the characteristics of subjects between treatment and comparison groups (see Campbell & Stanley, 1963). This threat to internal validity prevents us from interpreting correlations as causal relationships and attributing states' achievement gains to their accountability policy. Although most researchers acknowledged the limitations of ex post facto research design, their findings were not often viewed as tentative, falsifiable evidence that needs verification.

One of threats to the internal validity of previous studies comes from the fact that they fell short in defining and measuring state accountability policies as independent variables. They often failed to differentiate multiple types of educational accountability policies (e.g., teacher vs. student accountability) that coexisted in states and to examine their separate and joint effects. Although external, test-driven school accountability that rewards or sanctions whole schools for their academic improvement through high-stakes testing may have become a national prototype of accountability, this system is limited in its design and impact. Future research needs to broaden the notion of educational accountability from this fundamental question: Who is held accountable for what, how, and why? Along each of these dimensions, we can classify accountability policies into categories. For example, the logic of current external test-driven accountability policy appears to draw on rationalistic and behaviorist views of human behavior by positing that holding schools, teachers, or students (the question of who) accountable for academic

Downloaded from http://rer.aera.net at SUNY AT BUFFALO on May 1, 2009

performance as measured by standardized tests (the question of what), with incentives such as rewards and sanctions (the question of how), will inform, motivate, and reorient the behavior of schooling agents toward the goal (the question of why) (see Benveniste, 1985; Wise, 1979; Rowan & Miskel, 1999).

Researchers also should attempt to further refine their test-driven accountability policy measures by capturing variations within this narrow accountability subtype in terms of specific design features of the programs such as the measurement of status versus gain scores, the method of adjustment for factors out of school control that influence test results, and the magnitude of incentives (see Clotfelter & Ladd, 1996; Linn, 2001). Indeed, the simple dichotomization of states into high-stakes testing versus low-stakes testing categories is no longer relevant under NCLB, which has spread high-stakes testing to all states. Researchers also should explore advanced measurement models to construct more valid and reliable measures. There were previous attempts to use psychometric models, such as Rasch models, with the use of state policy survey data for measuring state activism in standards-based education reform as a package (see Lee, 1997; Swanson & Stevenson, 2002).[6]

The studies reviewed herein examined policy effects on achievement during the past two decades prior to NCLB, but some policies may still be too recent to make measurable effects. Student accountability polices may have more immediate impact, such as academic promotion or graduation depending on test performance. School accountability policies may take longer. Although states report on school performance, the reporting may not translate into real sanctions in immediate terms; there may be sanctions after 3 to 5 years of failing performance. In other words, accountability systems vary on the actual and immediate use of "high-stakes" measures. If these are designed for a longer term purpose, then they are not likely to have any effects. Clearly, there is a need for evaluating the policy effect over the long run.

Because not all states had participated in the NAEP before NCLB, the studies reviewed herein also raise questions about the generalizability of their findings to all states, including non-NAEP states. Will the estimated effect of accountability policy show up in other unexamined subjects and grades as well? Although the studies' use of a large-scale NAEP database with a statewide representative sample of students may contribute to their external validity, there are other potential limitations, such as nonparticipating states, limited time periods, and selected grades and subjects, that constrain the generalizability of the study findings. This problem may be ameliorated in future studies as the role and scope of NAEP has expanded to include all states with biennial testing as a result of NCLB. At the same time, however, NAEP may become less immune from the threat of test contamination as a result of its enhanced role to confirm state assessment results under NCLB, which will generate pressure for NAEP-driven achievement gains.

Whether high-accountability states averaged significantly greater gains on the NAEP than students in states with little or no accountability measures, reasons for the presence or absence of an expected effect remain to be investigated and explicated fully. Previous studies tended to take a purely empirical or atheoretical approach in evaluating the policy effect. Most studies did not present any theoretical or conceptual framework about the mechanism through which accountability policy might have affected student outcomes. This kind of black-box approach

Downloaded from http://rer.aera.net at SUNY AT BUFFALO on May 1, 2009

ignores variables that can moderate or mediate the relationship between state policy and student outcome variables. Part of this problem is attributable to the use of states as the unit of analysis and concerns about aggregation bias at the state level as well as concerns about limited sample size and loss of power as a result of adding more predictors.

Critics point out that the working theory behind test-based accountability system is fatally simple and that internal accountability must precede external accountability (Elmore, 2002; Newmann, King, & Rigdon, 1997; O'Day, 2002). Although some of the reviewed studies mentioned the importance of capacity building and funding in their literature reviews, their analyses lacked specification and testing of any mechanism by which test-driven accountability policy may have affected student achievement in a multilayered state educational system. One critical factor that might facilitate or constrain the effect of accountability policy on achievement is the level of state support to help schools meet the standards (Harris & Herrington, 2006). Lee (2006a) showed that the effect of accountability policy is moderated by school support; strong accountability states that provided relatively favorable schooling conditions in terms of class size, teacher qualification, and per pupil spending made larger gains than their counterparts that did not. Further investigations are needed to understand the circumstances in which test-driven external accountability policy works.

Finally, the important question remains as to how accountability policy compares with other educational policies in terms of effect and what it costs in comparison with such alternatives. Standards-raising education reform and accountability policies were popular with state legislators because they held out the hope of greater cost-effectiveness (Berliner & Biddle, 1995). It was argued that the cost of testing and accountability was small relative to the cost of other expensive educational programs such as class size reduction; the cost of paying for tests, publishing results, and writing and publishing the standards on which the tests are graded is about $5 per student on average (Hoxby, 2002). This estimate, however, included only the most basic part of an accountability system. Further studies are needed to address the cost of monitoring, identifying, assisting, rewarding, and/or punishing the target population of accountability according to test results and other related information.

Among the studies reviewed in this article, very few investigated adverse side effects of high-stakes testing policy. Carnoy and Loeb (2002) reported no harmful state accountability policy effect on student retention or high school completion rates. However, the comparability of such statistics as reported by state departments of education remains dubious in the absence of common criteria and reliable measures. Jacob (2001) showed that high school graduation exams increased the probability of dropout among the lowest ability students but not among the average students. Evidence on the effects of accountability policy on academic achievement must be weighed carefully with more evidence on potential harms and risks (see Heubert & Hauser, 1999; Lee, 2007; Madaus & Greaney, 1985; Orfield & Kornhaber, 2001; Shepard, 1991).

It is inappropriate to make decisions about students or schools on the basis of a single measure of achievement (American Educational Research Association, American Psychological Association, & National Council on Educational Measurement, 1999). Whereas states could previously use multiple sources of information to make

Downloaded from http://rer.aera.net at SUNY AT BUFFALO on May 1, 2009

accountability determinations and set their own timelines before NCLB, the law now prescribes both the nature of accountability measures and the timelines for achievement (Marion et al., 2003). Although states still have an option to use multiple assessments for accountability decisions, the U.S. Department of Education requires that they ensure that the assessments are aligned with state standards and are of acceptable technical quality. Some see this option as a compromise that might result in assessment patchwork, and only a few states considered use of locally selected and/or locally developed assessments (Erpenbach, Forte-Fast, & Potts, 2003). Others argue that classroom assessments, administered over the full course of a year, can provide more complete measures of the key standards identified by the state that are otherwise impractical to assess on a large-scale basis (Baker, 2003; Commission on Instructionally Supportive Assessment, 2001). Studies of test-driven accountability need to pay attention to the relative costs and benefits of using national, state, and/or local assessments for accountability decision making.

Implications for Research and Policy

Although NCLB calls for evidence-based education policy, the past research on the impact of test-driven accountability policy on achievement tends to fall short of meeting rigorous scientific research standards. Any causal attribution from such observational studies is not warranted because of many serious threats to internal validity. Moreover, this review has limitations in that it was restricted to cross-state causal-comparative or correlational studies that used comparable test results across states, such as the NAEP and NELS, to explore the effects of states' test-driven external accountability policy on reading and/or math achievement. The use of restrictive criteria for the inclusion of studies may have resulted in a loss of insights from relatively smaller scale studies that used state and local assessment results to assess the effect of high-stakes testing at the district, school, or classroom levels. Further study is needed to compare and synthesize the evidence from studies of accountability that have been conducted at different levels of the school system with different sources of data. Preliminary review of some recent studies that examined the impact of school and student accountability programs on academic achievement on the basis of state or local data suggest mixed and inconclusive evidence as well; for example, the sites of studies with such conflicting views and evidence include Chicago (Bryk, 2003; Jacob, 2003) and California (Betts & Danenberg, 2002; Hauser, 2002). These smaller scale studies of individual state or district data also were not free from the aforementioned limitations of larger scale studies of national data. Researchers who work with local data may be in a better position to capture the effects of programs in flux and investigate complex processes within schools and classrooms. At the same time, they need cautions because of potential biases in using results from high-stakes state and local tests that function as an intervention as well as a measure of student outcomes.

With these caveats in mind, this review of large-scale studies using national data can make a timely and important contribution to policy discussion by producing more generalizable knowledge on the effects of high-stakes testing policy on the basis of common national benchmarks for the comparison of student achievement outcomes. The meta-analysis of 76 effect-size estimates drawn from 14 selected studies showed a modestly positive policy effect on average but no significant

Downloaded from http://rer.aera.net at SUNY AT BUFFALO on May 1, 2009

effect on narrowing the racial achievement gap. More important, the analysis revealed substantial discrepancies among studies, particularly between grade- and age-based analyses of cross-sectional or repeated cross-sectional data and cohort-based longitudinal or quasi-longitudinal analyses. Therefore, educational policy makers and practitioners should be cautioned against relying exclusively on research that is consistent with their ideological positions to support or criticize the current high-stakes testing policy movement. They should become aware of potential biases arising from the uncertainty and variability of evidence in the literature. This article raises questions about the scientific basis of NCLB and state accountability policy and possible social consequences of the policy on the basis of inconclusive evidence and/or false premises about the policy impact on student achievement.

How large of an effect is enough to declare a policy a success? There is no one-size-fits-all criterion to evaluate the size of a policy effect, and many studies were vague about the benchmark that can be used to judge any practical import of their findings about policy effects. The contrast of effect-size estimates on the basis of student-level versus state-level distributions of test scores leads to different impressions about the practical significance of test-driven accountability policy effects. Furthermore, if we also evaluate the size of the reported policy effect relative to an announced policy goal (e.g., NCLB's goal of reaching 100% proficiency for all students by 2014), test-driven external accountability policy turns out to be far less effective. For White eighth graders in Carnoy and Loeb's (2002) study, for example, a two-step move (e.g., shifting from simple testing requirements to having moderate repercussions for schools and districts in combination with a high school exit test) would only bring about a 2.5 percentage point gain in the percentage of students reaching or exceeding the proficient level.

Past studies generally focused on the issue of whether high-stakes testing policy works in general, but they did not answer other important questions, such as under what circumstances the policy works and which groups benefit more or less from the policy. The synthesis of past studies reveals that there were no systematic differences between high-stakes and low-stakes testing states in their progress toward narrowing racial achievement gaps. If test-driven accountability policy left pernicious achievement gaps unchanged during the 1990s, it could signify policy deadlock to the past national progress in narrowing the Black–White and Hispanic–White achievement gaps during the 1970s and early 1980s (Lee, 2002). Although this article's review relied on a limited number of studies of state accountability policies during pre-NCLB period only, it challenges the core argument by proponents of test-driven accountability policy that the policy should help close the achievement gap by serving disadvantaged minority students most. Given tensions between improvement of academic excellence and equity, further studies are needed to explore whether and how external test-driven accountability policy can contribute to equity, particularly in narrowing the achievement gap among racial and social groups. Moreover, any hidden costs and adverse side effects of high-stakes testing and accountability policy need further investigations.

It needs to be noted that the studies in this review were able to explore the impact of state high-stakes testing prior to NCLB by comparing the first-generation states that adopted accountability policy prior to NCLB with the second-generation states that did not adopt similar policy until NCLB. To argue that the first-generation

Downloaded from http://rer.aera.net at SUNY AT BUFFALO on May 1, 2009

states adopting strong accountability policies prior to NCLB significantly improved academic achievement is not convincing until it can be demonstrated that the alleged effect can transfer to the second-generation states (see Lee, 2006b). Under NCLB, the existence of dual accountability systems and interactions between federal and state policies poses methodological challenges for the analysis of post-NCLB data. Hanushek and Raymond (2004) pointed out the irony that the implementation of NCLB essentially precludes analysis of further impacts of overall accountability systems by eliminating a comparison group of states without accountability systems, although the continuation of individual states' own locally developed schemes under NCLB may still allow for possible comparison of the impacts of alternative types of accountability systems.

At the same time, there is also the irony that NCLB has permitted national tests to be less of a gauge of state performance than it had been. By requiring that states tie school accountability systems to state-defined performance standards and state-chosen test results, NCLB raises the stakes for state assessments vis-à-vis other tests. If the nation and states continue the current policy course, academic proficiency is unlikely to improve significantly on the NAEP, but it is possible that the state assessment, which is the basis of accountability decisions, will continue to give a false impression of progress (Lee, 2006b). As Congress moves to reauthorize NCLB, it is poised to discuss the topic of increasing the rigor of state standards and tests by linking them to those set at the national level (Olson, 2007). Indeed, the NAEP gains greater importance in the current accountability policy debate under NCLB, because it is often seen as the single most reliable, valid, and readily available tool to compare results across states and to possibly confirm each state's own assessment results. However, this policy movement toward national standards and testing could increase the risk of mandating a level of learning measured by national tests such as the NAEP and transforming NAEP into a new layer of high-stakes testing.

Notes

[1]The use of scores on college entrance exams, such as the SAT and ACT, to evaluate the effect of a high school exit exam can be misleading, because a high school exit exam affects lower achieving students most. Even if the exit exam could affect college-bound students' achievement as well, using SAT or ACT results for the analysis of their achievement gains is problematic because the trends are influenced by changes in the composition of test takers.

[2]Some prior studies may have misled readers and the media by giving the impression that the accountability policies did or did not work, even when the studies lacked scientific rigor and strong evidence. Raymond and Hanushek (2003) called for more rigorous evaluation of studies by the media and policy communities. This problem of what they called "no accountability for research" also can be ameliorated when study authors themselves fully acknowledge limitations and issue strong caveats or warnings about possible misinterpretations.

[3]Between 1985 and 1995, the number of states that required passing exams for new teacher licenses doubled, from 13 to 29 for basic skills tests, from 11 to 24 for professional knowledge tests, and from 14 to 24 for subject knowledge tests (Goertz, 1986; Council of Chief State School Officers, 1996).

(text continues on page 639)

Downloaded from http://rer.aera.net at SUNY AT BUFFALO on May 1, 2009

Appendix

Descriptions of cross-state causal-comparative and correlational studies of the effects of high-stakes testing and external test-driven accountability policies on reading and math achievement

Study	Independent variables (policy)	Dependent variables (achievement)	Samples	Analytic methods	Effect sizes (standardized group mean differences)	Methods to calculate effect sizes
Fredericksen (1994)	Minimum competency testing	NAEP 1978–1986 math scores	9-, 13-, and 17-year-old students in 25 to 28 states (classified into three groups as high-, moderate-, and low-stakes states)	Age-based analysis of math gain scores in high-stakes vs. low-stakes states	Positive (null to small) 0.22 for 9-year-old math routine 0.13 for 9-year-old math nonroutine 0.08 for 13-year-old math routine 0.12 for 13-year-old math nonroutine 0.02 for 17-year-old math routine 0.05 for 17-year-old math nonroutine	Dividing the average gain score differences between high-stakes and low-stakes states by the standard deviations of student scores
Lee (1998)	Standards-based education reform including high-stakes testing policies in the 1980s	NAEP 1992 math scores	4th and 8th graders in 40 states	Cross-sectional analysis of the relationship between state policy and achievement	Negative (small) −0.33 for 4th grade math −0.46 for 8th grade math	Converting the correlations between state policy activism and state average math achievement into standardized group mean differences

(continued)

Downloaded from http://rer.aera.net at SUNY AT BUFFALO on May 1, 2009

Downloaded from http://rer.aera.net at SUNY AT BUFFALO on May 1, 2009

Appendix (continued)

Study	Independent variables (policy)	Dependent variables (achievement)	Samples	Analytic methods	Effect sizes (standardized group mean differences)	Methods to calculate effect sizes
Grissmer & Flanagan (1998)	Standards-based assessments and school accountability system adopted in the 1980s and 1990s	NAEP 1990–1996 math scores and 1992–1994 reading scores	4th and 8th graders in North Carolina and Texas vs. the average state	Case study (indirect estimation of policy effects on gain scores as unexplained by changes in student and school factors)	Positive (null to small) 0.22 for 1990–1996 8th grade math 0.25 for 1992–1996 4th grade math 0.07 for 1992–1994 4th grade reading	Subtracting the average standardized gain scores of the two high-stakes states (North Carolina and Texas) from the average of all states' standardized gain scores
Bishop et al. (2001)	School accountability (stakes for schools and teachers), MCE, EOCE	NAEP 1998 reading and 1996 math scores	4th and 8th graders in 35 to 43 states (comparing New York and North Carolina with EOCE/MCE with others)	Cross-sectional analysis of three types of accountability policy effects with control for state demographic background variables	Positive for school accountability (small) 0.49 for 1998 4th grade reading 0.22 for 1998 8th grade reading 0.27 for 1996 4th grade math 0.13 for 1996 8th grade math Insignificant for MCE 0.00 for 1998 4th grade reading 0.14 for 1998 8th grade reading 0.33 for 1996 4th grade math 0.04 for 1996 8th grade math	Dividing the regression coefficients for state accountability policy dummy variables by the standard deviations of test scores

(continued)

Appendix (continued)

Study	Independent variables (policy)	Dependent variables (achievement)	Samples	Analytic methods	Effect sizes (standardized group mean differences)	Methods to calculate effect sizes
					Positive for EOCE/MCE (medium to large) 0.79 for 1998 4th grade reading 1.24 for 1998 8th grade reading 1.20 for 1996 4th grade math 0.54 for 1996 8th grade math	
Jacob (2001)	High school graduation exams (MCEs) adopted in the late 1970s and 1980s	NELS 1988–1992 reading and math scores	8th graders and 12th graders in 15 states with high school exit exams in comparison with the other states without such exams	Longitudinal analysis of student achievement gains (with controls for student, school, and state characteristics)	Insignificant 0.04 for 1988 8th grade to 1992 12th grade math 0.001 for 1988 8th grade to 1992 12th grade reading	Dividing the regression coefficients for graduation test dummy variables by the standard deviations of 12th grade scores
Amrein & Berliner (2002)	Index of high-stakes testing policies adopted from the 1970s to the 1990s, including graduation exams, public report cards, rewards, or sanctions	NAEP 1992–2000 math, 1990–2000 math, and 1992–1998 reading scores	4th and 8th graders in 18 high-stakes testing states in comparison with the nation as a whole	Grade-based and cohort-based analyses of relative gain scores (as deviations from national average gain scores)	Mixed (small) 0.35 for 1992–2000 4th grade math 0.14 for 1990–2000 8th grade math 0.28 for 1992–1998 4th grade reading	Dividing the average gain scores of 18 high-stakes states (as deviations from the national average gain scores) by their standard deviations[b]

(continued)

Downloaded from http://rer.aera.net at SUNY AT BUFFALO on May 1, 2009

Downloaded from http://rer.aera.net at SUNY AT BUFFALO on May 1, 2009

Appendix (continued)

Study	Independent variables (policy)	Dependent variables (achievement)	Samples	Analytic methods	Effect sizes (standardized group mean differences)	Methods to calculate effect sizes
					–0.38 for 1996 4th grade to 2000 8th grade math 0.02 for 1994 4th grade to 1998 8th grade reading	
Carnoy & Loeb (2002)	External accountability policy index (5-point scale), including high school exit exam, school report cards, rewards and sanctions as of 1999–2000	NAEP 1996–2000 math achievement levels (basic and proficient)	4th and 8th graders in 25 to 37 states (the number of states varied among racial groups)	Grade-based analysis of gains in percentage basic by racial group (adjustment for baseline achievement, demographics, per pupil revenue, and changes in the population)	Positive (small to large) 0.10 for 1996–2000 4th grade White 0.77 for 1996–2000 4th grade Black 0.54 for 1996–2000 4th grade Hispanic 0.78 for 1996–2000 8th grade White 0.95 for 1996–2000 8th grade Black 1.04 for 1996–2000 8th grade Hispanic	Dividing doubled regression coefficients for accountability policy by the standard deviation of state gain scores[a]
Raymond & Hanushek (2003)	Amrein & Berliner's list of high-stakes testing states	NAEP 1992–2000 math scores	4th and 8th graders in 34 to 36 states	Grade-based analysis of gain scores (adjustment for changes in exclusion rates, spending on education, parents' education)	Positive (large) 1.16 for 1992–2000 4th grade math 0.71 for 1996–2000 4th grade math 0.79 for 1992–2000 8th grade math 0.74 for 1996–2000 8th grade math	Dividing the average difference between high-stakes states and nonaccountability states by the standard deviations of state gain scores

(continued)

Appendix (continued)

Study	Independent variables (policy)	Dependent variables (achievement)	Samples	Analytic methods	Effect sizes (standardized group mean differences)	Methods to calculate effect sizes
Rosenshine (2003)	Amrein & Berliner's list of "clear" high-stakes testing states, excluding states with changing exclusion rates	NAEP 1996–2000 math and 1994–1998 reading scores	4th and 8th graders in 20 to 26 states	Grade-based analysis of state average gain scores	Positive (small to moderate) 0.35 for 1996–2000 4th grade math 0.79 for 1996–2000 8th grade math 0.62 for 1994–1998 4th grade reading	Dividing the average gain score difference between Amrein & Berliner's clear high-stakes states and other states by the standard deviations of gain scores
Amrein-Beardsely & Berliner (2003)	"Clear" high-stakes testing states (excluding "unclear" states with increased exclusion rates)	NAEP 1996–2000 math and 1994–1998 reading scores	4th and 8th graders in 12 to 16 states	Grade-based descriptive analysis of state average gain scores	Mixed (small to large) 1.2 for 1996–2000 4th grade math 0.77 for 1996–2000 8th grade math –0.33 for 1994–1998 4th grade reading	Dividing the average gain score difference between newly identified clear high-stakes states and other states by the standard deviations of gain scores
Braun (2004)	Amrein & Berliner's list of high-stakes testing states	NAEP 1992-2000 math scores	4th and 8th graders in 33 states	Grade-based and cohort-based analyses of relative gain scores (as deviations from the national average)	Mixed (small to large) 0.96 for 1992–2000 4th grade math 0.81 for 1992–2000 8th grade math –0.67 for 1992 4th grade to 1996 8th grade math –0.31 for 1996 4th grade to 2000 8th grade math	Dividing the average gain score difference between high-stakes and low-stakes testing states by the standard deviations of gain scores

(continued)

Downloaded from http://rer.aera.net at SUNY AT BUFFALO on May 1, 2009

Downloaded from http://rer.aera.net at SUNY AT BUFFALO on May 1, 2009

Appendix (continued)

Study	Independent variables (policy)	Dependent variables (achievement)	Samples	Analytic methods	Effect sizes (standardized group mean differences)	Methods to calculate effect sizes
Lee & Wong (2004)	A composite factor of state activism in test-driven accountability policy during the 1990s	NAEP 1992–2000 math scores	8th graders in 31 states	Grade-based trend analysis of gain scores (adjusted for baseline status, demographics, schooling conditions)	Positive (small) 0.28 for 1992–2000 8th grade all 0.27 for 1992–2000 8th grade White 0.39 for 1992–2000 8th grade Black 0.18 for 1992–2000 8th grade Hispanic	Dividing adjusted regression coefficients for accountability policy by the standard deviation of test scores[c]; the effect size for each racial group was estimated in the same way
Hanushek & Raymond (2004)	A time-varying dummy variable of whether states both report results and attach consequences to school performance	NAEP 1992–2002 reading and math scores	4th and 8th graders in 42 states	Cohort-based analysis of gain scores (adjusted for changes in exclusion rates, school spending, educational attainment)	Positive (small to medium) 0.22 for all 0.21 for White 0.09 for Black 0.54 for Hispanic	Dividing regression coefficients for consequential accountability variable by standard deviations of test scores for converting the coefficients into standardized group mean differences
Nichols et al. (2006)	Time-varying measures of EPRs through reviews of state portfolios	NAEP 1990–2003 reading and math scores (grade-based analysis); 1994–2002 reading and 1992–2000 math scores (cohort-based analysis)	4th and 8th graders in 25 states	Grade-based and cohort-based analyses of gain scores	Mixed (null to medium) 0.08 for 4th grade reading Black 0.49 for 4th grade math Black 0.04 for 8th grade reading Black	Converting correlations between EPR changes and NAEP gains into standardized group mean differences

(continued)

Appendix (continued)

Study	Independent variables (policy)	Dependent variables (achievement)	Samples	Analytic methods	Effect sizes (standardized group mean differences)	Methods to calculate effect sizes
					0.00 for 8th grade math Black –0.12 for 4th grade reading Hispanic 0.63 for 4th grade math Hispanic 0.30 for 8th grade reading Hispanic 0.32 for 8th grade math Hispanic 0.20 for 4th grade reading White 0.39 for 4th grade math White 0.20 for 8th grade reading White 0.16 for 8th grade math White 0.37 for 4th to 8th grade reading Black 0.43 for 4th to 8th grade math Black –0.56 for 4th to 8th grade reading Hispanic	

(continued)

Downloaded from http://rer.aera.net at SUNY AT BUFFALO on May 1, 2009

Downloaded from http://rer.aera.net at SUNY AT BUFFALO on May 1, 2009

Appendix (continued)

Study	Independent variables (policy)	Dependent variables (achievement)	Samples	Analytic methods	Effect sizes (standardized group mean differences)	Methods to calculate effect sizes
					0.32 for 4th to 8th grade math Hispanic 0.14 for 4th to 8th grade reading White 0.06 for 4th to 8th grade math White	

Note. Studies are listed in chronological order. Only one study, that of Braun (2004), reported the effect size *d*, and it was used for this analysis directly. The effect sizes for all other studies above were calculated or converted by the author of this review on the basis of information available from the original studies. NAEP = National Assessment of Educational Progress; MCE = minimum competency exam; EOCE = end-of-course exam; NELS = National Educational Longitudinal Study; EPR = expert pressure rating.

a. A two-step move in accountability (e.g., from 1 to 3) was deemed a significant change from weak to strong accountability status, and this change was associated with gain scores. Therefore, regression coefficients (i.e., estimates of change in the gain score with a 1-unit increase in the accountability policy variable) were multiplied by 2 to estimate the average difference between weak and strong accountability states.

b. Among the 18 high-stakes states identified by the study, the numbers of states that had available data were 10 for the analysis of 8th grade gain from 1990 to 2000 and 13 for the analyses of 4th grade gain from 1992 to 2000, 4th grade reading gain from 1992 to 1998, 4th to 8th grade math gain from 1996 to 2000, and 4th to 8th grade reading gain from 1994 to 1998.

c. Regression coefficients (the estimates of yearly average gain with a 1 standard deviation increase in accountability factor score from 1992 to 2000) were adjusted by multiplying by 16 to obtain estimates of cumulative policy effect over the 8-year period between strong and weak accountability states (2 standard deviation changes).

[4]The reported sample size of 37 used for the analysis of the eighth grade mathematics gain scores for White students from 1996 to 2000 should have been 33 states. At the same time, the sample size for the analysis of the fourth grade mathematics gain scores for Whites from 1996 to 2000 should have been 35 rather than the 36 states reported by the authors. There is also a discrepancy in the list of states included in their sample for the analysis of Black and Hispanic achievement gains.

[5]Although a successive cohort comparison method that compares average student performance at the same grade level over time has been used widely in evaluating school-level academic growth, the volatility of gain scores obtained through this method is very severe (Kane & Staiger, 2002; Lee & Coladarci, 2002; Linn & Haug, 2002). Although this grade-based method should produce more reliable estimates of achievement gains at the state level, concerns about a cohort artifact confounding the gain estimates remain.

[6]The application of Rasch models to state policy survey data affords researchers several ways to check the validity of policy measures. First, policies can be hierarchically ordered on the basis of the chance of being adopted by states, and the construct validity of policy measure can be assessed by examining this obtained policy difficulty order against expected order. The adoption of a policy at one point in time may affect the likelihood that states would adopt policies at another time, and the independence of policies can be verified by examining the fit of state policy data with the model. See Lee (1997) for details regarding Rasch model application.

References

Achieve. (2004). *The expectations gap: A 50-state review of high school graduation requirements*. Retrieved January 9, 2005, from http://www.achieve.org/files/coursetaking.pdf

Adams, J. E., & Kirst, M. W. (1999). New demands and concepts for educational accountability: Striving for results in an era of excellence. In J. Murphy & K. S. Louis (Eds.), *Handbook of research on educational administration* (pp. 463–490). San Francisco, CA: Jossey-Bass.

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

Amrein, A. L., & Berliner, D. C. (2002). High-stakes testing, uncertainty, and student learning. *Education Policy Analysis Archives*, *10*(18). Retrieved June 14, 2003, from http://epaa.asu.edu/epaa/v10n18/

Amrein-Beardsley, A. A., & Berliner, D.C. (2003). Re-analysis of NAEP math and reading scores in states with and without high-stakes tests: Responses to Rosenshine. *Education Policy Analysis Archives*, *11*(25). Retrieved August 24, 2003, from http://epaa.asu.edu/epaa/v11n25/

Baker, E. L (2003). Multiple measures: Toward tiered systems. *Educational Measurement: Issues and Practice*, *22*(2), 13–17.

Benveniste, G. (1985). The design of school accountability systems. *Educational Evaluation and Policy Analysis*, *7*(3), 261–279.

Berliner, D. C., & Biddle, B. J. (1995). *The manufactured crisis: Myth, fraud, and the attack on America's public schools*. Reading, MA: Addison-Wesley.

Downloaded from http://rer.aera.net at SUNY AT BUFFALO on May 1, 2009

Betts, J. R., & Danenberg, A. (2002). School accountability in California: An early evaluation. In D. Ravitch (Ed.), *Brookings papers on education policy 2002* (pp. 123–180). Washington, DC: Brookings Institution.

Bishop, J. H., Mane, F., Bishop, M., & Moriatry, J. (2001). The role of end-of-course exams and minimum competency exams in standards-based reforms. In D. Ravitch (Ed.), *Brookings papers on education policy 2001* (pp. 267–330). Washington, DC: Brookings Institution.

Braun, H. (2004, January 5). Reconsidering the impact of high-stakes testing. *Education Policy Analysis Archives*, *12*(1). Retrieved March 10, 2004, from http://epaa.asu.edu/epaa/v12n1/

Bryk, A. (2003). No Child Left Behind, Chicago-style. In P. E. Peterson & M. R. West (Eds.), *No Child Left Behind? The politics and practice of school accountability* (pp. 242–268). Washington, DC: Brookings Institution.

Campbell, D. T., & Stanley, J. C. (1963). *Experimental and quasi-experimental designs for research*. Chicago: Rand McNally.

Carnoy, M., & Loeb, S. (2002). Does external accountability affect student outcomes? *Educational Evaluation and Policy Analysis*, *24*(4), 305–331.

Carnoy, M., Loeb, S., & Smith, T. (2001). *Do higher scores in Texas make for better high school outcomes?* (CPRE Research Report No. RR-047). Philadelphia: Consortium for Policy Research in Education.

Clotfelter, C. T., & Ladd, H. F. (1996). Recognizing and rewarding success in public schools. In H. Ladd (Ed.), *Holding schools accountable* (pp. 23–63). Washington, DC: Brookings Institution.

Cohen, D. K., & Haney, W. (1980). Minimums, competency testing, and social policy. In R. M. Jaeger & K. T. Carol (Eds.), *Minimum competency achievement testing* (chap. 1). Berkeley, CA: McCutchan.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). New York: Academic Press.

Commission on Instructionally Supportive Assessment. (2001). *Building tests that support instruction and accountability: A guide for policymakers*. Washington, DC: Author.

Council of Chief State School Officers. (1996). *Key state education policies on K–12 education*. Washington, DC: Author.

Darling-Hammond, L. (1989). Accountability for professional practice. *Teachers College Record*, *91*, 59–80.

Darling-Hammond, L. (2000). Teacher quality and student achievement: A review of state policy evidence. *Education Policy Analysis Archives*, *8*(1). Retrieved June 14, 2003 from http://epaa.asu.edu/epaa/v8n1/

Dorn, S. (1998). The political legacy of school accountability systems. *Education Policy Analysis Archives*, *6*(1). Retrieved November 5, 2003, from http://epaa.asu.edu/epaa/v6n1/

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

Elmore, R. F. (2002, September-October). Testing trap. *Harvard Magazine Forum*. Retrieved May 31, 2006, from http://www.harvard-magazine.com

Erpenbach, W. J., Forte-Fast, E., & Potts, A. (2003). *Statewide educational accountability under NCLB: Central issues arising from an examination of state accountability workbooks and U.S. Department of Education reviews under the No Child Left Behind Act of 2001*. Washington, DC: Council of Chief State School Officers.

Downloaded from http://rer.aera.net at SUNY AT BUFFALO on May 1, 2009

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

Fuller, B., Gesicki, K., Kang, E. & Wright, J. (2006). *Is the No Child Left Behind Act working? The reliability of how states track achievement* (PACE Working Paper 06-1). Berkeley: University of California.

Goertz, M. E. (1986). *State educational standards: A 50-state survey*. Princeton, NJ: Educational Testing Service. (ERIC Document Reproduction Service No. ED 275 726)

Grissmer, D., & Flanagan, A. (1998). *Exploring rapid achievement gains in North Carolina and Texas*. Washington, DC: National Education Goals Panel.

Grissmer, D., Flanagan, A., Kawata, J., & Williamson, S. (2000). *Improving student achievement: What state NAEP test scores tell us*. Santa Monica, CA: RAND.

Haney, W. (2000). The myth of the Texas miracle in education. *Educational Policy Analysis Archives*, *8*. Retrieved March 3, 2001, from http://epaa.asu.edu/epaa/v8n41/

Hanushek, E. A., & Raymond, M. E. (2004). Does school accountability lead to improved performance? *Journal of Policy Analysis and Management*, *24*(2), 297–327.

Harris, D. N., & Herrington, C. D. (2006). Accountability, standards, and the growing achievement gap: Lessons from the past half-century. *American Journal of Education*, *112*, 209–238.

Hauser, R. M. (2002). Comment by Robert M. Hauser. In D. Ravitch (Ed.), *Brookings papers on education policy 2002* (pp. 184–194). Washington, DC: Brookings Institution.

Hedges, L. V. (1990). Directions for future methodology. In K. W. Watcher & M. L. Straf (Eds.), *The future of meta-analysis* (pp. 11–26). New York: Russell Sage.

Hedges, L. V., Laine, R. D., & Greenwald, R. (1994). Does money matter? A meta-analysis of studies of the effects of differential school inputs on student outcomes. *Educational Researcher*, *23*(3), 5–14.

Heubert, J. P. (2000). Graduation and promotion testing: Potential benefits and risks for minority students, English-language learners, and students with disabilities. *Poverty and Race*, *9*(5), 1–2, 5–7.

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

Hoxby, C. M. (2002). *The cost of accountability* (National Bureau of Economic Research Working Paper 8855). Retrieved September 10, 2003, from http://www.nber.org/papers/w8855/

Jacob, B. A. (2001). Getting tough? The impact of high school graduation exams. *Educational Evaluation and Policy Analysis*, *23*(2), 99–121.

Jacob, B. A. (2003). A closer look at achievement gains under high-stakes testing in Chicago. In P. E. Peterson & M. R. West (Eds.), *No Child Left Behind? The politics and practice of school accountability* (pp. 269–291). Washington, DC: Brookings Institution.

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

Kane, T. J., & Staiger, D. O. (2002). Volatility in school test scores. In D. Ravitch (Ed.), *Brookings papers on education policy 2002* (pp. 235–284). Washington, DC: Brookings Institution.

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.

Koretz, D., & Barron, S. I. (1998). *The validity of gains on the Kentucky Instructional Results Information System (KIRIS)* (MR-792-PCT/FF). Santa Monica, CA: RAND.

Downloaded from http://rer.aera.net at SUNY AT BUFFALO on May 1, 2009

Ladd, H. F. (1999). The Dallas school accountability and incentive program: An evaluation of its impact on student outcomes. *Economics of Education Review*, *18*, 1–16.

Langenfeld, K. L., Thurlow, M. L., & Scott, D. L. (1996). *High stakes testing for students: Unanswered questions and implications for students with disabilities* (Synthesis Report No. 26). Retrieved January 10, 2005, from http://education.umn.edu/NCEO/OnlinePubs/Synthesis26.htm

Lee, J. (1997). State activism in education reform: Applying the Rasch model to measure trends and examine policy coherence. *Educational Evaluation and Policy Analysis*, *19*(1), 29–43.

Lee, J. (1998). State policy correlates of the achievement gap among racial and social groups. *Studies in Educational Evaluation*, *24*(2), 137–152.

Lee, J. (2002). Racial and ethnic achievement gap trends: Reversing the progress toward equity? *Educational Researcher*, *31*(1), 3–12.

Lee, J. (2006a). Input-guarantee vs. performance-guarantee approaches to school accountability: Cross-state comparisons of policies, resources, and outcomes. *Peabody Journal of Education*, *81*(4), 43–64.

Lee, J. (2006b). *Tracking achievement gaps and assessing the impact of NCLB on the gaps: An in-depth look into national and state reading and math outcome trends*. Cambridge, MA: The Civil Rights Project at Harvard University.

Lee, J. (2007). Revisiting the impact of high-stakes testing on student outcomes from an international perspective. In L. Deretchin & C. Craig (Eds.), *International research on the impact of accountability systems* (pp. 65–82). Lanham, MD: Rowman & Littlefield.

Lee, J., & Coladarci, T. (2002). *Using multiple measures to evaluate the performance of students and schools: Learning from the cases of Kentucky and Maine*. Orono: University of Maine.

Lee, J., & Wong, K. K. (2004). The impact of accountability on racial and socioeconomic equity: Considering both school resources and achievement outcomes. *American Educational Research Journal*, *41*(4), 797–832.

Linn, R. L. (2001). *The design and evaluation of educational assessment and accountability systems* (CSE Technical Report 539). Los Angeles: Center for the Study of Evaluation.

Linn, R. L. (2003). Accountability: Responsibility and reasonable expectations. *Educational Researcher*, *32*(7), 3–13.

Linn, R. L., Baker, E. L., & Betebenner, D. W. (2002). Accountability systems: Implications of requirements of the No Child Left Behind Act of 2001. *Educational Researcher*, *31*, 3–16.

Linn, R. L., & Haug, C. (2002). Stability of school-building accountability scores and gains. *Educational Evaluation and Policy Analysis*, *24*(1), 29–36.

Madaus, G. (1988). The influence of testing on the curriculum. In L. Tanner (Ed.), *Critical issues in curriculum: 87th yearbook of the NSSE Part 1*. Chicago: University of Chicago Press. (ERIC Document Reproduction Service No. 263 183)

Madaus, G., & Greaney, V. (1985). The Irish experience in competency testing: Implications for American education. *American Journal of Education*, *93*(2), 268–293.

Marion, S. F., White, C., Carlson, D., Erpenbach, W. J., Rabinowitz, S., Sheinker, J. (2002). *Making valid and reliable decisions in the determination of adequate yearly progress*. Washington, DC: Council of Chief State School Officers.

McLendon, M. K., Hearn, J. C., & Deaton, R. (2006). Called to account: Analyzing the origins and spread of state performance-accountability policies for higher education. *Educational Evaluation and Policy Analysis*, *28*(1), 1–24.

Downloaded from http://rer.aera.net at SUNY AT BUFFALO on May 1, 2009

National Commission on Excellence in Education. (1983). *A nation at risk: The imperative for education reform*. Washington, DC: U.S. Government Printing Office.

Newmann, F. M., King, M. B., & Rigdon, M. (1997). Accountability and school performance: Implications from restructuring schools. *Harvard Educational Review*, *67*(1), 41–74.

Nichols, S. L., Glass, G. V, & Berliner, D. C. (2006). High-stakes testing and student achievement: Does accountability pressure increase student learning? *Education Policy Analysis Archives*, *14*(1). Retrieved May 1, 2006, from http://epaa.asu.edu/epaa/v14n1/

North Central Regional Educational Laboratory. (1996). *State student assessment programs database*. Oak Brook, IL: Author.

O'Day, J. (2002). Complexity, accountability, and school improvement. *Harvard Educational Review*, *72*(3), 293–329.

Olson, L. (2007, January 11). New bills would prod states to take national view on standards. *Education Week*. Retrieved January 29, 2007, from http://www.edweek.org

Orfield, G., & Kornhaber, M. (Eds.) (2001). *Raising standards or raising barriers? Inequality and high-stakes testing in public education*. New York: Century Foundation.

Osborne, D., & Gaebler, T. (1992). *Reinventing government: How the entrepreneurial spirit is transforming the public sector*. Reading, MA: Addison-Wesley.

Peterson, P. E. (Ed.). (2006). *Generational change: Closing the test score gap*. Lanham, MD: Rowman & Littlefield.

Peterson, R. A., & Brown, S. P. (2005). On the use of beta coefficients in meta-analysis. *Journal of Applied Psychology*, *90*(1), 175–181.

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

Porter, A., & Chester, M. (2002). Building a high-quality assessment and accountability program: The Philadelphia example. In D. Ravitch (Ed.), *Brookings papers on education policy 2002* (pp. 285–315). Washington DC: Brookings Institution.

Raymond, M. E., & Haushek, E. A. (2003, Summer). High-stakes research. *Education Next*. Retrieved January 20, 2004, from http://www.educationnext.org

Rosenshine, B. (2003). High-stakes testing: Another analysis. *Education Policy Analysis Archives*, *11*(24). Retrieved December 8, 2004, from http://epaa.asu.edu/epaa/v11n24/

Rosenthal, R. (1994). Parametric measures of effect size. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 231–244). New York: Russell Sage.

Rowan, B., & Miskel, C. G. (1999). Institutional theory and the study of educational organizations. In J. Murphy & K. S. Louis (Eds.), *Handbook of research on educational administration* (pp. 359–384). San Francisco: Jossey-Bass.

Shavelson, R. J., & Towne, L. (Eds.). (2002). *Scientific research in education*. Washington, DC: National Academy Press.

Shepard, L. A. (1991). Will national tests improve student learning? *Phi Delta Kappan*, *72*, 232–238.

Skrla, L., Scheurich, J. J., Johnson, J. F., & Koschoreck, J. W. (2004). Accountability for equity: Can state policy leverage social justice? In L. Skrla & J. J. Scheurich (Eds.), *Educational equity and accountability: Paradigms, policies, and politics* (pp. 51–78). New York: Routledge Falmer.

Slavin, R. E. (2002). Evidence-based education policies: Transforming educational practice and research. *Educational Researcher*, *31*(7), 15–21.

Downloaded from http://rer.aera.net at SUNY AT BUFFALO on May 1, 2009

Swanson, C. B., & Stevenson, D. L. (2002). Standards-based reform in practice: Evidence on state policy and classroom instruction from the NAEP state assessments. *Educational Evaluation and Policy Analysis*, *24*(1), 1–28.

Valencia, R. R., Valenzuela, A., Sloan, K., & Foley, D. (2004). Let's treat the cause, not the symptoms: Equity and accountability in Texas revisited. In L. Skrla & J. J. Scheurich (Eds.), *Educational equity and accountability: Paradigms, policies, and politics* (pp. 29–38). New York: Routledge Falmer.

Wise, A. E. (1979). *Legislated learning: The bureaucratization of the American classroom*. Berkeley: University of California Press.

Author

JAEKYUNG LEE is an associate professor in the Graduate School of Education at the University at Buffalo, 409 Baldy Hall, Buffalo, NY 14260; e-mail *jl224@buffalo.edu*. His research focuses on educational policy for accountability and equity, particularly the issue of closing the achievement gap. He is the recipient of 2007 American Educational Research Association Early Career Award.

Downloaded from http://rer.aera.net at SUNY AT BUFFALO on May 1, 2009