Educational Policy
http://epx.sagepub.com/

# Evidence, Methodology, Test-Based Accountability, and Educational Policy : A Scholarly Exchange Between Dr. Eric A. Hanushek and Drs. John Robert Warren and Eric Grodsky

Eric A. Hanushek, John Robert Warren and Eric Grodsky

*Educational Policy* 2012 26: 351 originally published online 30 May 2012
DOI: 10.1177/0895904812447892

The online version of this article can be found at:
http://epx.sagepub.com/content/26/3/351

Published by:
SAGE
http://www.sagepublications.com

On behalf of:
Politics of Education Association

Educational Policy
26(3) 351–368

DOI: 10.1177/0895904812447892
http://epx.sagepub.com

# Evidence, Methodology, Test-Based Accountability, and Educational Policy: A Scholarly Exchange Between Dr. Eric A. Hanushek and Drs. John Robert Warren and Eric Grodsky

**Eric A. Hanushek[1], John Robert Warren[2], and Eric Grodsky[2]**

## Abstract

This exchange represents a follow-up to an article on the effects of state high school exit examinations that previously appeared in this journal (Warren, Grodsky, & Kalogrides 2009). That 2009 was article featured prominently in a report by the National Research Council (NRC) that evaluated the efficacy of test-based accountability systems. Hanushek (2012) was highly critical of the NRC's interpretation of the existing evidence, including Warren, Grodsky, & Kalogrides' 2009 piece. Here, Warren and Grodsky explain why they believe that Hanushek incorrectly evaluated their original article on state high school exit examinations. Hanushek then responds.

## Keywords

evidence, testing, accountability, methods

[1]Stanford University, Stanford, CA, USA
[2]University of Minnesota, Minneapolis, MN, USA

**Corresponding Author:**
John Robert Warren, University of Minnesota, Department of Sociology, 909 Social Sciences, 267 19th Ave S., Minneapolis, MN, 55455
Email: warre046@umn.edu

## Introductory Note From the Editor

It is the responsibility of editors of journals like ours to provide the educational communities of scholars and researchers, practitioners, and policy makers with appraisals of educational trends and circumstances that broaden and deepen our understanding of educational issues. As both a professional service and a public service, a scholarly deliberation like the exchange between Eric Hanushek and John Robert Warren and Eric Grodsky as follows, is intended to contribute to our knowledge and resolution of critical issues facing American education, educational research, and educational policy.

Our readers should be mindful of the genesis of this exchange. Drs. Warren and Grodsky approached me as editor of *Educational Policy* with a proposal to respond to Dr. Hanushek's critique of their research in "Grinding the Antitesting Education Ax," which appeared in *Education Next* (Spring 2012, Vol. 12, No. 2), and to invite Dr. Hanushek to reply to their account. Dr. Hanushek's critique focused on Warren's and Grodsky's research featured in an article with Demetra Kalogrides published in *Educational Policy*, "State High School Exit Examinations and NAEP Long-Term Trends in Reading and Mathematics, 1971-2004" (2009, vol. 23, no. 4). I extended the invitation to Dr. Hanushek, who graciously agreed. What follows is Warren's and Grodsky's response to Hanushek's original critique, which is then followed by Hanushek's response. There has been no back-and-forth between the authors after submitting what appears in these pages.

> As editor of *Educational Policy* it is my hope that this open and collegial consideration of how data are gathered, analyzed, and are used as a basis for educational policy making will provoke and stimulate further consideration of these issues, all for the betterment of our children's educational experiences and the greater civic good.
>
> Ana M. Martínez-Alemán, Editor

## No Axe to Grind: A Response to Hanushek

*John Robert Warren and Eric Grodsky, University of Minnesota*

Dr. Eric Hanushek (2012) recently published a critique of a 2011 National Research Council (NRC) report *Incentives and Test-Based Accountability in Education* Hout & Elliott, 2011. That NRC report sought to "review and synthesize research about how incentives affect behavior" and to "consider the implications of that research for educational accountability systems that attach

incentives to test results" (Hout & Elliott, 2011, p. 1). The NRC concluded that "test-based incentive programs. . . . have not increased student achievement enough to bring the United States close to the levels of the highest achieving countries" and that "high school exit exam programs . . . decrease the rate of high school graduation without increasing achievement" (pp. 4-5). Dr. Hanushek's main critique was that "the NRC's strongly worded conclusions are only weakly supported by scientific evidence" (pp. 49-50). In particular, he faults the NRC for relying so heavily on just one study about the impact of high school exit exams (HSEEs) on students' academic achievement: a study that we published in this forum (Warren, Grodsky, & Kalogrides, 2009).

We will not comment here on Dr. Hanushek's broader conclusions about the quality or fairness of the NRC's report. Instead, we respond here (a) to several specific claims that Dr. Hanushek made about our 2009 article, and (b) to Dr. Hanushek's broader conclusions about the state of the empirical evidence concerning the efficacy and consequences of state HSEEs. In short, we disagree with virtually all of his claims.

We have no axe to grind. As we concluded in a different 2009 article (Warren & Grodsky, 2009, p. 649): "We came to our work on exit exams not as policy advocates but as researchers. We believed that the claims proponents made about the benefits of HSEEs were just as plausible as those made by opponents of those policies. We still believe that arguments in favor of exit exams as policy levers may have merit." Our empirical findings (and our interpretations of other researchers' findings) drive our conclusions, not preconceived ideas or political views. We fully welcome healthy discussion about research design, the quality of evidence, and the interpretation of results.

## Response to Claims About Our Article

Briefly, the goal of our 2009 article was to assess the claim that states' HSEEs improve students' academic achievement. We used data on 13- and 17-year-olds' reading and mathematics achievement from the nationally representative 1971 through 2004 Long-Term Trend National Assessment of Education Progress (LTT-NAEP) combined with data about the implementation timing and level of difficulty of states' HSEEs. We found no evidence for any effects of HSEEs on achievement in either reading or mathematics at the mean or at the 10th, 20th, 80th, or 90th percentiles of the achievement distribution. In italics below are quotations from Dr. Hanushek's (2012) article; we have used ellipses to indicate when we have omitted less pertinent words or passages. In each case, our response follows.

> *The long-term NAEP . . . was designed to provide consistent score information to judge achievement of the nation as a whole. It was not designed to be used to evaluate the schools of any particular state or district.* (p. 54)

In our article, we do not make inferences or draw conclusions about any particular state, district, or school. We compare two groups of students: those who lived in states with HSEEs and those who lived in states without them. There is no reason to suppose that students in the LTT NAEP are not representative of these two groups of students. The LTT NAEP sample is a nationally representative sample.

In fact, however, the LTT NAEP samples *are* state representative over time. They just contain small numbers of students in some states in some years. This is a matter of statistical *efficiency*, not *consistency*. Every student 13 or 17 years of age attending public or private school in the United States at the time each wave of data were collected had a known, nonzero probability of being selected into the LTT NAEP sample. Aggregating students to the state level therefore provides noisy but unbiased estimates of state means, estimates that become less noisy as additional observations are added over time. By using LTT NAEP, we arrive at less efficient estimates than we otherwise might have. However, there is no statistical reason to suppose that our estimates are biased.

> *As a result, NAEP never collected in its long-term trend assessment a representative sample of students for any specific state, and the median number of tested students in each state was very small*. (p. 54)

First, in the case of 13-year-olds we always have more than 100 students per state per year, and for 17-year olds we always have at least 40. In all, our samples are quite large (in the tens or hundreds of thousands).

Second, to reiterate, this critique speaks to the *efficiency* of our estimates, and not their *bias*. As we say in our 2009 article (p. 598), "Uncertainty associated with estimates of the effect of HSEE policies across states owing to small samples of states with similar HSEEs will manifest itself in standard errors and confidence bounds. LTT NAEP will prove useful so long as the confidence bounds are small enough to capture substantively important changes in student achievement." However, those confidence bounds are not especially large. As we say later, in discussing the results (p. 605), "Our power to pick up effects of state HSEEs does not appear to be substantially constrained. Our least reliable estimates are for 17-year-olds in reading, where we are able to distinguish main effects as

small as 6.3 points (or 16% of a standard deviation) from noneffects." We have power to detect much smaller effects in our analysis of data on 13-year-olds. We detect no such effects.

> *Grodsky et al. pretend that the NAEP provides them with just that: a representative sample of students for each state. They assume that the average performance of students in each state on the long-term NAEP provides an accurate measure of the average performance of students in that state, thereby violating the first principle of statistical sampling.* (p. 54)

Consider an analogy: Imagine that we wanted to use U.S. Census data to compare the earnings of female Native Hawaiians who are exactly 26 years old to those of other 26-year-old women. That comparison would be unbiased (presuming that the U.S. Census is a nationally representative sample). Would that be an *efficient* comparison? Certainly not; there are probably only a hand-full of 26-year-old Native Hawaiian women in the Census. If one were interested in coming up with a maximally efficient estimate, one would draw a sample with about 50% Native Hawaiian female 26-year-olds and about 50% other female 26-year-olds.

Dr. Hanushek claims that we violate the "first principle of statistical sampling" by assuming that "the average performance of students in each state on the long-term NAEP provides an accurate measure of the average performance of students in that state." This is incorrect, just as in the Native Hawaiian example. If we had claimed that "the average performance of students in each state on the long-term NAEP provides an *efficient* measure of the average performance of students in that state" then we would be at fault. In fact, we were quite clear about this point.

> *Only 1 percent of the observations included in their analysis are for states that had an exit exam rated at the 9th-grade level or higher, as most current examinations are*. (p. 54)

One percent of hundreds of thousands is still a lot of students. On average, those students performed no better than otherwise similar students in states without HSEEs. Our confidence bounds around these estimates are not large.

Do state HSEEs that are now in place, or that were implemented since 2004, have larger and positive effects on achievement? Perhaps. However, in science we are confined to making claims based on the available data rather than on what we want to believe. Our analyses, using extant data available to other researchers, leave us with little reason to believe that state HSEEs enhanced student achievement—at least through 2004.

> *Any attempt to see the effects of state tests should compare the changes that occur in the states that introduce them with changes in the states that do not. But the Grodsky study effectively tosses out all the information available for the 27 states that do not have an exit examination before 2004*. (p. 54)

In fact, three quarters of the students in our sample attended school in states without an HSEE. Moreover, we do precisely what Dr. Hanushek suggests we ought to do—compare changes in states with HSEE policies to changes in states without them. We identify state HSEE effects as deviations from the temporal trajectory in average scores for states with HSEE policies compared to states without them, adjusting for difference in state average test scores over time as well as the individual-level covariates available to us (race/ethnicity, sex, parental education, and an index measuring educational resources in the home).

> *As important, the analysis does not consider any measures of state policies except for exit exams, implying that any other policy changes for the three decades between 1971 and 2004 are either irrelevant for student performance or are not correlated with the introduction and use of exit exams.* (p. 54)

This is a true statement—we do not include other measures of states' policies. It is also true that states that have implemented HSEEs tend to be states that have implemented other policies (e.g., more course requirements, higher ages of compulsory schooling). However, this does not bias our results—unless it biases them in favor of finding positive effects of state HSEEs on achievement.

Assume for the moment that our null finding—no effect of HSEEs on achievement—is driven by our failure to control for some set of state policies. For this scenario to transpire there would have to be some state policies that (a) are more likely to be implemented in states that adopt HSEEs, (b) are initially implemented at the same time as HSEEs, and (c) have a sufficiently *negative* effect on achievement to undermine the positive effects of HSEEs. That is, if those omitted state policies that are enacted in the same year as a HSEE raise achievement then our results are biased in favor of finding *positive* effects of HSEEs. We are unable to think of policies that (a) tend to get adopted when HSEEs get adopted, and (b) reduce academic achievement levels.

> *The central finding is that exit exams do not have a statistically significant effect on test scores. But this insignificance could arise because of any or all of the above-mentioned problems rather than the absence of an effect of exit exams, as the NRC committee wants us to presume*. (p. 54)

We do not claim to have conclusive evidence about the impact of state HSEEs on student achievement, nor do we presume to make claims about the potential for HSEEs to enhance student academic success. We do, however, claim to have better data than most and a sound research design for evaluating the impact of the state HSEEs that actually existed through (at least) 2004. We discuss not only significance levels but also confidence bounds. Those bounds are sufficiently modest to lead us to conclude that state HSEEs had no meaningful positive impact on student academic achievement through at least 2004.

## Response to Broader Claims About the Evidence on Exit Exams

One of the NRC's (Hout & Elliott 2011, pp. 4-5) two main conclusions is that "the evidence . . . suggests that high school exit exam programs, as currently implemented in the United States, decrease the rate of high school graduation without increasing achievement." The authors of that report make no policy claims based on that conclusion; theirs is simply a summary of the evidence. Nonetheless, Hanushek (2012, p. 52) writes that "the committee objects to state laws that require students to pass an examination for a high school diploma" (p. 49) and that "the panel strongly suggests that states that impose an exit exam should repeal this requirement" (p. 52). We see no basis for either of these claims; in our opinion, they represent a distortion of the content and purposes of the NRC's report. We challenge Dr. Hanushek to substantiate these claims with quotations or other evidence from the report—something he did not do in his 2012 article.

Hanushek faults the NRC for not considering the full body of evidence concerning the consequences of state HSEEs. Again, his claim is that the NRC's conclusion about HSEEs is "only weakly supported by scientific evidence" (pp. 49-50). He goes on to say that "some of the excluded studies use the well-regarded quasi-experimental technique known as regression discontinuity analysis" (p. 54). Hanushek suggests that "the impact of an exit examination is of special interest for exactly those students on the cusp of adequate levels of achievement. While these excluded studies are not really

appropriate for studying achievement, they tend to show little impact of exit exams on dropout behavior or graduation outcomes" (p. 54).

First, while we agree that the subpopulation of students on the cusp of the passing threshold on state HSEEs is interesting, we believe that HSEE policies are targeted to (and matter for) a much broader range of students than those who happen to barely pass or fail them. Unlike regression discontinuity designs, interrupted time series models of the sort that we and Reardon et al. (Reardon, Atteberry, Arshan, & Kurlaender, 2009) employ facilitate inferences about the effects of these policies on all students, not just on a narrow subset of them.

Second, our interpretation of the results of recently published studies based on regression discontinuity designs is markedly different than Professor Hanushek's interpretation. Recent empirical results based on regression discontinuities suggest that HSEEs reduce overall graduation rates by 1 (Ou, 2010) or 2 (Clark & See, 2011) percentage points, consistent with previous findings based on national data and an observational design (Warren, Jenkins, & Kulick, 2006). Effects appear much more negative for Hispanic and African American students and for economically disadvantaged students in general (Ou, 2010) and economically disadvantaged urban students in particular (Papay, Murnane, & Willett, 2010). For example, Papay et al. (2010) found that the difference between the (null) effect of failing the math portion of Massachusetts' HSEE on 5-year graduation rates for suburban students and the negative effects for low-income urban students was about 7 percentage points. Policies like state HSEEs that at best do no harm—or harm only marginalized populations of students—while offering no tangible benefits for achievement or other valued outcomes undermine the goals of enhancing quality and equality in educational outcomes. Dr. Hanushek fails to cite any credible evidence—indeed, any evidence at all—of any positive effects of state HSEEs on achievement. In short, had the NRC paid more attention to recent regression discontinuity-based studies it would not likely have changed its conclusion.

We welcome Professor Hanushek's engagement with our work and stand by our published results. However, we regard our project as one among many that have undermined claims about the effectiveness of HSEE policies. The more important point is borne out by the consistency of results across these studies: HSEE policies harm some students without benefiting others and should either be substantially revised or entirely abandoned.

## Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

## Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

## References

Clark, D., & See, E. (2011). The impact of tougher education standards: Evidence from Florida. *Economics of Education Review, 30*, 1123-1135.

Warren, J. R., Grodsky, E., & Kalogrides, D. (2009). State high school exit examinations and NAEP long-term trends in reading and mathematics, 1971-2004. *Educational Policy, 23*(4), 589-614.

Hanushek, E. A. (2012). Grinding the antitesting ax: More bias than evidence behind NRC panel's conclusions. *Educational Next, 12*(2), 49-55.

Ou, D. (2010). To leave or not to leave? A regression discontinuity analysis of the impact of failing the high school exit exam. *Economics of Education Review, 29*(2),171-186.

Papay, J. P., Murnane, R. J., & Willett, J. B. (2010). The consequences of high school exit examinations for low-performing urban students: Evidence from Massachusetts. *Educational Evaluation and Policy Analysis, 32*(1), 5-23.

Reardon, S. F., Atteberry, A., Arshan, N., & Kurlaender, M. (2009). *Effects of the California High School Exit Exam on student persistence, achievement, and graduation*. Palo Alto, CA: Institute for Research on Education Policy and Practice.

Warren, J. R., & Grodsky, E. (2009). Exit exams harm students who fail them—and don't benefit students who pass them. *Phi Delta Kappan, 90*(9), 645-649.

Warren, J. R., Jenkins, K. N., & Kulick, R. (2006). High school exit examinations and state-level completion and GED rates, 1973-2000. *Educational Evaluation and Policy Analysis, 28*, 131-152.

# A Flawed Analysis of Unrepresentative State Achievement Data

*Eric A. Hanushek, Stanford University*

In 2011, as political debates about the reauthorization of the No Child Left Behind Act of federal educational accountability were becoming more heated, the National Research Council (NRC) released a report about test-based accountability (Hout & Elliott, 2011). This report argued that the scientific evidence about accountability did not support common accountability statutes relevant to both schools and students. I wrote a critique of this report for nontechnical readers that highlighted both the NRC reliance on poor-quality scientific analysis and their overinterpretation of the available evidence (Hanushek, 2012). One of the two conclusions of the report focused on high school exit examinations, resting heavily on a prior paper in this journal by John Robert Warren, Eric Grodsky, and Demetra Kalogrides (Grodsky, Warren, & Kalogrides, 2009; subsequently GWK). GWK, in their part of this exchange, suggest that I was mistaken in my critique of their analysis. Unfortunately, they obfuscate several central issues and potentially mislead readers about the relevance of this study for policy.

High school exit exams (HSEE) have been used since the 1970s to assess the knowledge of graduating students. They have historically been minimum competency tests with levels of the assessments set below ninth grade. In the past decade, however, exit exams gained new popularity as the educational standards movement took hold with policy makers. In 2010, half of the states employed exit exams and, except for three, set them at the 10th-grade level or higher.[1]

The NRC report concluded that there was no evidence that exit exams led to improved student achievement, but there was evidence that exit exams led to increased school dropouts—thus raising serious questions about the advisability of using them. The single study cited to support the lack of evidence of positive effects of exit exams was that by GWK. The GWK study looks at whether achievement of 13-year-olds or 17-year-olds in reading and mathematics tends to change after states introduce high school exit exams. The analysis relies on a newly constructed database that merges repeated cross-sections from the National Assessment of Educational Progress (NAEP) between 1971 and 2004 with information on when exit exams became binding

in each state. The empirical analysis finds a statistically insignificant relationship between the presence of exit exams and achievement.

The question addressed in this discussion is how this estimation of an insignificant relationship should be interpreted.

## *Assessing the Impact of High School Exit Exams*

Estimating how exit exams affect student performance is problematic for several reasons. First, because exit exams apply uniformly to all students in a state, none of the within-state variation across students in achievement is informative.[2] Second, because states differ from each other in many ways, it is difficult simply to compare outcomes across states with and without exit exams. Third, consistent data on student achievement across states and over time are not readily available before 1992 and then not for all states until 2003. Fourth, the high-stakes use of exit exams is generally announced long in advance of their becoming binding, making it difficult to know how to "date" any potential effects of these exams.[3]

GWK structure their analysis to focus on the time of introduction of exit exams and then proceed to compare achievement differences before and after that introduction. Their statistical model relates achievement of student $i$ in state $s$ at time $t$ ($y_{ist}$) to a vector of student-level covariates ($x_{ist}$); a state fixed effect ($\gamma_s$), time ($t$), and time squared ($t^2$); the existence of a high school exit exam ($\text{HSEE}_{st}$); and an error terms ($e_{ist}$) as in

$$y_{ist} = \alpha + x_{ist}\beta + \gamma_s + \lambda_o t + \lambda_1 t^2 + \delta \text{HSEE}_{st} + e_{ist} \quad (1)$$

They employ various parameterizations of whether or not the state has an exit exam (HSEE), but for now it is sufficient just to think of an indicator variable in each state at time $t$. The statistical analysis involves estimating the unknown parameters $\alpha$, $\beta$, $\lambda_0$, $\lambda_1$, and $\delta$, but the focus of attention is restricted exclusively to the magnitude and statistical significance of $\delta$.

*a. Data issues.* The GWK estimation strategy relies on before-after comparisons of the introduction of exit exams, requiring data that precede the time when states started their exams. The common source of data for achievement comparisons across states and across time is the National Assessment of Educational Progress (NAEP), which periodically has tested students in math, reading, and other subjects over time.[4] But the NAEP testing of representative samples of students by state did not begin until 1992, when assessments were introduced at the fourth and eighth grades with voluntary state participation. Unfortunately, a majority (19) of the states using exit exams

had already introduced them by 1992, limiting the usefulness of NAEP data for the estimation of Equation 1.[5]

To circumvent this lack of consistent state data, GWK take a unique approach—decomposing information on the long-term trend (LTT) data from NAEP in order to develop state samples. The long-term trend NAEP was the original national assessment of U.S. student achievement. Since 1969, NAEP has conducted ongoing nationwide assessments of student achievement in various subject areas, including reading, writing, mathematics, science, U.S. history, and world geography. Students were tested at age 9, 13, and 17 on roughly a 4-year cycle in reading, science, and math, and more intermittently in the other subject areas.

GWK employ the student-level microdata from the LTT NAEP to create a state panel data set based on 9 cross-sections for mathematics and 10 cross-sections for reading. (Note that these are not panels that follow individual students but instead are repeated cross-sections of different students at the same age.) They then must implicitly assume that the observations for each state-year cell are representative of the performance in each state, allowing them to estimate Equation 1.

Unfortunately, the LTT NAEP data cannot support this analysis as the sampling procedure does not yield representative samples for each state. As the U.S. Department of Education reports in describing their development of a separate sampling and assessment (the "main NAEP") that would provide state data:

> Since 1990, NAEP assessments have also been conducted on the state level. States that choose to participate receive assessment results that report on the performance of students in that state. In its content, the state assessment is identical to the assessment conducted nationally. However, because *the national NAEP samples were not, and are not currently designed to support the reporting of accurate and representative state-level results*, separate representative samples of students are selected for each participating jurisdiction/state.[6] (emphasis added)

This statement contrasts rather sharply with the unsubstantiated GWK assertion in their contribution to this discussion, "In fact, however, the LTT NAEP samples *are* state representative over time" (emphasis in original).

The GWK argument is that, because there is a known probability of sampling all U.S. students, with enough time and a large enough sample, the sampling must be an unbiased measure of the performance in each state. From this technically correct statement, they suggest that it is just some small

samples gathered in each round of the LTT NAEP about which one must worry. And, viewed this way, any sampling issue is simply reduced to a matter of efficiency in estimating the parameters of Equation 1. They then use an example of looking at a small cell of the U.S. census, which is a representative sample of individuals, to illustrate their assertion.

But just because there is a known probability of sampling students in each state does not imply that the complicated, multistage cluster sampling of NAEP yields representative samples of students for each state and year. Indeed, GWK at times acknowledge this. On the other hand, GWK do not recognize that the large sample property of consistency of their statistical estimator (which they assert but probably is not true for reasons given below) may not be particularly relevant when their time series estimates rely on 9 or 10 observations and when the estimation issues turn on breaks in the individual state time series.

They do introduce the secondary argument that it really does not matter what is going on with the sampling in any state because they are just contrasting states with HSEEs and those without them and the NAEP data are most likely representative of those two groups. At the outset, they point out in their original article that states with high school exit exams do not look like a representative group of states, a fact that might influence this assertion about representativeness of NAEP samples. But that is not really the relevant issue because the identification of the HSEE parameter relies on the contrast of performance before and after introduction *in each state*.[7] For that, one would like to know that the before and after scores were somehow representative of what was happening in each state, as opposed to a very likely unrepresentative sample of students in a state that fit into the overall national and demographic sampling scheme in the different years.

The second important data issue relates to the measurement of student achievement. GWK separately estimate Equation 1 for reading and math performance of 13-year-olds and 17-year-olds. One might question whether the 13-year-old sample is relevant. Have students begun to focus on high school exit exams in middle school? Would we expect the schools to do something different than they otherwise would have done in middle school when faced with a high school exit exam? In other words, do middle schools only attempt to teach middle school math when they know it will be important down the line for graduation? On the other hand, the 17-year-old sample reflects just the portion of students who are still in school at the end of the school year when they are age 17. This selection problem arising from school dropouts varies both across states and over time, adding some other concerns to the analysis. In simplest terms, both student samples are questionable in attempting to assess the achievement impacts of HSEEs.

GWK acknowledge that there is a possible issue with the relatively small samples of students in some states. They indicate that there is a median number of students per year of 100-144 (depending on age and subject). They do not report whether there are significant numbers of state–year combinations where there are no students participating or where there is a *de minimis* number. They also do not report the frequency of finding states that appear to be all urban or all rural in any given year (and possibly switching categories across time). Such sampling issues are very relevant because it is just the state–time variation that is important for this problem and not the numbers of individual students.

*b. Analytical issues.* The biggest analytical issue—overwhelming problems of model misspecification—can be seen in Equation 1. An enormous amount of work has been done to understand the determinants of student achievement (see, for example, Hanushek, 2002). The canonical conceptual model is that individual achievement is a function of families, peers, schools, and ability. Compare this to the model of Equation 1. While GWK have data on individual family backgrounds (socioeconomic background, race, and age), their only explicit measure of school inputs is state-level existence of a high school exit exam. In other words, in their analysis no other school characteristic is relevant for student achievement. The singular and dominant influence of high school exit exams set at the eighth-grade level as a driver of student achievement is hardly a thought that either many researchers or many policy makers would be likely to espouse.

Their defense of the model specification issue does not quite go that far. GWK challenge the reader to identify some other school policy that might be correlated with the introduction and use of high school exit exams. They state: "We are unable to think of policies that (1) tend to get adopted when HSEEs get adopted and (2) reduce academic achievement levels."

Without reviewing the extensive research into student achievement, it seems sufficient simply to note that considerable work has been done on the impact of state policies about school accountability and that might likely be correlated with HSEEs.[8] The work on school accountability is even reviewed in their article. Of course, this is not the only policy change that might be relevant over the period 1971-2004. As taught in all graduate statistical methods courses, the key issue is whether omitted variables are correlated with the explanatory variables of interest, something that is hard to determine in the abstract in a complicated panel of state performance.[9] Indeed, such identification by introspection has quite gone out of favor, particularly in areas where causation is an issue and where a considerable amount of prior work exists.

We do know that since 1992, when state representative data first became available, states have followed very different patterns of achievement growth.[10] They have also followed quite different policies going beyond simple school accountability regimes. Is the combination of policy changes over the sample period of GWK (1971-2004) correlated with the use of HSEEs? It seems quite plausible that omitted policy influences are important. Moreover, it does not have to be the same set of omitted policies across all states to produce serious problems.

GWK claim that they are comparing states with and without HSEEs, but that is not apparent from Equation 1. By including state fixed effects in the model, they are restricting the estimation to the within-state variance in performance over time. Said differently, the states that never institute HSEEs cannot contribute to the identification of the main effect of HSEEs. The state fixed effects control for any policies that are constant across the sample for each state, implying that states that never introduce an exit exam or are always observed to have an exit exam provide no useful information for the estimation of the impacts of HSEEs. They are effectively throwing away any information on performance in states that never have an exit exam.[11]

There is a modified version of Equation 1 that could be interpreted in a "difference-in-difference" framework that provided comparisons across states with and without exit exams, although the requirements for identifying such a model are certainly violated by the specification problems discussed previously. Among other things, one would have to believe that the HSEE states and non-HSEE states were otherwise similar except for the introduction of the exit exams. For this it is necessary to identify the time patterns of achievement for the different states. GWK do include time and time squared in the model, but these variables clearly do not characterize the time patterns of the NAEP results or the differences among any of the 50 states. Thus, this difference-in-differences interpretation cannot resurrect the underlying model specification.

## *Policy Conclusions*

The conclusion that GWK wish to reach—and that the NRC incorporates in its report—is that the statistical insignificance of the HSEE variable in Equation 1 indicates that exit exams have no influence on student achievement. But the GWK estimation (a) uses inappropriate data that are prone to huge and systematic errors, and (b) relies on models that on their face must produce biased estimates of the relevant exam parameter. That a poorly identified parameter estimated with an imprecise and potentially unrepresentative data set yields statistically insignificant results should not

be a huge surprise—completely independent of the underlying truth of whether or not there are important achievement effects of HSEEs. Perhaps GWK should have concluded their remarks in this exchange with their frank admission, "We do not claim to have conclusive evidence about the impact of state HSEEs on student achievement, nor do we presume to make claims about the potential for HSEEs to enhance student academic success." Unfortunately, they continue beyond that to make the same case as the NRC report, based on the same unconvincing evidence of dubious scientific validity.

The prior discussion has focused on why the analytical results should not be interpreted in the way that GWK or the NRC interprets them. Yet the policy interpretations of the NRC report and of GWK in their contribution to this forum go beyond this simple interpretation. Remember that the vast bulk of the evidence of GWK refers to minimal-competency examinations at the eighth-grade level or below. Yet the current policy discussions are about the use of noticeably higher-level exams—at the 10th-grade level or above.

The evidence provided by GWK simply has little to do with the consideration of these higher-level exams. They argue that, while only 1% of their observations come from states with a ninth grade or higher exam, they have such large numbers of students that it is still a relevant finding. Again, however, the only thing that really enters into their estimation is the proportion of state-years where these higher-level tests were used—which appears very small.

The NRC report had two conclusions about the use of test-based accountability. The conclusion related to student accountability through exit exams is the subject of this discussion—and clearly comes up wanting. The other conclusion related to school accountability, discussed extensively in Hanushek (2012), suffers equally, although it is beyond this discussion.

## Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

## Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

## Notes

1. U.S. Department of Education (2011), Table 176.
2. It would potentially be possible to use variation across subsets of students to investigate whether some were more affected than others by exit exams, something the GWK pursue.

3. GWK do not address this issue, but they implicitly assume that the time when the tests become binding is the relevant date.
4. See http://nationsreportcard.gov/about.asp
5. Data on prior use of exit exams can be found in U.S. Department of Education (1995), Table 152. Presumably it would be possible to estimate Equation 1 using any states that subsequently stopped using exit exams or that changed the character of these exams, although that would introduce even further issues into the estimation. GWK do not discuss how they treat either the cessation of use of exams or the switch across levels of difficulty—something that is an issue in some of their parameterizations of exit exams in Equation 1.
6. U.S. Department of Education (2002), p. 2.
7. While their estimator averages the effects across all of the states with exit exams, it fundamentally relies on the separate within-state comparisons.
8. See, for example, Carnoy & Loeb (2002), Hanushek & Raymond (2005), or the references and conclusions of Hout & Elliott (2011).
9. There are reasons to believe that the model specification problems of GWK tend to be particularly severe in the kinds of models like Equation 1 because the variable of interest is aggregated to the state level where overall omitted policy variables are most important (Hanushek, Rivkin, & Taylor, 1996).
10. The state-representative data are available only for fourth and eighth graders, leaving open the question of variations during high school.
11. Data on students in the non-HSEE states can contribute to estimating β, the impacts of student demographics, but this is largely irrelevant for estimating δ the impact of HSEE in Equation 1.

## References

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

Hanushek, E. A. (2002). Publicly provided education. In A. J. Auerbach & M. Feldstein (Eds.), *Handbook of public economics* (Vol. 4, pp. 2045-2141). Amsterdam, Netherlands: North Holland.

Hanushek, E. A. (2012). Grinding the antitesting ax: More bias than evidence behind NRC panel's conclusions. *Education Next, 12*(2) 49-55.

Hanushek, E. A., & Raymond, M. E. (2005). Does school accountability lead to improved student performance? *Journal of Policy Analysis and Management, 24*(2), 297-327.

Hanushek, E. A., Rivkin, S. G., & Taylor, L. L. (1996). Aggregation and the estimated effects of school resources. *Review of Economics and Statistics, 78*(4), 611-627.

Hout, M., & Elliott, S. W. (Eds.). (2011). *Incentives and test-based accountability in education*. Washington, DC: National Academies Press.

U.S. Department of Education. (1995). *Digest of education statistics, 1995*. Washington, DC: National Center for Education Statistics.

U.S. Department of Education. (2002). *Digest of education statistics, 2001*. Washington, DC: National Center for Education Statistics.

U.S. Department of Education. (2011). *Digest of education statistics, 2010*. Washington, DC: National Center for Education Statistics.

Warren, J. R., Grodsky, E., & Kalogrides, D. (2009). State high school exit examinations and NAEP long-term trends in reading and mathematics, 1971-2004. *Educational Policy, 23*(4), 589-614.

## Bios

**Eric A. Hanushek** is the Paul and Jean Hanna Senior Fellow at the Hoover Institution of Stanford University. His research focuses on economic analysis of educational policy issues.

**John Robert Warren**, PhD, is Professor of sociology at the University of Minnesota and the Training Director of the Minnesota Population Center. His recent research focuses on the measurement of grade retention and high school dropout rates.

**Eric Grodsky**, PhD, is Associate Professor of sociology at the University of Minnesota. His current work focuses on inequality in higher education, including changes over time in the dimensions of academic merit most important to attendance and completion, college mismatch and the role of postsecondary remediation in degree completion.