EDUCATION POLICY ANALYSIS ARCHIVES

A peer-reviewed scholarly journal
Editor: Sherman Dorn
College of Education
University of South Florida

Volume 14 Number 31 November 20, 2006 ISSN 1068–2341

No More Aggregate NAEP Studies?

Sherman Dorn, Editor
Education Policy Analysis Archives

Citation: Dorn, S. (2006). No more aggregate NAEP studies? [editorial]. *Education Policy Analysis Archives, 14*(31). Retrieved [date] from http://epaa.asu.edu/epaa/v14n31/.

Abstract
This editorial reviews recent studies of accountability policies using National Assessment of Educational Progress (NAEP) data and compares the use of aggregate NAEP data to the availability of individual-level data from NAEP. While the individual-level NAEP data sets are restricted-access and do not give accurate point-estimates of achievement, they nonetheless provide greater opportunity to conduct more appropriate multi-level analyses with state policies as one set of variables. Policy analysts using NAEP data should still look at exclusion rates and the non-longitudinal nature of the NAEP data sets.
Keywords: accountability; multi-level analysis; multiple imputation; National Assessment of Educational Progress (NAEP).

Resumen
Este trabajo editorial examina estudios recientes sobre políticas de responsabilidad de gestión que usan datos de la Evaluación Nacional del Progreso Educativo (NAEP) y compara el uso de datos agregados de la NAEP con datos por nivel individual de la misma NAEP. Aún cuando los sets de datos por nivel individual de la NAEP son de acceso restringido y no proporcionan puntos de estimación de logro académico precisos, estos datos proporcionan una buena oportunidad para realizar análisis multinivel de las políticas educativas estatales constituidas como un set de variables. Los que hacen análisis de políticas usando los datos proporcionados por la NAEP deben siempre tener el cuidado de observar las tasas

de exclusión y la naturaleza no longitudinal de los sets de datos de la NAEP.
Palabras clave: responsabilidad de gestión; análisis multinivel; imputación múltiple; Evaluación Nacional del Progreso Educativo (NAEP).

With Marchant, Paulson, and Shunk's (2006) analysis of National Assessment of Educational Progress (NAEP) results aggregated at the state level, *Education Policy Analysis Archives* publishes its tenth article that analyzes education accountability policy using state-level NAEP data (see Amrein & Berliner, 2002; Amrein-Beardsley & Berliner, 2003; Braun, 2004; Camilli, 2000; Klein, Hamilton, McCaffrey, & Stecher, 2000, 2005; Nichols, Glass, & Berliner, 2006; Rosenshine, 2003; Toenjes, 2005). Until recently, individual-level data were unavailable, and aggregate NAEP data has served as a fundamental basis for policy discussions of high-stakes accountability.

Research using the aggregate-level data has expanded both our knowledge of accountability's effects and the questions that are worth investigating. While test scores do not capture all the consequences of high-stakes accountability, analyzing student achievement is important in deciding whether the policies have "face validity"—does high-stakes accountability influence what its advocates think is important? Carnoy and Loeb (2002) and Grissmer, Flanagan, Kawata, and Williamson (2000) used aggregate NAEP level to claim beneficial effects for high-stakes accountability. Klein, Hamilton, McCaffrey, and Stecher (2000) focused specifically on Texas, suggesting that Grissmer et al.'s analysis overestimated the effects. Amrein and Berliner (2002) argued that quasi-longitudinal measures of achievement on NAEP with two-group measures of stakes did not suggest positive consequences of high-stakes accountability. Rosenshine (2003) disagreed and the original study authors responded (Amrein-Beardsley & Berliner, 2003). Nichols, Glass, and Berliner (2006) and now Marchant, Paulson, and Shunk (2006) suggest that national evidence of the effects of high-stakes accountability is relatively weak, especially for reading, and that the only NAEP aggregate evidence supporting effects from high-stakes accountability (either for raising achievement in general or for closing the achievement gap) appears for math (also see Hanushek & Raymond, 2006, for math only).

There are four sticking points with NAEP research cited above. One methodological and substantive issue is the definition and measurement of high-stakes accountability. Amrein and Berliner (2002), Carnoy and Loeb (2002), Clarke et al. (2003), Pedulla et al. (2003), Swanson & Stevenson (2002), and Nichols et al. (2006) have worked with different measures of accountability's consequences for students and educators. State accountability policies are shifting, complex entities; measures of stakes will always include a qualitative measure of judgment combining both written policies and evidence of perceived pressures by educators (as "street-level bureaucrats;" Weatherly & Lipsky, 1977). The most intensive efforts by Nichols et al. (2006) used Torgerson's (1960) method of distilling comparative judgments into a single scale. While they had the resources to calculate such judgments for a set of state policies, they did not have long-range, year-by-year judgments. Nichols et al. then used an expert's judgment whose general judgments by state correlated highly with the Torgerson measure of accountability pressures. Given the methodological difficulties and multiple perspectives, Nichols et al. replicated a portion of their analysis using Carnoy and Loeb's (2002) scale, a step that responsible researchers in this area should follow.

A second sticking point is the non-longitudinal nature of NAEP. The National Assessment of Educational Progress samples students in each state, and there is no follow-up with individual students from assessment to assessment. Analysts have tackled this problem in various ways. The approach of Marchant et al. (2006) is perhaps typical, looking at single cross-sections, changes in a

single grade from assessment to assessment, and quasi-cohort measures from fourth grade to eighth grade four years later. The implicit reasoning of multiple approaches is that if multiple "slices" of NAEP lead to similar results, then those different slices provide confirming evidence for a conclusion. None of those approaches has the advantages of a longitudinal sampling design, but NAEP does not afford that luxury.

A third sticking point is the differential rate of exclusions from NAEP samples. To some extent, differences in aggregate achievement measures are an artifact of changing exclusion rates (Carnoy & Loeb, 2002). This conflation of exclusion rates with underlying achievement makes comparisons more difficult, whether between states, between years within a state and an individual grade, or between years and grades within an individual state (the quasi-cohort approach). Whether via multiple imputation (Rubin, 1987) or through econometric selection models, modeling the selection bias of differential rates of exclusion depends on individual-level data, which are not accessible for state-level analyses.

The fourth sticking point is the aggregate nature of freely-accessible NAEP data. The only unit of analysis available (the state) may not be appropriate either for the most commonly implied research question or for more sophisticated policy analyses. While the main research question of this growing body of research is at the state level—do state-level high-stakes testing policies lead to higher achievement?—the context of the research does not make clear whether the key measure of interest should be at the state level (aggregate achievement or some summary of the achievement gap) or whether it should be at the individual level, whether individual student achievement in itself or measures of achievement gaps at the individual level. State-level analysis phrased in terms of individual achievement—whether high-stakes testing leads to higher achievement or lower gaps for *individual* students—would be perhaps an expected slip but an ecological fallacy nonetheless. In addition, recent research on accountability strongly suggests that the local context is crucial in determining educators' responses to high-stakes accountability (e.g., Carnoy, Elmore, & Siskin, 2003; Mintrop, 2004; Mintrop & Trujillo, 2005). State-level analyses—which are important for questions about overarching policy—cannot address local context.

The last two sticking points are directly related to the aggregate nature of the existing analyses, tied to the sampling design of NAEP assessments and the perception that such sampling restricted the relevant unit of analysis to the state level. Such restrictions no longer exist. The National Assessment on Educational Progress now makes individual-level data available for restricted access by researchers. While point estimates of individual achievement are not available, the data still are useful:

> To reduce the test-taking burden on individual students, NAEP administers only a subset of items to each student. Hence, individual students' achievement is not measured reliably enough to be assigned a single "score." Instead, using Item Response Theory (IRT), NAEP estimates a distribution of plausible values for each student's proficiency, based on the student's responses to administered items and other student characteristics. When analyzing NAEP achievement data, separate analyses are conducted with the five plausible values assigned to each student. The five sets of results are then synthesized, following Rubin (1987) on the analysis of multiply-imputed data. (Lubienski, 2006, p. 8).

While securing access to and working with restricted-access data is more onerous and requires greater infrastructure support than researchers' working with aggregate data, recent research in other areas suggests the viability of using the new individual-level data for policy research (e.g., Lubienski, 2006). New software, such as *AM* (American Institutes of Research, n.d.), has the facility to work with the new individual-level sets.

Using the individual-level plausible-value data sets for NAEP would address the ecological problems of existing analyses. To some extent, the individual-level data may also address selection problems and contextual effects by allowing more sophisticated modeling and multi-level analyses. The existence of individual-level data is not a panacea. Modeling the exclusion bias will still be difficult, and the sampling design of NAEP makes identifying the proper level of contextual analysis difficult. Nor does individual-level data solve the non-longitudinal nature of NAEP assessments, and in some ways makes them worse by reducing most (but not all) analyses to cross-sections. I will leave the solutions of such problems to more sophisticated researchers. In addition, the availability of individual-level data sets does not address the question of how one measures *high stakes*.

Nonetheless, regardless of the questions and problems involved, the existence of individual-level data for NAEP creates a burden of proof for researchers who continue to rely on aggregate data. As an editor, I will look for manuscripts that use the new form of NAEP data as an opportunity to conduct more sophisticated analyses. This desire to see quantitative policy researchers use individual-level data does not imply that *Education Policy Analysis Archives* will only publish individual-level analyses in the future, but it does mean that the editor and reviewers will be looking for an acknowledgment of individual-level data and a justification for why aggregate-level analyses are superior. I suspect editors and reviewers of other journals will have similar reactions.

References

American Institutes of Research. (n.d.). *AM statistical software* [program]. Washington, DC: Author. Retrieved October 30, 2006, from http://am.air.org/.

Amrein, A.L. & Berliner, D.C. (2002). High-stakes testing, uncertainty, and student learning *Education Policy Analysis Archives, 10*(18). Retrieved October 30, 2006, from http://epaa.asu.edu/epaa/v10n18/.

Amrein-Beardsley, A., & Berliner, D. (2003). Re-analysis of NAEP math and reading scores in states with and without high-stakes tests: Response to Rosenshine. *Education Policy Analysis Archives, 11*(25). Retrieved October 30, 2006, from http://epaa.asu.edu/epaa/v11n25/.

Braun, H. (2004). Reconsidering the impact of high-stakes testing. *Education Policy Analysis Archives*, *12*(1). Retrieved October 30, 2006, from http://epaa.asu.edu/epaa/v12n1/.

Camilli, G. (2000). Texas gains on NAEP: Points of light? *Education Policy Analysis Archives*, *8*(42). Retrieved October 30, 2006, from http://epaa.asu.edu/epaa/v8n42.html.

Carnoy, M., Elmore, R., & Siskin, L. S., (Eds.). (2003). *The new accountability: High schools and high-stakes testing*. New York: RoutledgeFarmer.

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

Clarke, M., Shore, A., Rhoades, K., Abrams, L., Miao, J., & Li, J. (2003). *Perceived effects of state-mandated testing programs on teaching and learning: Findings from interviews with educators*

in low-, medium-, and high-stakes states. Boston, MA: Boston College, National Board on Educational Testing and Public Policy. Retrieved October 30, 2006, from http://www.bc.edu/research/nbetpp/statements/nbr1.pdf.

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

Hanushek, E. A., & Raymond, M. E. (2006). Early returns from school accountability. In P. E. Peterson, Ed., *Generational change: Closing the test score gap* (pp. 143-166). Lanham, MD: Rowman & Littlefield Publishers, Inc.

Klein, S. P., Hamilton, L. S., McCaffrey, D. F., & Stecher, B. M. (2000). What do test scores in Texas tell us? *Education Policy Analysis Archives, 8*(49). Retrieved October 30, 2006, from http://epaa.asu.edu/epaa/v8n49/.

Klein, S. P., Hamilton, L. S., McCaffrey, D. F., & Stecher, B. M. (2005). Response to "What do Klein et al. tell us about test scores in Texas?" *Education Policy Analysis Archives, 13*(37). Retrieved October 30, 2006, from http://epaa.asu.edu/epaa/v13n37/.

Lubienski, S. T. (2006). Examining instruction, achievement, and equity with NAEP mathematics data. *Education Policy Analysis Archives, 14*(14). Retrieved October 30, 2006, from http://epaa.asu.edu/epaa/v14n14/.

Marchant, G. J., Paulson, S. E., & Shunk, A. (2006). Relations between high-stakes testing policies and student achievement after controlling for demographic factors in aggregated data. *Education Policy Analysis Archives, 14*(30), retrieved November 15, 2006, from http://epaa.asu.edu/epaa/v14n30/.

Mintrop, H. (2004). *Schools on probation: How accountability works (and doesn't work)*. New York: Teachers College Press.

Mintrop, H. & Trujillo, T.M. (2005).Corrective action in low performing schools: Lessons for NCLB implementation from first-generation accountability systems. *Education Policy Analysis Archives, 13*(48). Retrieved October 30, 2006, from http://epaa.asu.edu/epaa/v13n48/.

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 October 30, 2006, from http://epaa.asu.edu/epaa/v14n1/.

Pedulla, J. J., Abrams, L. M., Madaus, G. F., Russell, M. K., Ramos, M. A., & Miao, J. (2003). *Perceived effects of state-mandated testing programs on teaching and learning: Findings from a national survey of teachers*. Boston, MA: Boston College, National Board on Educational Testing and Public Policy. Retrieved October 30, 2006, from http://www.bc.edu/research/nbetpp/statements/nbr2.pdf.

Rosenshine, B. (2003). High-stakes testing: Another analysis. *Education Policy Analysis Archives, 11* (24). Retrieved October 30, 2006, from http://epaa.asu.edu/epaa/v11n24/.

Rubin, D. (1987). *Multiple imputation for nonresponse in surveys.* New York, NY: Wiley.

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), 1–27.

Toenjes, L. A. (2005). What do Klein et al. tell us about test scores in Texas? *Education Policy Analysis Archives, 13*(36). Retrieved October 30, 2006, from http://epaa.asu.edu/epaa/v13n36/.

Torgerson, W. S., (1960). *Theory and methods of scaling. New York:* John Wiley.

Weatherley, R., & Lipsky, M. (1977). *Street-level bureaucrats and institutional innovation: Implementing special education reform in Massachusetts*. Cambridge, MA: Joint Center for Urban Studies of the Massachusetts Institute of Technology and Harvard University.

About the Author

Sherman Dorn
University of South Florida

Email: epaa-editor@shermandorn.com

Sherman Dorn is editor of *Education Policy Analysis Archives* and a member of the social foundations faculty in the University of South Florida College of Education.

EDUCATION POLICY ANALYSIS ARCHIVES **http://epaa.asu.edu**