11-20-2006

Education Policy Analysis Archives 14/31

Arizona State University
University of South Florida

Follow this and additional works at: http://scholarcommons.usf.edu/coedu_pub
Part of the Education Commons

Scholar Commons Citation
http://scholarcommons.usf.edu/coedu_pub/617

This Article is brought to you for free and open access by the College of Education at Scholar Commons. It has been accepted for inclusion in College of Education Publications by an authorized administrator of Scholar Commons. For more information, please contact scholarcommons@usf.edu.
No More Aggregate NAEP Studies?

Sherman Dorn, Editor
Education Policy Analysis Archives


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...
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


---

**About the Author**

**Sherman Dorn**

University of South Florida

Email: epaa-editor@shermann.dorn.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

## English-language Graduate-Student Editorial Board

<table>
<thead>
<tr>
<th>Name</th>
<th>University/Institution</th>
</tr>
</thead>
<tbody>
<tr>
<td>Noga Admon</td>
<td>New York University</td>
</tr>
<tr>
<td>Jessica Allen</td>
<td>University of Colorado</td>
</tr>
<tr>
<td>Cheryl Aman</td>
<td>University of British Columbia</td>
</tr>
<tr>
<td>Anne Black</td>
<td>University of Connecticut</td>
</tr>
<tr>
<td>Marisa Cannata</td>
<td>Michigan State University</td>
</tr>
<tr>
<td>Chad d'Entremont</td>
<td>Teachers College Columbia University</td>
</tr>
<tr>
<td>Carol Da Silva</td>
<td>Harvard University</td>
</tr>
<tr>
<td>Tara Donahue</td>
<td>Michigan State University</td>
</tr>
<tr>
<td>Camille Farrington</td>
<td>University of Illinois Chicago</td>
</tr>
<tr>
<td>Chris Frey</td>
<td>Indiana University</td>
</tr>
<tr>
<td>Amy Garrett Dikkers</td>
<td>College of St. Scholastica</td>
</tr>
<tr>
<td>Misty Ginicola</td>
<td>Yale University</td>
</tr>
<tr>
<td>Jake Gross</td>
<td>Indiana University</td>
</tr>
<tr>
<td>Hee Kyung Hong</td>
<td>Loyola University Chicago</td>
</tr>
<tr>
<td>Jennifer Lloyd</td>
<td>University of British Columbia</td>
</tr>
<tr>
<td>Heather Lord</td>
<td>Yale University</td>
</tr>
<tr>
<td>Shereeza Mohammed</td>
<td>Florida Atlantic University</td>
</tr>
<tr>
<td>Ben Superfine</td>
<td>University of Michigan</td>
</tr>
<tr>
<td>John Weathers</td>
<td>University of Pennsylvania</td>
</tr>
<tr>
<td>Kyo Yamashiro</td>
<td>University of California Los Angeles</td>
</tr>
</tbody>
</table>
Archivos Analíticos de Políticas Educativas

Associate Editors
Gustavo E. Fischman & Pablo Gentili
Arizona State University & Universidade do Estado do Rio de Janeiro

Founding Associate Editor for Spanish Language (1998—2003)
Roberto Rodríguez Gómez

Editorial Board

Hugo Aboites
Universidad Autónoma Metropolitana-Xochimilco

Dalila Andrade de Oliveira
Universidade Federal de Minas Gerais, Belo Horizonte, Brasil

Alejandro Canales
Universidad Nacional Autónoma de México

Erwin Epstein
Loyola University, Chicago, Illinois

Rollin Kent
Universidad Autónoma de Puebla. Puebla, México

Daniel C. Levy
University at Albany, SUNY, Albany, New York

María Loreto Egaña
Programa Interdisciplinario de Investigación en Educación

Grover Pango
Foro Latinoamericano de Políticas Educativas, Perú

Angel Ignacio Pérez Gómez
Universidad de Málaga

Diana Rhoten
Social Science Research Council, New York, New York

Susan Street
Centro de Investigaciones y Estudios Superiores en Antropología Social Occidente, Guadalajara, México

Antonio Teodoro
Universidade Lusófona Lisboa, Portugal

Adrián Acosta
Universidad de Guadalajara, México

Alejandra Birgin
Ministerio de Educación, Argentina

Ursula Casanova
Arizona State University, Tempe, Arizona

Mariano Fernández
Enguita Universidad de Salamanca. España

Walter Kohan
Universidade Estadual do Rio de Janeiro, Brasil

Nilma Limo Gomes
Universidade Federal de Minas Gerais, Belo Horizonte

Mariano Narodowski
Universidad Torcuato Di Tella, Argentina

Vanilda Paiva
Universidade Estadual Do Rio De Janeiro, Brasil

Mónica Pini
Universidad Nacional de San Martin, Argentina

José Gimeno Sacristán
Universidad de Valencia, España

Nelly P. Stromquist
University of Southern California, Los Angeles, California

Carlos A. Torres
UCLA

Claudio Almonacid Avila
Universidad Metropolitana de Ciencias de la Educación, Chile

Teresa Bracho
Centro de Investigación y Docencia Económica-CIDE

Sigfredo Chiroque
Instituto de Pedagogía Popular, Perú

Gaudéncio Frigotto
Universidade Estadual do Rio de Janeiro, Brasil

Roberto Leher
Universidade Estadual do Rio de Janeiro, Brasil

Pia Lindquist Wong
California State University, Sacramento, California

Iolanda de Oliveira
Universidade Federal Fluminense, Brasil

Miguel Pereira
Catedratico Universidad de Granada, España

Romualdo Portella do Oliveira
Universidade de São Paulo

Daniel Schugurensky
Ontario Institute for Studies in Education, Canada

Daniel Suarez
Laboratorio de Políticas Publicas-Universidad de Buenos Aires, Argentina

Jurjo Torres Santomé
Universidad de la Coruña, España