Using PISA Data to Measure School and System Characteristics and Their Relationships with Student Performance in Science

Laura S. Hamilton
RAND Corporation

May 2009

Send all correspondence to:
Laura S. Hamilton
RAND Corporation
4570 Fifth Avenue, Suite 600
Pittsburgh, PA 15213-2665
Voice: 412-683-2300 x4403
Fax: 412-683-2800
E-mail: laurah@rand.org

DRAFT:

This paper has not been formally reviewed and should not be cited without author permission.

INTRODUCTION

Recent efforts to improve the achievement of students in the United States and in other nations have addressed a range of contextual factors and conditions that reformers believe are likely to promote student learning. Governance reforms such as standards-based accountability and charter schools have been especially popular in recent years, reflecting a widespread belief in the power of incentives and autonomy to promote effective practices. Chapter 5 of the 2006 Programme for International Student Assessment (PISA) report, titled “School and System Characteristics and Student Performance in Science,” uses PISA data to explore relationships between students’ science achievement and a number of conditions at the school and system levels. Research-based evidence on how these various conditions relate to achievement could contribute to more effective policy and practice decisions, so this research is timely and potentially valuable. At the same time, several features of the PISA research design and data limit the extent to which strong conclusions can be made on the basis of these analyses. This paper explores these limitations and discusses implications for future efforts to understand the correlates of science learning in a cross-national context.

The focus of this paper is on the measurement of school and system characteristics and the analyses of their relationships with student achievement. There are a number of additional methodological issues that need to be consider when using international survey data. These include the methods for sampling, assigning weights, dealing with missing data, setting criteria for inclusion of special needs students, and collecting demographic data for students and schools. This paper does not address these topics. Moreover, it does not examine the content or technical quality of the achievement tests themselves. Other papers in this symposium, as well as earlier reports on PISA and other international assessments (e.g., Porter & Gamoran, 2002; Loveless, 2009), explore these topics. Despite the lack attention to the quality of the achievement measures in this paper, it should be kept in mind that high-quality tests are critical to ensuring appropriate inferences based on the kinds of modeling presented in Chapter 5. The accuracy of the translations, the appropriateness of the content for students with different cultural backgrounds, and the alignment between the test items and schools' curricula will all affect the validity of inferences drawn from analyses of relationships between achievement and school and system characteristics.

In the next section I briefly recap some of the key findings from Chapter 5 and discuss why these school and system characteristics are relevant to policy. The paper then examines a set of issues that may limit the validity of inferences from the analyses described in Chapter 5. It concludes with implications and recommendations for improving the utility of information from PISA and other large-scale international education surveys.

School and System Characteristics and Student Science Achievement: Key Findings

Chapter 5 explores six sets of policies and practices, examining their distributions within and across countries,[1] their relationships with student achievement in science, and the extent to which differences in these factors are associated with differences in the magnitude of the relationship between achievement and socioeconomic status. The selected categories include (1) admittance, selection, and grouping practices; (2) public and private management and financing of schools; (3) role of parents, including school choice; (4) accountability; (5) school management and the involvement of stakeholders in decision making; and (6) school resources.

[1] Throughout this paper, "countries" refers both to the Organisation for Economic Co-operation and Development (OECD) countries and the partner countries/economies.

Studying the implementation of these policies and practices in an international context has the potential to provide evidence that can help decision makers at all levels of the education system make good choices regarding the funding and operations of schools and school systems. Moreover, to the extent that these policies and practices can be examined jointly rather than in isolation, there may be opportunities to help policymakers think more strategically about all of the factors that influence the effectiveness of school systems rather than considering specific policies or approaches in isolation. For example, much of the research on accountability focuses on how test-score data are used to monitor, reward, and sanction schools, but this work typically fails to examine how these test-based accountability systems interact with policies regarding parental choice of schools, even though decisions about where to send a child to school are likely to be influenced by the accountability context. Chapter 5 includes separate analyses examining each category of policies and practices, along with a set of joint analyses that bring the separate categories together. Some of the key findings include the following.

- Tracking into different institutions or programs at an early age is associated with inequality in achievement as a function of student socioeconomic status (SES). In other words, the relationship between SES and achievement is larger in countries with extensive tracking. Between-school variation in student performance is largest in countries that begin tracking at an early age. The PISA data cannot provide definitive information on whether this type of tracking promotes better achievement for all students, but Poland is cited in the chapter as one example of a country that reduced tracking and experienced achievement gains for both high- and low-achieving students.

- Analyses of within-school ability grouping show that in several countries, students in schools without grouping or with grouping in only some subjects perform better than students in schools that group for all subjects.
- Students in private schools outperform public school students in most countries, but the magnitude of the relationship is reduced when controlling for students' family backgrounds, and the direction of the relationship is reversed when controlling for both family background and school-level SES.
- In 16 countries where a parent survey was administered, parental satisfaction was generally high. Competition among schools for students is associated with better performance at the system level but not at the student level when controlling for SES. Parental pressure on schools is not associated with performance after controlling for SES, and neither pressure nor competition is related to equity as measured by the achievement/SES relationship.
- Existence of a standards-based external examination is associated with better performance, though the relationship is positive but not significant after controlling for demographic and socioeconomic characteristics. This positive relationship has been documented by others, such as John Bishop (1998). Public posting of test results was also positively associated with achievement. None of the data and accountability measures was associated with differences in the relationship between achievement and SES.
- At the country level, the analysis showed positive and statistically significant relationships between greater school-level autonomy in educational content and in budgeting and student achievement. This relationship was not observed at the student

level, and there was no association between autonomy and the SES/achievement relationship.

- There is extensive variation across countries on all of the resource measures. Particularly noteworthy are the differences in access to science courses, time spent in those courses, and content covered (e.g., general science vs. biology, chemistry, or physics).
- The analyses showed several relationships between resources and achievement, especially educational resources focusing on science. Computers and learning time are also associated with the SES/achievement relationship: SES is less strongly related to achievement in schools with more computers per student and in schools with less in-class learning time.
- Several school characteristics were significantly associated with student achievement in a combined regression model after accounting for demographic characteristics. These include ability grouping within schools (negative relationship), high academic selectivity (positive relationship), public posting of achievement data (positive), and learning time at school (positive). One system-level characteristic showed a significant relationship—higher levels of autonomy in budgeting were associated with higher student achievement.

Together, these findings suggest that student achievement in science as measured by the PISA tests is related to several school and system characteristics that could be manipulated by policies enacted at the national or local levels. If the findings could be interpreted to suggest that policy levers related to these six areas could be used to improve student performance, they would provide important guidance to help policymakers and education administrators design more effective programs and policies. However, as noted earlier, the policy relevance of these findings is dependent on the extent to which the data, study design, and analyses have the characteristics necessary to support causal inferences. The remainder of this paper explores several issues related to the measurement of policies and practices as well as the approaches used to link policies and practices with student achievement data, and discusses their implications for efforts to use findings from PISA to make policy decisions.

Methodological Issues in Interpreting Results

The analyses described in Chapter 5 raise several concerns related to the data and analytic approaches used to examine relationships among policies, practices, and student achievement. It is important to point out that the chapter's authors recognize many of the limitations and are careful to acknowledge them at the beginning of the chapter as well as in the discussions of individual findings. At the same time, the presentation of results and the discussions of implications could lead readers to make conclusions that are not warranted based on the data and analyses used. For example, the first paragraph poses the question, "what can schools and school policies do to raise overall student performance?" (p. 214). Readers who do not pay attention to the analytic details or who do not understand them might overestimate the extent to which the findings provide answers to these kinds of questions.

The issues can be grouped into several categories. Each of these is discussed below, along with one or more illustrative examples from the chapter.

Interpretation of Questionnaire Items: The Issue

The use of questionnaires such as those used in PISA is common in education research. This approach to data collection facilitates the gathering of information from large, representative samples of administrators or teachers across multiple settings and is an inexpensive alternative to methods such as interviews or observations, which require in-person contact. It is probably the only feasible approach for collecting data on the scale required by PISA. It is important to recognize, however, that the use of questionnaires has some drawbacks. Perhaps most important is the observation that educators do not always interpret terms or phrases consistently or in ways that the survey developers intended (Spillane & Zeuli 1999; Hill, 2005). One group of researchers interviewed teachers about their responses to surveys and discovered that teachers sometimes rephrased questions in ways that changed their meaning (Le et al., 2006). Although the PISA questionnaires are administered to principals rather than teachers, concerns about interpretation are still relevant. The cross-national nature of PISA data collection makes these concerns especially salient because responses could be influenced by linguistic and cultural differences that are not always fully addressed despite careful attention to the quality of translation (Loveless, 2009).

Even when questionnaire items provide reasonably accurate information about the use of specific approaches or practices, they typically do not address variation in the implementation of those practices or the quality of the implementation. This criticism has been leveled frequently at teacher surveys that measure instructional practices. These surveys often ask about specific approaches such as cooperative groups or use of open-ended assessment techniques, but the usual frequency-based items cannot distinguish between teachers who use these approaches in ways that elicit complex cognitive activity and those who elicit less complex responses (Hamilton & Martinez, 2007). A similar concern applies to surveys of principals or other administrators who provide information about the adoption of policies or practices but without any details about what those look like in practice.

Examples

One example of survey results that provide somewhat ambiguous information can be found in the section on accountability. Principals were asked to respond to several questions about how achievement data are used. Information about data use could be helpful for understanding the extent to which schools and individual educators are likely to face incentives to raise achievement test scores, but the questions lack sufficient detail to get a sense of how these data actually affect educators' day-to-day work. Moreover, it is not clear whether all respondents interpreted the items about data use in consistent ways. For example, principals were asked whether achievement data were used to "evaluate teachers' performance." Forty-two percent of U.S. students are in schools where principals report using data this way, but the meaning of "evaluate" could range from relatively low-stakes, formative uses to high-stakes applications such as pay for performance. The stakes attached to the data for evaluation purposes are likely to affect teachers' responses, so the lack of information about how scores are used for this purpose makes these findings hard to interpret. There are also differences in what kinds of data are gathered—e.g., the quality of the achievement measures and the grades and subjects in which they are administered. These differences would be expected to influence educators' behaviors and the extent to which practices such as "teaching to the test" would lead to desirable or undesirable consequences. And these effects could vary within individual schools if teachers experience different testing regimes as a function of subject and grade level. An additional limitation is that the survey questions only address large-scale external assessments; they do not examine other sources of student assessment data, such as formative assessments (Black & Wiliam, 1998), that have been shown to be associated with improved instruction and achievement.

Another example comes from the section on school admittance, selection, and ability grouping. Principals were asked to report on their use of a set of possible criteria for consideration when making admissions decisions. The authors classify a school as having "high academic selectivity" if the principal reports that students' prior academic records or recommendations from prior schools are prerequisites for admission. It is likely that some schools that use these criteria have student populations that are fairly homogeneous and high-achieving, but others (such as many Catholic schools in the United States) may require prior achievement data but use cut scores that lead to a wide range of ability levels being represented among the admitted students. The content and use of prior records and recommendations is likely to vary within and across countries, making it difficult to determine what "high academic selectivity" means in practice.

This is not an exhaustive list of examples. In fact the same concerns apply to almost every analysis described in Chapter 5, and caution is warranted when making inferences about the meaning of most of the school and system characteristics covered in this chapter.

Lack of Study Design that Supports Causal Inference: The Issue

The analyses of relationships between school and system characteristics and student achievement rely on multilevel regression models using cross-sectional data. While these models are appropriate for estimating relationships in a way that takes into account the nested nature of PISA data, they do not support the kinds of causal inferences that most readers would like to make. A host of unmeasured factors could influence the magnitude and even the direction of an observed relationship between achievement and a school or system characteristic. In addition, as Raudenbush and Kim (2002) point out in the context of opportunity-to-learn measures, some of the factors that are presumed to influence achievement may be outcomes of prior learning in addition to predictors of subsequent learning. They use the example of curriculum: “A nation’s curriculum represents not only an externally imposed ‘treatment,’ but also a historically conditioned set of expectations about how much students will know at any age. The curriculum is thus an endogenous variable. Standard methods of statistical analysis generally cannot reveal the causal impact of such endogenous treatment effects” (p. 288).

While the chapter’s authors do attempt to include some controls in their regression models, most notably student as well as school-level SES, these controls alone are unlikely to address all of the possible confounding factors. The fact that PISA does not gather longitudinal achievement data for individual students makes it especially difficult to parse out important confounders. The possibility of unmeasured influences exists at the individual student, school, and country levels, which complicates the interpretation of relationships. A final concern is that even when a particular practice exerts a strong influence on achievement in one country, the adoption of that practice in another country might not lead to the same effects. Raudenbush and Kim (2002) note that inferring that what works in one country will work in others requires extrapolations beyond what the available data can support.

Examples

The discussion of school admittance policies describes the primary criteria used to determine which students attend which schools, but the consequences of how these criteria are applied are not evident from the survey responses. To illustrate, in the United States, 80% of principals listed student residence as a prerequisite or high priority, but the extent to which residential assignment contributes to segregation along racial/ethnic, socioeconomic, or achievement lines cannot be determined and undoubtedly varies across states and districts. This problem also affects the interpretation of relationships among institutional tracking, SES, and performance. In many jurisdictions there may be no official policy on institutional differentiation but other factors such as residential segregation may lead to schools that end up offering different opportunities in response to the perceived strengths and weaknesses of the students who attend those schools.

Comparisons between privately and publicly managed schools have attracted much attention in the United States, and Chapter 5 provides some additional data relevant to this comparison. However, the PISA data do not support strong conclusions regarding whether or not private schools are more effective at promoting learning. The difference in the relationship between private management and student performance when school-level SES is or isn't in the models could reflect a peer effect, which could suggest that a student is (on average) better off in a private school even if some of the benefits stem from being around wealthier peers rather than from superior instruction. It could also reflect unmeasured differences in family or student motivation or other factors that make some families more likely than others to choose private education.

Single-Year Focus of Surveys and Cumulative Nature of Assessments: The Issue

Policies and practices change over time to reflect changes in funding, cultural context, public priorities, and other factors. Moreover, even when there is stability in policies and practices in a specific school, students' exposure to these policies and practices changes as they change schools or grade levels within a school. PISA is designed to capture information about a student's exposure to specific school and system characteristics in the year in which he or she takes the PISA achievement tests, which allows researchers to examine what kind of schooling the student received just before taking the tests. PISA data cannot be used to measure students' cumulative exposure to different types of schooling.

This single-year focus might not be problematic for analyses examining achievement growth on content covered in a single year. However, PISA does not measure student growth but instead provides a measure of a student's achievement at a single point in time. Moreover, as noted by PISA analysts and other scholars (e.g., Loveless, 2009), the PISA tests are not aligned to the curriculum offered in a specific year, but instead measure cumulative knowledge and skill development that occurs over many years. The cumulative nature of these tests is almost inevitable, given the need to create measures that can be administered across different countries with different curriculum content and varying sequences in which material is presented over the course of a student's time in school. The tests are also limited in the extent to which they can measure everything that is taught in school science classes. They typically exclude content that is unique to one or a small number of countries, and tend to emphasize breadth rather than depth of coverage, thereby ignoring more complex skills such as the ability to solve problems with multiple steps (Porter & Gamoran, 2002).

In addition, some studies suggest that science achievement test items often draw on knowledge and experiences acquired outside formal school settings (e.g., National Reesearch Council, 2009; Hamilton, 1998). This may be especially true for PISA; the assessment is described in an official brochure as "forward looking, focusing on young people's ability to use their knowledge and skills to meet real-life challenges, rather than merely on the extent to which they have mastered a specific school curriculum."[2] Students' experiences in applying learning to real-life situations may be to a large degree a function of outside-of-school experiences. The alignment between PISA tests and the curriculum to which students are exposed in school will affect the strength of relationships between schooling factors and achievement. This alignment is

[2] http://www.pisa.oecd.org/dataoecd/51/27/37474503.pdf

likely to vary as a function of the types of courses offered, the amount of time students spend taking science, national priorities regarding what aspects of science to emphasize, and other factors.

Efforts to link achievement with school and system characteristics are hindered by these features of the PISA tests. We would expect a specific characteristic of the school or system measured at one point in time to exert a limited influence on students' test scores which reflect knowledge and skills gained over many years and across school-based and outside-of-school contexts. This concern is especially salient in light of the fact that some students (albeit a minority) are not even taking science during the data-collection year, and many others are in a science course for 2 hours or less per week, as discussed in the section on learning time.

Examples

The breadth and scope of the achievement test, and the lack of any pretest measure, affect all of the analyses described in this chapter. As an example, consider the findings regarding resources for learning. Average time spent in instruction in science, mathematics, and language; time for self-study or homework; and activities that are intended to promote science learning are all associated with higher levels of science achievement, whereas learning time for out-of-school lessons is negatively related. Without a pretest score or a clear link between the PISA science assessment and the science curriculum to which students are exposed in school, it is impossible to determine whether the directions and magnitudes of the coefficients in the regression models accurately reflect the effects of each of these resources.

Reliance on Principals as Source of Information about Policies and Practices: The Issue

A final challenge to understanding how science achievement is associated with school and system characteristics is PISA's reliance on principal surveys as the primary source of information about many of these characteristics (student and parent responses are also considered but to a lesser extent). Although principals may be the best source of information about decisions that occur primarily at the school level, the inability to examine within-school variation or to collect information from teachers, who are closest to the teaching and learning process, hinders efforts to understand the mechanisms through which school and system characteristics may influence learning. It is common to find large within-school variation in teachers' reports of school and classroom conditions (e.g., Hamilton et al., 2008), so treating all students in a school as if they are experiencing the same conditions results in the loss of potentially important information about their actual experiences. In addition, teachers, principals, and district administrators often respond to survey questions differently (Desimone, 2006), so collecting information from only principals may lead to an incomplete picture of the educational environment. Another way of thinking about the problem is that principals' reports provide some evidence of students' opportunity to learn (OTL), particularly through questions about resources, but other aspects of OTL, including exposure to specific topics in the classroom as well as outside-of-school experiences that may influence performance, are not measured (see Floden, 2002 for a discussion of how OTL has been conceptualized in international studies of student achievement).

Missing data from other levels of the education system, such as local education authorities or municipalities, is also problematic. In the U.S. context, Porter and Gamoran (2002) note that the importance of states in making governance decisions makes it difficult to make inferences from comparisons at the national level. Many of the domains explored in Chapter 5 address decisions that in the United States are often made at the district or state level, and as illustrated in the example below, it is not always clear how to interpret principals' responses to these questions.

Example

One puzzling finding is that more than three fourths of U.S. students are in schools whose principals reported that only schools have responsibility for establishing teachers' starting salaries. For most U.S. public schools, these decisions are made at the district level rather than at the school level, using a set of rules or guidelines such as a district-wide salary schedule. This suggests that at least in the case of U.S. principals, the interpretation of "school" in these questions might include decisions made at the district level. A similar issue is evident in U.S. principals' responses to the question about who has authority to dismiss teachers. Almost all U.S. students are in schools whose principals say that schools have this authority, but many principals across the United States complain that they have very little control over decisions about teacher dismissal due to factors such as district regulations and tenure policies. So the meaning of decision-making authority in this case is ambiguous.

Another set of questions for which data from higher levels would be informative is the set discussed in the section on stakeholder involvement. Principals may not be aware of the nature or extent of influence exerted by various groups. For example, in the United States, business and industry may affect local curriculum through efforts to influence the writing of state standards, but principals might not necessarily be aware of this influence.

The absence of data at the classroom level also has implications for many of the analyses described in Chapter 5. The chapter includes some intriguing findings related to ability grouping within schools, but the specific subjects in which grouping occurs are not specified on the principal survey, making it difficult to determine the mechanisms through which grouping could influence achievement in science. There is also no way to assess the extent of grouping that occurs within classrooms, either through formal policies or as a result of the decisions made by individual teachers. The principals' responses provide at best an incomplete picture of how schools stratify students by ability.

An additional example comes from the section on resources, which discusses students' exposure to science instruction in the year in which the data were collected. There is no information about the content of the instruction or about the extent to which students have opportunities to engage in activities that would promote the skills measured by PISA. These opportunities are likely to vary within schools as well as within and across countries. The findings regarding the SES/achievement slope and its relationships with resources are especially difficult to interpret in the absence of longitudinal test-score data and of information on within-school differences in access to these resources.

RECOMMENDATIONS

The issues discussed in this paper raise doubts about the extent to which PISA can be used to support causal inferences about education policies and practices. Other scholars have reached similar conclusions about PISA and about international assessments more generally (Loveless, 2009; Smith, 2002; Haertel, 1997). However, the results presented in Chapter 5 provide some suggestive evidence that, when combined with other data, could help inform decision making and future research. Below are some recommendations for using PISA to inform research and decision making related to science teaching and learning.

Interpret findings in the context of other literature addressing the same topics

Smith (2002) notes that some useful inferences can be made based on international surveys if inferences can be supported by findings from other research and grounded in theory.

For example, the negative relationship between ability grouping and student achievement should be examined in relation to other research on ability grouping. What does the preponderance of evidence suggest? Are there conditions under which ability grouping seems to work well? What groups of students benefit the most? To the extent possible, differences in findings across countries should also be examined in the context of country-specific literature.

Examine differences in within-country relationships to understand how contextual factors might mediate these relationships

One intriguing finding was that despite the negative relationship between ability grouping and achievement found in the overall sample, the relationship was positive in several countries including the United States. In fact, the U.S. relationship was one of the three largest (along with Korea and Poland). These cross-nation differences may stem from unmeasured differences in the nature of the practice or the way it is carried out, or from other factors that have not been considered in this study. These differences make it difficult to provide policy recommendations. For making policy decisions in the United States, a focus on the within-U.S. relationships rather than the cross-national relationships might be most informative, and supplementing these findings with data gathered through richer, qualitative or more-detailed quantitative studies could improve their utility for decision making.

Use findings as basis for further research

As Raudenbush and Kim (2002) point out, the real value of the kinds of analyses described in this report might be to provide guidance for designing within-country studies using approaches that support strong causal inference. Obviously some of the school and system characteristics explored in Chapter 5 would be difficult or impossible to study using a randomized design, but many of them could be subject to more-rigorous experimental or quasiexperimental studies in specific countries, and these kinds of studies would provide information to support or refute the hypotheses that emerge from the PISA analyses.

Explore exposure to science instruction in greater detail

Some of the strongest predictors of achievement after controlling for SES address the amount of access students have to science learning opportunities. It would be informative to look at cumulative exposure to science learning rather than exposure during a single year, especially since the PISA tests are likely to draw on skills and knowledge to which students may have been exposed in earlier years. This is especially relevant as U.S. states and other nations are rethinking their approaches to accountability. There is evidence that in the United States, the No Child Left Behind (NCLB) Act's focus on mathematics and reading has been accompanied by a decrease in science instruction, so any evidence regarding the importance of time spent in science instruction, as well as the content of that instruction, would be informative.

Compare results across the three tested subjects

The authors state that they examined mathematics and reading achievement in addition to science, but that the results were similar and therefore the chapter focuses on science. It is not clear what "similar" means. It is not surprising that predictors of science achievement also predict achievement in reading and mathematics, especially given the common focus on real-life applications and the significant reading load in all three subjects. Still, it would be helpful to examine patterns of differences to determine, for example, whether resources devoted to science instruction are more strongly related to science achievement than to achievement in other subjects, or whether ability grouping is associated with achievement in the same way across subjects.

These comparisons may be especially important for the analyses of instructional resources. In the implications section the authors note that the positive relationship with science learning time suggests that schools may want to increase time on science, and that this can be done in all schools rather than requiring resources to be shifted from some students to other students. It might, however, require resources to be shifted across subjects, and so a widespread gain in science learning might come at the expense of learning in other subjects, some of which are not measured by standardized tests. Even though the focus of this report is on science, any policy decisions should be informed by information about the likely effects on student learning in other subjects. Of course, PISA does not measure achievement in all school subjects, so this analysis would be somewhat limited but still informative.

CONCLUSION

This paper summarized the key findings from the PISA analysis of school and system characteristics, and raised some concerns about interpretations of those findings. The discussion of these issues should not be taken as an indictment of the measures and analyses reported in Chapter 5, but rather as a set of cautions that need to be considered by readers who are interested in using the findings to make decisions about policy or practice. The findings provide some intriguing evidence of how student achievement is related to the kinds of policy decisions made at the national and local levels, but they need to be treated as one incomplete source of information that should be included alongside a broader set of evidence to guide decision making.

REFERENCES

Bishop, J. (1998). *Do curriculum-based external exit exam systems enhance student achievement?* (CPRE Research Rep. RR-40). Philadelphia: Consortium for Policy Research in Education.

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education, 5*(1), 7-74.

Desimone, L.M. (2006). Consider the source: Response differences among teachers, principals, and districts on survey questions about their education policy environment. *Educational Policy, 20*(4), 640-676.

Floden, R.E. (2002). The measurement of opportunity to learn. In A.C. Porter & A. Gamoran (Eds.), *Methodological advances in cross-national surveys of educational achievement* (pp. 229-266). Washington, DC: National Academy Press.

Haertel, E.H. (1997). Exploring and explaining U.S. TIMSS performance. Paper prepared for *Learning from TIMSS: An NRC Symposium on the Results of the Third International Mathematics and Science Study.* Washington, DC: National Research Council.

Hamilton, L.S. (1998). Gender differences on high school science achievement tests: Do format and content matter? *Educational Evaluation and Policy Analysis, 20,* 179-195.

Hamilton, L.S., & Martinez, J.F. (2007). What can TIMSS surveys tell us about mathematics reforms of the 1990s? In T. Loveless (Ed.), *Lessons learned: What international assessments tell us about math achievement* (pp. 127-174). Washington, DC: Brookings.

Hamilton, L.S., Stecher, B.M., Russell, J.L., Marsh, J.A., & Miles, J. (2008). Accountability and teaching practices: School-level actions and teacher responses. In B. Fuller, M.K. Henne, & E. Hannum (Eds.), *Strong state, weak schools: The benefits and dilemmas of centralized accountability (Research in the Sociology of Education, Vol. 16*, pp. 31-66). St. Louis, MO: Emerald Group Publishing.

Hill, H.C. (2005). Content across communities: Validating measures of elementary mathematics instruction. *Educational Policy*, *19*(3), 447-475.

Le, V., Stecher, B.M., Lockwood, J.R., Hamilton, L.S., Robyn, A., Williams, V., Ryan, G., Kerr, K., Martinez, F., & Klein, S. (2006). *Improving mathematics and science education: A longitudinal investigation of the relationship between reform-oriented instruction and student achievement*. Santa Monica, CA: RAND.

Loveless, T. (2009). How well are American students learning? *The 2008 Brown Center Report on American Education*. Washington, DC: Brookings.

National Research Council. (2009). *Learning science in informal environments: Places, people, and pursuits*. Washington DC: National Academies Press.

Porter, A.C., & Gamoran, A. (2002). Progress and challenges for large-scale studies. In A.C. Porter & A. Gamoran (Eds.), *Methodological advances in cross-national surveys of educational achievement* (pp. 3-23). Washington, DC: National Academy Press.

Raudenbush, S.W., & Kim, J. (2002). Statistical issues in analysis of international comparisons. In A.C. Porter & A. Gamoran (Eds.), *Methodological advances in cross-national surveys of educational achievement* (pp. 267-294). Washington, DC: National Academy Press.

Smith, M.S. (2002). Drawing inferences for national policy from large-scale cross-national education surveys. In A.C. Porter & A. Gamoran (Eds.), *Methodological advances in cross-national surveys of educational achievement* (pp. 295-317). Washington, DC: National Academy Press.

Spillane, J.P., & Zeuli, J.S. (1999). Reform and teaching: Exploring patterns of practice in the context of national and state mathematics reforms. *Educational Evaluation and Policy Analysis, 21*(1), 1-27.