Moving Beyond Country Rankings in International Assessments: The Case of PISA

by Nancy E. Perry & Kadriye Ercikan — 2015

A main goal of this themed issue of Teachers College Record (TCR) is to move the conversation about PISA data beyond achievement to also include factors that affect achievement (e.g., SES, home environment, strategy use). Also we asked authors to consider how international assessment data can be used for improving learning and education and what appropriate versus inappropriate inferences can be made from the data.

There is widespread interest among a variety of stakeholders, including parents, teachers, policy makers, and the general public, about what and how well students are learning in educational systems around the world and how well educational systems are preparing students for life outside school (Organization for Economic Cooperation and Development, OECD, 2009). Student achievement is often monitored at the national level, but nations are increasingly interested in cross-national educational comparisons as well. Perhaps in response to increasing globalization in both social and economic terms, stakeholders want to understand their country's education system within a broader international context (OECD, 2009; 2010). What are its relative strengths and weaknesses? Is it preparing citizens to participate in a globalized economy? Is it valuing high quality learning opportunities and distributing them equitably among children and youth? Is it sufficiently resourced in terms of personnel and materials? Are teachers prepared and supported to work with diverse and high needs student populations?

The Programme for International Student Assessment (PISA) was designed by the OECD to evaluate the quality, equity, and efficiency of school systems around the world (OECD, 2010). PISA assesses 15-year-old students' reading, mathematics, and science literacy. Age 15 is chosen because students this age are nearing the end of their compulsory education in most countries and it seems an appropriate time to judge the extent to which they have acquired knowledge and skills in reading, mathematics, and science that they will need in their adult lives (OECD, 2009). PISA was first administered in 2000 and has been administered on a 3-year cycle ever since (in 2000, 2003, 2006, etc.). Each cycle foregrounds one academic domain (e.g., in 2006, two thirds of the testing time was spent on science), although all three domains are included in the assessment at each cycle.

Importantly, the PISA sells itself not just as a test of basic skills and not just as a test of knowledge. According to its developers, the OECD member countries (OECD, 2009, p. 9): "The PISA assessment takes a broad approach to measuring knowledge, skills, and attitudes [and] . . . while it does assess students' knowledge, PISA also examines their ability to reflect, and to apply their knowledge and experience to real-life issues." Also, in addition to assessing specific subject matter knowledge, the PISA collects information about students' home backgrounds, their approaches to and motivation for learning, their learning environments, and their access to and familiarity with new technologies for learning (i.e., computers). As a result, the PISA not only provides information about achievement outcomes, but also offers information from key stakeholders (e.g., students, principals, parents) about how those outcomes are related to key demographic, social, economic, and educational variables. However, the preponderance of reports involving PISA data focus on the achievement/performance variables; specifically, they focus on cross-national comparisons of achievement variables. For example, data concerning students' approaches to learning (e.g., their metacognition and strategic action) and their motivation for learning (e.g., their interest/enjoyment of subjects and tasks) are almost never reported.

Our main goal for this themed issue of *Teachers College Record* (TCR) was to move the presentation of the PISA data beyond an achievement/performance focus to also examine factors that may affect that performance and measurement of that performance, using the broad range of data available through PISA (e.g., data that can inform our understandings of how students engage with tasks like those posed by the PISA, data that informs our understandings of students' home environments and school experiences). Also we wanted to examine opportunities and challenges in using PISA data to inform education policy and practice. Therefore, when we invited authors to contribute to this issue, we asked them to focus on variables other than achievement, but related to achievement, from the PISA assessment in 2009. The PISA 2009 was administered to 15-year-old students in 75 jurisdictions (OECD, 2009), assessing their knowledge of reading, math, and science, with the primary focus on reading. The assessment consisted, primarily, of paper and pencil tasks, although a computerized assessment of reading electronic texts was carried out in a subset of jurisdictions representing a range of economies. Response formats included multiple choice, short/closed answer, as well as constructed response that allowed for a wider range of acceptable answers. As you will see, the authors who contributed to this issue took up our invitation in a variety of ways. Together, their contributions relate a wide range of factors to achievement outcomes in reading, mathematics, and science, both within and across nations.

Of course, in addition to very valuable comparative inferences, freely available international assessment data lend themselves to inappropriate interpretations, such as how correlational relationships between different factors and literacy levels within some countries can be used to identify strategies to improve learning in other countries.

Therefore, considering both appropriate and inappropriate interpretations of PISA data was our second goal for this issue. This is a main focus in Ercikan and colleagues' article, but we also asked the other authors to distinguish appropriate from inappropriate implications of their analyses as they relate to teaching and learning, policy, and/or achievement testing. Furthermore, we asked authors to provide contextual information (e.g., information about school systems or cultural practices in the countries represented in their analyses) as this is relevant to interpretations of results, and we asked authors to explain how the characteristics of PISA (the framework and design) limited or constrained their investigation of particular variables.

We believe that discussions of findings and issues across chapters are critical aspects of themed issues on focused topics such as ours. To this end, we invited two prominent scholars in the field of educational policy and educational measurement to write commentaries on the articles in this issue: David Berliner and Stephen Sireci. Finally we sought to promote TCR to an international audience by focusing on a topic that has international appeal and involving international scholars. In this spirit, we invited and obtained contributions from scholars in Canada, Germany, Korea, and the United States, some of who subsequently included multiple nations in the analyses reported in their articles.

THE ARTICLES

Ercikan, Roth, and Asil wrote a cautionary tale about using international assessment data to draw conclusions about the effectiveness of particular educational systems or generate policy for practice. Specifically they asked: Do country rankings really reflect the quality of education in different countries? And should we look to higher performing countries to identify strategies for improving students' proficiency/performance in relatively lower performing countries? They addressed these questions with examples from five countries that scored very differently on the reading portion of the 2009 PISA: Canada, Germany, Shanghai-China, Turkey, and the United States. They compared country rankings from the PISA with other quality of education indicators, such as dropout rates and student satisfaction with schools. Also they examined correlates of high performance in countries with different levels of achievement. They concluded country rankings reported from international assessments should not be used to judge the quality of education students receive in particular countries because there is heterogeneity of education outcomes and processes within countries and because basing judgments about quality of education on achievement outcomes is too limiting. Moreover, for similar reasons, and because of the nature of correlational associations, they concluded that correlates of higher performance in high-ranking countries should not be used to formulate strategies for improving educational outcomes in lower ranking countries.

Self-regulation, which involves metacognition, motivation, and strategic action, is relevant to discussions about enhancing learning and instruction (a stated goal of the PISA), and a central theme in discussions of principles for *21st Century Learning* (Dumont, Instance, & Benavides, 2010). A good deal of scholarship demonstrates self-regulation is a significant source of achievement differences among students, and interventions aimed at increasing students' knowledge and use of strategies that support self-regulation have proven efficacious in improving achievement outcomes (Zimmerman & Schunk, 2011). In this issue, Artelt and Schneider used self-report data from the 2009 PISA to examine relationships among students' metacognitive knowledge, strategy use, and reading competence, as measured in the PISA. Their analyses included data from 15 year olds in 34 OECD countries. Of interest to them was the cross-country generalizability of assumptions that substantial relationships exist among metacognitive knowledge, strategy knowledge, and reading performance. Their findings support such assumptions, but suggest metacognitive knowledge (i.e., students' awareness of personal strengths and weaknesses in relation to the demands of tasks, and their knowledge of strategies for coping with challenges that tasks present) was a particularly powerful predictor of students' reading performance in the PISA.

Similarly, Lim and colleagues (Lim, Bong, & Woo) examined whether and how factors, such as sex, access to literacy resources in the home, and parents' attitudes toward reading, related to students' attitudes toward (motivation for) reading and subsequent engagement in reading activities. They limited their analyses to a nationally representative sample of Korean adolescents in the PISA 2009 database. Data concerning students' motivation for reading, strategies used while reading, and general reading behaviors (e.g., time spent, engagement with particular types of materials) came from student self-report questionnaires. Parents provided self-report data about their attitudes toward reading, and reports of home resources for reading were collected from parents and students. The results of Lim et al.'s analyses indicated sex, home resources, and parents' attitudes were positive predictors of 15-year-old Korean students' attitudes toward reading and students with positive attitudes toward reading were more likely to report using effective strategies for reading (i.e., using memorization, elaboration, and control strategies). Parents' attitudes also predicted students' use of learning strategies and teachers' practices (reported by students) were indirectly related to students' reading behaviors. Whether and how Korean adolescents' reported attitudes and actions translated into performance on the 2009 PISA needs to be corroborated. (Numerous scholarly publications suggest what learners say they do may not provide reliable and valid data about what they actually do, c.f., Winne, Jamieson-Noel & Muis, 2002; Winne & Perry, 2000). However, the addition of even self-report measures of metacognitive, motivation, and strategy measures in international achievement testing represents a significant advancement in efforts to understand not only achievement levels of students across a wide range of international contexts, but also how differences in student motivation, interest, and learning strategies may contribute to these achievement levels.

The focus in the remaining three articles is equity, due to societal context in the case of Chiu and cross-cultural issues in testing in the case of Ruiz-Primo and Li and Solano-Flores and Wang. Ming Ming Chiu's article focuses on issues of economic equality, or inequality, to explain differences in students' mathematics achievement across jurisdictions. Specifically, he used the PISA 2009 data to study relationships among family inequality, school inequality, economic segregation, and students' mathematics

achievement in 65 of the participating jurisdictions. Importantly, his analyses considered not just how children from poor families or poor nations are at a disadvantage relative to their wealthier peers, but also how economic and resource disparities relate to outcomes for all learners in a nation. His findings indicate students in countries with high levels of economic inequality had lower *overall*[1] achievement in mathematics relative to students in countries with more egalitarian economies. Chiu uses two microeconomic theories to explain the results: rent-seeking and diminishing marginal returns. A detailed explanation of each theory is provided in Chiu's article. Here we emphasize that lower mathematics achievement scores were observed not just in countries having high rates of family inequalities, but also high rates of disparities in material resources and teacher quality across schools, and high rates of economic segregation (i.e., placing high socioeconomic status, SES, and low SES students in separate schools). Importantly, these results held for different groups of students participating in the PISA (high and low SES students; high and low achieving students). It appears inequalities did not benefit any group of students. In fact, they proved detrimental to all groups of students.

As we indicated earlier, the PISA seeks not only to evaluate students' subject specific knowledge, but also their ability to apply that knowledge to real life circumstances. Large-scale standardized assessments are often accused of presenting students with questions and problems that bear little resemblance to tasks they face in their day-to-day lives. However, embedding test items in problems in societal/cultural contexts creates challenges of its' own, given students who write the PISA are from many different countries and, consequently, societal and cultural contexts. Ruiz-Primo and Li examined this issue in their article, using data from the PISA 2006 and 2009 science tests for six countries: Australia, Canada, New Zealand, the United Kingdom, and the United States. Their findings indicate students' performance on these assessments were indeed associated with the contexts in which items were presented. They found evidence of differential performance related to item context across countries and between male and female respondents. Ruiz-Primo and Li argue that more needs to be known about how to appropriately construct items *in context*, and about relationships between context characteristics and student performance to ensure assessments are fair and valid for diverse groups of students.

Solano-Flores and Wang also examined how the presentation of science questions/problems in the PISA related to student performance. Specifically, where illustrations were presented, they were interested in whether and how the complexity of the illustrations affected observed item difficulty for students from Mexico, Shanghai-China, and the United States. Their findings revealed a negative correlation between the number of illustration features and item difficulty for students from the United States and Mexico, but not for students from Shanghai-China. Moreover, they found the magnitude and direction of the correlations they observed paralleled the three countries' rankings on the test, suggesting students' ability to interpret illustrations may have played an important role in some students' performance on the PISA science test.

In summary, the authors of this issue have taken up a wide range of topics in an effort to move discussions of the PISA data beyond cross-national comparisons of achievement/performance. They have delved deeply into data the PISA provides to begin to examine and formulate hypotheses about how and why students, who are living and learning in diverse contexts, perform as they do on tasks involving knowledge and skills for reading, math, and science. Of course the authors were limited by the data the PISA collects, as well as what kinds of analyses and inferences can be made based on such data. Primarily, the international, cross-cultural, and multilingual nature of the PISA data create challenges for researchers and policy makers that are associated with incomparability of student samples, instruments, and scores within and across countries and jurisdictions (see Ercikan et al., this issue). Importantly, incomparability of data due to these factors limits meaningful ranking of country performances as well as combining data across countries and using "country" as one of many variables in a multivariate analysis. Also, users and interpreters of the PISA data need to exercise caution when interpreting country rankings as indicators of effectiveness of education systems. Ercikan et al. (this issue) present several examples that demonstrate problems with such interpretations, including the inadequacy of international assessment scores as indicators of the quality of education and education systems in a country, and heterogeneity of education outcomes within countries.

Two more assessment design issues in PISA limit making two types of inferences that can be informative for policy and practice. First, there is the lack of classroom level data in the PISA, because 15 year olds are sampled across different grades within a school, so students come from different classrooms. As a result, teacher surveys that could provide valuable information for connecting instructional processes to student achievement levels are not available and researchers are limited in the kinds of analyses they can conduct to inform classroom practices. Second, the cross-sectional nature of the PISA data prohibits making causal claims about "how things work" or why some education systems, policies and strategies appear to be more effective than others. Without longitudinal data or experimental designs such claims are not possible. Correlational investigations that focus on identifying variables and strategies that make some countries more successful than others on international assessments are some of the most common uses of such data. Ercikan et al. show that the variables with high correlations with achievement, for example attitudes towards reading with reading scores, may be highly correlated in both high- and low-performing countries, and in fact low-performing countries may have higher standings on such variables than high-performing countries. Therefore, identifying what policies or practices are responsible for "success" in high-performing countries is more complex than looking at correlational relationships. Furthermore, even if the strategies used in successful nations were emulated and implemented in less successful nations, an expectation of improvement based solely on those strategies would be naïve at best.

As we write this introduction and put the finishing touches on this themed issue, we are aware that reports from the 2012 PISA are emerging (OECD, 2013). This fifth wave of data collection, like the others, assessed 15-year-old students' facility in reading, math, and science, this time with a major focus on mathematical literacy (65 countries participated). Consistent with previous waves of the assessment, new measures have been added to enrich the PISA database (e.g., students completed a computerized assessment of problem solving and an assessment of financial literacy and responded to an educational career questionnaire, which covers students' perceived preparation for

future careers). How should researchers analyze such data to inform policy and practice in meaningful ways? And how can findings from this wave of the PISA be used to inform measurement in the next wave to continually improve the information it provides?

Based on our work on this issue and consideration of appropriate and inappropriate uses of the PISA data, we offer several recommendations. First, as Ercikan et al. (this issue) indicate, comparability of scores needs to be investigated before any cross-country analyses are conducted (see Ercikan & Lyons-Thomas, 2013, for methodologies for investigating such comparability). Second, researchers need to resist using causal language when describing correlational relationships. In particular, we need to resist the temptation to draw solid conclusions, or proffer recommendations about what strategies will improve education systems based on correlational relationships and incomparable data. Rather we should use correlational findings to guide small-scale in-depth studies that examine relationships in particular contexts and countries.

Third, the OECD and other organizations that develop international assessments can take steps to improve the potential for these assessments to inform policy and practice in meaningful ways. In particular, they should address the issue of comparability of data across countries, languages, and cultures by reporting research results that present levels of comparability of student assessment and survey data. Currently, very little information is provided about comparability and, in particular, no information about incomparability has been reported, despite the high levels of incomparability researchers find when analyzing data from these assessments. Also, guidelines that safeguard appropriate and inappropriate uses of the data are needed; particularly, guidelines concerning what kinds of analyses will lead to identifying effective policy and practice strategies. As the researchers in this special issue show, simple correlational designs that focus on high-ranking countries will not lead to meaningful inferences or, likely, changes.

Finally, if one goal of international assessments, such as the PISA, is identifying strategies to improve teaching and learning in schools, why not address this issue directly. Currently these assessments provide direct measures of students' learning, but rely on indirect measures (i.e., students' and others' reports) about factors generally believed to affect learning, which have limited reliability and validity (Winne et al., 2002; Winne & Perry, 2000). Hypotheses about effective strategies for improving learning could be tested as part of the survey based research, *in situ*. For example, technology provides opportunities for examining how students use metacognitive strategies while responding to items/solving problems on a particular assessment, rather than relying on what they say they do through self-reports. The result would be a direct measure of the relationship between strategy use and problem solving and a deeper understanding about what learners actually do in the context of test taking and problem solving.

Ideally, PISA and other international assessments gather rich data to inform research, teaching, and policy agendas about many aspects of education around the world. Each wave of data collection presents an opportunity and a responsibility to improve on our learning from the last. We trust the articles and commentary in this issue have moved the presentation of the PISA data beyond an almost exclusively achievement/performance focus and generated a wide range of hypotheses to test about factors, both individual and contextual, that contribute to achievement. Also, we trust our critical consideration of appropriate and inappropriate uses of the PISA data will advance methods, analyses, and interpretations applied to the PISA and other international assessments in the future.

Notes

1. Emphasis is Chiu's.

References

Dumont, H., Istance, D., & Benavides, F. (Eds.). (2010). *The nature of learning: Using research to inspire practice. Practitioner guide from the Innovative Learning Environments Project*. Paris: OECD, Centre for Educational Research and Innovation. Retrieved from http://www.oecd.org/edu/ceri/50300814.pdf

Ercikan, K., & Lyons-Thomas, J. (2013). Adapting tests for use in other languages and cultures. In K. Geisinger (Ed.), *APA handbook of testing and assessment in psychology* (Vol. 3, pp. 545–569). Washington, DC: American Psychological Association.

OECD. (2009). *PISA 2009 assessment framework: Key competencies in reading, mathematics and science*. Retrieved from http://www.oecd.org/pisa/pisaproducts/44455820.pdf

OECD. (2010). *PISA 2009 results: Learning to learn—student engagement, strategies, and practices (Volume III)*. Retrieved from http://www.oecd.org/pisa/pisaproducts/48852630.pdf

OECD. (2013). *PISA 2012 assessment and analytic framework: Mathematics, reading, science, problem solving and financial literacy*. Retrieved from http://www.oecd.org/pisa/pisaproducts/PISA%202012%20framework%20e-book_final.pdf

Winne, P. H., Jamieson-Noel, D. L., & Muis, K. (2002). Methodological issues and advances in researching tactics, strategies, and self-regulated learning. In P. R. Pintrich & M. L.

Maehr (Eds.), *Advances in motivation and achievement: New directions in measures and methods* (Vol. 12, pp. 121–155). Greenwich, CT: JAI Press.

Winne, P. H., & Perry, N. E. (2000). Measuring self-regulated learning. In M. Boekarts, P. Pintrich, & M. Zeidner (Eds.), *Handbook of self-regulation* (pp. 532–566). Orlando, FL: Academic Press.

Zimmerman, B. J., & Schunk, D. H. (2011). Self-regulated learning and performance: An introduction and an overview. In B. Zimmerman & D. Schunk (Eds.), *Handbook of self-regulation of learning and performance* (pp. 1–12). New York: Routledge.

Purchase Reprint Rights for this article or review