Chapter 8
Prospects for the Future: A Framework and Discussion of Directions for the Next Generation of International Large-Scale Assessments

Henry Braun

Introduction

There is an old adage, "Be careful what you wish for." In the case of education policy, the old lament that the results of international large-scale assessments (ILSAs) were not a "front burner" issue has been replaced by the lament that they are now too politicized. Whether or not this is the case, it is certainly true that education policy debates stemming from international comparisons have attained unprecedented prominence, partly because of the ascendancy of the human capital model of competitive advantage among nations. In fact, in some countries, the reports of ILSAs have been key drivers of reform. The continuing expansion of the number of participating jurisdictions testifies to their global importance. Indeed, ILSAs are seen as providing unique, credible information that can—and should—inform broad policy decisions. In this landscape, holding a conference in March 2011 in Princeton, NJ, on the role of ILSAs in education policy was both timely and much needed. As this volume reveals, a broad range of topics was covered and different suggestions for future innovations put forward. The principal aim of this chapter is to offer a preliminary framework for considering ILSA-related issues, and to situate the chapters of this book—based on presentations given at the conference at Educational Testing Service in Princeton—within this framework. It concludes with some thoughts on future directions.

The largest and most influential global ILSAs are Trends in International Mathematics and Science Study (TIMSS) and Progress in International Reading Literacy Study (PIRLS), both sponsored by the International Association for the Evaluation of Educational Achievement (IEA), and the Programme for International Student Assessment (PISA) and the International Adult Literacy Survey (IALS), both spon-

H. Braun (✉)
Center for the Study of Testing, Evaluation and Education Policy, Lynch School of Education
Boston College, 140 Commonwealth Ave, 02467 Chestnut Hill, MA, USA
e-mail: braunh@bc.edu

M.von Davier et al. (eds.), *The Role of International Large-Scale Assessments: Perspectives from Technology, Economy, and Educational Research,*
DOI 10.1007/978-94-007-4629-9_8, © Springer Science+Business Media Dordrecht 2013

sored by the Organisation for Economic Co-operation and Development (OECD). At the conference there was also discussion of the International Civics and Citizenship Survey (ICCS) and the forthcoming Programme for the International Assessment of Adult Competencies (PIAAC). It was noted that an important role is played by regional large-scale assessments confined to nations in western, southern, and eastern Africa, and Latin America and the Caribbean. Although there are certainly policy issues specific to each, this chapter aims to address issues that apply to most, if not all, global ILSAs.

To begin at the beginning, the primary purpose of education is to adequately prepare all children to lead productive, satisfying lives that contribute to the common good. The role of education policy is to design, manage and monitor the education system so it accomplishes its purpose. Braun and Kanjee (2006) posited that this purpose subsumes four component goals, namely: access, quality, effectiveness, and efficiency. ILSAs typically have been used to address the goals of quality and effectiveness of educational systems. In particular, they help to answer three key questions:

1. What are the essential skills, dispositions and habits of mind required for success in the 21st century?
2. In view of the response to No. 1, how does each nation fare in comparison to other participating nations or jurisdictions?
3. What can be expected with respect to growth over time and attainment of these essential precursors to success?

With respect to quality, the rigorous and intensive process that precedes agreement on the blueprint for an ILSA represents an international consensus on valued outcomes for the focal cohort of students. Individual countries can examine their curricula to gauge alignment with these outcomes. Turning to effectiveness, comparisons with other countries with respect to both current level and trend provide at least a rough indication of the relative effectiveness of a country's education system. (Of course, more nuanced interpretations require due consideration of contextual differences.) As far as growth over time, the spectrum of results offers nations a choice of targets, both short term and long term.

Thus, the answers to the three questions can inform policymakers' deliberations. To this point, Ritzen (this volume) provides empirical evidence on the differential impact of PISA 2006 results on policy formation across participating countries. Not surprisingly, evidence, however credible and relevant, is not sufficient to drive macro-level educational policy.

More recently, ILSAs have also been used to address one aspect of the efficiency goal. In particular, various authors have sought to identify some characteristics common to the education systems of the jurisdictions that are at or near the top of the league tables, or have achieved substantial and sustained improvement in their standings over time (see, for example, Paine and Schleicher 2011). The implication is that other jurisdictions would do well to emulate these exemplars. I will return to this point below.

ILSAs: Theory of Action

It is evident that the primary contribution of ILSAs is to facilitate direct international comparisons of achievement; that is, in the absence of a common assessment, each nation's system remains "hermetically sealed," and it is well nigh impossible to make meaningful comparisons among them. Differences in high school completion rates, for example, are potentially confounded by differences in requirements, economic conditions, and so on. Thus, policy leaders are free to make assertions regarding their nation's relative standing in regard to educational achievement without fear of contradiction.

With this in mind, Ritzen (this volume) argues for the importance of the transparency provided by ILSAs and suggests different mechanisms by which they can serve as agents for change. Of course, transparency can be a double-edged sword (as if WikiLeaks didn't demonstrate that sufficiently). In the present instance, the most common presentation of ILSA results is in the form of a ranking of jurisdictions based on score means, the so-called league tables. These rankings can be overinterpreted or misinterpreted, with possibly negative consequences. What is called for is a more nuanced examination of the results at various levels of aggregation—but this is rarely done by reporters, pundits, or legislators. Although the sponsors not only publish massive tomes after each administration to provide supplemental analyses and greater insight, but also supply data files for secondary analyses, these rarely get the attention of the league tables and the accompanying commentary. A key issue, then, is how ILSAs can evolve both to mitigate negative outcomes and to better contribute to constructive change.

But how can transparency lead to improvements in education? Theoretically, the process should work like this: The surveys generate and disseminate widely accepted evidence on the relative performance of different jurisdictions on relevant constructs such as student knowledge and skills in reading, mathematics, or science. This "transparency" spurs reflection and review on the part of government officials, policymakers and other stakeholders in education. A consensus is reached on appropriate modifications to policy and practice that are informed, at least in part, by the policies and practices of the most successful jurisdictions. Moreover, the publicity resulting from the release of the results on a fixed cycle supports the political will to allocate sufficient resources over a long period of time to achieve sustainable improvement.

What are the essential conditions for such a theory to approximate reality? There are at least five. They are as follows:

1. The reported outcomes are considered credible, relevant, and sufficiently accurate.
2. There is acknowledgment of the correspondence between these outcomes and the national goals.
3. The interpretations of the outcomes, both absolutely and comparatively, are approximately correct.

4. Stakeholders are inspired (or spurred) by the results, as well as the accompanying public reaction, to propose new policies and allocate (or reallocate) resources.
5. Policymakers maintain a sustained but flexible focus on these policies.

Let us consider each one in turn, with references to chapters in this volume as appropriate.

Credibility and Relevance

Before each administration, there is a lengthy process, which typically involves all participating jurisdictions, to achieve consensus on the operational definitions of the target constructs and carry out a test development process that results in an instrument appropriate to a heterogeneous set of student populations comprising many different educational, cultural, and linguistic traditions. Both the rigor of the process and its products contribute to the credibility of the outcomes. In addition, because comparability is the touchstone for utility, such factors as sample selection and degree of participation, accuracy, and appropriateness of the translations/adaptations, candidate motivation, and fidelity to administrative protocols are addressed and monitored. Although these factors were not central to any of the presentations at the conference, the impact of any major changes in the design of an ILSA on these factors would have to be evaluated. Relevance is supported by the rationale proposed for each target construct, which includes an argument linking proficiency to success in further academic studies and/or in the workplace and civic life (Kirsch et al. 2007). Clearly, doubts about credibility undermine the argument for relevance.

Conference presenters did address different facets of both credibility and relevance. With respect to the latter, Hanushek and Woessman (this volume) argue that the core cognitive skills measured by ILSAs are key components of human capital and assert the importance of the direct measurement of skills in contrast to statistics on proxies for achievement, such as educational attainment and the like. There are at least two main difficulties with distal indicators such as educational attainment. First, they are not comparable across jurisdictions and, second, there is wide variation in the distribution of proficiency at each level of attainment. See, for example, results from the National Adult Literacy Survey (1993). Further, the authors cite empirical findings that relate country-level variation in human capital to differences in economic growth and development. They do acknowledge, however, that returns to skills vary by country due to differences in such factors as level of development, political structure, cultural issues, and the like.

With respect to credibility, critics of standardized testing typically focus on the twin criteria of depth and breadth. The former is usually framed in terms of construct representation. That is, the tests fail to address the more complex facets of the target constructs, leading to an incomplete, and too optimistic, portrait of achievement. With regard to breadth, the argument is that the target constructs are too nar-

rowly construed, with the consequence that important skills do not receive the necessary attention and resources.

Presenters addressed the issue of breadth and credibility, making the case for particular ensembles of constructs: Torney-Purta and Amadeo (this volume) speak to the importance of civic engagement and citizenship, while Levin (this volume) speaks to noncognitive skills. Although the authors certainly acknowledge the enduring importance of foundational skills, they argue that other constructs deserve considerably more attention if we are to capture the full spectrum of human capital relevant to success in the 21st century.

The chapter by Torney-Purta and Amadeo argues for looking beyond purely economic considerations to measuring dispositions related to civil society and participatory democracy. They provide a useful review of past assessments, making the case that they attained a high level of psychometric quality and, moreover, that secondary analyses of the results has yielded important insights with respect to both crossnational and subnational comparisons. In particular, it has been possible to identify multidimensional profiles of individuals with distinctly different beliefs and attitudes. Such findings complement the empirical findings that higher levels of cognitive skills are associated with greater participation in the economic, social, and civic life of the state.

The chapter by Levin urges that so-called noncognitive skills be assessed along with cognitive skills because there are strong theoretical and empirical rationales for the important roles that these skills play in individual success both in school and work. Thus, these skills should be considered integral components of human capital and deserving of attention. He also cites evidence that schooling influences the development of these skills, strengthening the argument that they should be included as target constructs in the design of school-based surveys. Further, Levin makes the important point that neglect of these constructs can skew policy choices.

As is the case with the assessment of civic dispositions, there is ample precedent for including noncognitive skills in ILSAs. For example, an instrument labeled an "Inventory of Student Approaches to Learning" was administered as part of PISA 2000 (OECD 2003). The instrument assessed such constructs as motivation, self-related beliefs, and approaches to learning. Psychometric and other analyses indicate that the instrument met the stringent criteria required for an international study. Since then, there has been considerable activity in this arena, as documented in a recent review (Author 2008).

In contrast to accountability-related assessments, the low stakes associated with ILSAs (at least for students) make them a suitable vehicle for assessing noncognitive skills. As ILSAs transition to computer-based delivery, the potential for high-quality measurement of a broad array of such skills and dispositions is greatly enhanced. It should also be acknowledged that the distinction between "cognitive" and "noncognitive" is increasingly viewed as anachronistic: Many noncognitive skills have a strong cognitive component, and cognitive skills are applied most effectively when noncognitive skills are engaged. Thus, ILSAs should consider adopting a more expansive and holistic view of their focal constructs.

Role of Technology

Bill Gates is said to have remarked that "we overpredict the impact of technology in the short run and underpredict its impact in the long run." That rings true in the case of educational assessment in the United States, despite some undeniable advances in introducing computer delivery in a few states, as well as introducing it to such sectors as graduate admissions testing and professional licensure. With the continuing development of cheaper and more powerful mobile computing/communication devices and the completion of the next generation of the Internet and communication networks, one can reasonably hope that we are leaving the short run and entering the long run.

In the context of a particular ILSA, the strategic use of technology depends on a holistic view of the goals of the program and a realistic view of the constraints under which it operates: Would the introduction of computer delivery lead to improvements in construct representation and data utility that are sufficiently compelling to justify a major initial investment and, perhaps, larger operating costs? Could it lead to unintended biases? How would it affect participation of jurisdictions and of certain subpopulations in different jurisdictions?

Notwithstanding these and other related questions, there is a general sense that the introduction of technology in the administration of ILSAs is both inexorable and to be welcomed. Beller (this volume) shares that view. She offers a useful, comprehensive review of technology initiatives at the national and international levels. In particular, she briefly describes a number of interesting technology-based supplemental assessments undertaken or planned by both IEA and OECD.

There are a number of goals that can be envisioned for technology-based assessments. These include improving alignment and accuracy for measures of current target constructs, and facilitating the measurement of new constructs, such as problem-solving and computer/information literacy. These two, as well as other constructs that lend themselves more to technology-based assessments, could contribute to increased credibility and relevance of technology-based ILSAs, not least by strengthening links to the world outside schools. However, the assessment of new constructs will certainly raise challenging methodological issues. The introduction of more complex stimuli, as well as the desire to evaluate both processes and outcomes, will call for more sophisticated psychometric models and data-analytic strategies.

The conjunction of more ambitious targets of assessment and new means of delivery will also require different ways of organizing the work. The dynamics of the interactions among the various specialists are bound to become more complex as well. Technology will help here. As Beller points out, technology can increase efficiency and cost effectiveness by supporting new methodologies for collaborative assessment design and development, machine scoring of open-ended responses, and dissemination of results. Further advances are on the horizon. However, she does acknowledge the formidable challenges in conducting a computer-based administration internationally.

On this point, PIAAC, which is in the field in 2012, is a bellwether as it has been designed from the outset to be fully computer delivered.[1] Many lessons (some painful) have already been learned about conducting an ILSA on a new technology platform. If PIAAC can be carried out with reasonable success, it will surely provide an impetus for a broader move to computer-based ILSAs. Presumably, the infrastructure built for PIAAC can be leveraged for other OECD initiatives. The example of PIAAC demonstrates that, despite the challenges, many countries are eager to participate in a next-generation assessment.

Informing Policy

It is certainly true that volumes can be written concerning both the proper use of ILSA results and decrying the misuse of those same results. As mentioned earlier, ILSA results are most commonly viewed through the lens of league tables. Such tables are clear and irresistible, and appear to tell a very simple story. Too often, however, commentators focus on ranks (or changes in ranks) without due regard to the corresponding score differences. In many cases, substantially different ranks may mask small score differences (Bracey 2004). Although crossnational comparisons are of obvious interest, subnational comparisons may have greater immediate use. Unfortunately, these are too rarely given equal attention. An interesting hybrid is the simultaneous crossnational comparison of both levels of achievement and within jurisdiction variation (Sum et al. 2002) that directly addresses issues of equity.

ILSAs offer a well-designed framework to instantiate important constructs, and the outcomes do offer compelling examples of the high level of accomplishment that large proportions of students can reach in some jurisdictions. The contrasts among jurisdictions can be a powerful call to action, with a natural tendency to look to leading nations for policy prescriptions. Indeed, there is now burgeoning mini-industry based on culling "lessons learned" from the study of high-performance education systems. Delegations from lagging jurisdictions have been routinely dispatched to such destinations as Finland, Singapore, and Ontario to ferret out the secrets of their success. Commissioned reports drawing on the policies and practices of several leading nations purport to have distilled the keys to improved achievement, See for example, the reports by McKinsey (2007, 2010) and by Paine and Schleicher (2011).

Despite the enthusiasm of the authors and the certainty they communicate, caution is in order. Policy prescriptions implicitly rely on some form of causal attribution. As Hanushek and Woessman (this volume) acknowledge, there are serious impediments to making unassailable causal inferences from ILSAs. Although par-

[1] There is provision to administer a paper-and-pencil form when computer administration is infeasible or inadvisable.

ticipating jurisdictions form a natural experiment, high rankings or rapid progress are likely due to a confluence of factors, both educational and extraeducational. Focusing on certain common features of education policy offers only a partial and perhaps misleading picture. Hargreaves (2011) makes this point by noting that the Canadian provinces of Alberta, Ontario, and Quebec all do well in ILSAs but have rather different policies. He speculates that their advantage over the United States may be as much a function of economic, social and community conditions as the specifics of their educational systems.

From a methodological perspective, a necessary (but not sufficient) step would be to analyze the policies of a comparable group of "laggard" jurisdictions and determine that they indeed differ systematically from those of the leaders. Further, one would have to amass evidence to discredit alternative explanations for the differences in outcomes (Campbell 1957; Braun 2008). Another issue is whether differences in PISA outcomes truly reflect differences in performance of different jurisdictions or whether they are also due, in part, to the fact that the meaning of the background variables characterizing individuals and groups may vary across jurisdictions. Unfortunately, analyses that examine whether background characteristics of students in countries can be directly compared are rarely done. The patterns highlighted in the various reports, then, may be suggestive and even "common-sensical," but they are not scientifically impregnable. Caveat emptor is the watchword.

In a useful counternarrative, Klieme (this volume) offers a thoughtful analysis of the difficulties inherent in making inferences from ILSAs that are directly relevant to policymakers. He notes that the cross-sectional nature of ILSAs limit the strength of any causal claims and, in particular, points out the futility of carrying out credible value-added analyses. In a more positive vein, he suggests there is the possibility of a productive dynamic between ILSAs and what he terms "education effectiveness research." This is illustrated by an example of how Germany enhanced the value of an ILSA through the addition of a carefully designed and executed longitudinal component. More generally, there can be real value in secondary analyses of ILSA results, especially through focused subnational comparisons.

Because of the comprehensiveness of the data collected, going well beyond the cognitive results, ILSAs and related surveys offer a rich treasure trove for secondary analysts and have yielded important insights not available from single-country data. Hanushek and Woessman (this volume) cite an example from macroeconomic policy, but there are many others. As usual, special care is required in drawing conclusions from these data.

Policymakers and other stakeholders should not underestimate the obstacles to profiting from participation in an ILSA. Given the inertia inherent in education systems and the time lag in effecting meaningful reforms, successful change requires a sustained focus that, in turn, requires a broad political consensus on a long-term plan. Strategies should incorporate intermediate milestones whose attainment can maintain interest and support. Periodic assessments can be helpful in this regard, particularly if the results accurately reflect a trend of incremental improvement.

Policy Action

Ritzen (this volume) presents a general model of the different channels through which ILSA results provide information to various groups of stakeholders and, in the case of PISA 2006, categorizes different jurisdictions in terms of its impact on their education policies. He speculates on the various factors that determine whether ILSAs have a substantial impact in a particular jurisdiction. Where there is sufficient interest and concern, however, ILSAs can provide both impetus and direction, as illustrated by some of the examples cited by Ritzen. Klieme (this volume) makes the case that leveraging an ILSA through complementary studies can substantially enhance the utility of the findings and, thereby, play a greater role in policymakers' decisions.

Of course, a country's decision making in the educational arena depends on the interaction of multiple factors including the political context, national ambition, and competing priorities. However, the increasing prominence of ILSA results makes it more difficult for political leaders to dismiss them as irrelevant and disengage from the collaboration. At the same time, good intentions must be matched by sustained commitment and sufficient capacity. Poorer nations and those beset by political instability can experience difficulties in providing the former and building the latter. ILSA sponsors should rethink how they can provide the necessary support and encouragement to a broader array of within-country champions, recognizing that there are political considerations involved in engaging stakeholders through non-governmental channels.

Looking Ahead

On balance, in my view, ILSAs have had a positive impact on global educational systems. The critical question is whether and how that positive impact can be increased. There are at least three different paths:

- Provide more useful information.
- Enhance the value of that information.
- Extend the reach of the ILSAs.

Presumably, forward-looking strategies should encompass all three paths. With respect to the first two, conference presenters argued for extending the range and depth of the target constructs. Meaningful progress in this direction will likely involve some combination of computer-based delivery and special studies. As mentioned earlier, this will surely require the development and implementation of more powerful methodologies to assure sponsors of the accuracy and comparability of the results.

Although technical issues were raised only peripherally at the conference, they merit serious attention in any strategic planning exercise. Setting more ambitious

assessment goals may call for the introduction of adaptive testing algorithms, new psychometric models, or expert systems for evaluating complex student responses. Given the long lead times typically incorporated into ILSA schedules, there should be ample opportunity for sponsors and contractors to review the current status of these technologies, to project near-term developments, and to conduct pilot studies to obtain empirical results that can inform design choices. In any real-world setting, there will be conditions or demands that constrain what may be feasible from a technical point of view, necessitating various tradeoffs. Again, the experience with PIAAC can provide useful guidelines for future innovations.

An ILSA can also serve as the anchor for various time-linked complementary surveys conducted by individual countries or groups of countries sharing a common interest. One example, already cited, was provided by provided by Klieme (this volume). Earlier exemplars include the TIMSS teacher video study (Stigler et al. 1999), the OECD school leadership study (Pont et al. 2008), and the OECD teacher study (2009). Another direction is to link an ILSA to a national assessment, as has been done with TIMSS and NAEP (Phillips 2009; NCES 2011).

Such extensions greatly enhance the value of the core ILSA results. Further enhancements would accrue if ILSAs provided more interpretable descriptions of different levels of performance. This could be accomplished through a modified behavioral anchoring of selected points along the score scale or through segmenting the score scale and providing descriptions of the modal student in each segment. The former approach was pioneered with the National Assessment of Educational Progress (Beaton and Allen 1992) and is currently employed by TIMSS and PIRLS. The latter was developed for the NALS (Kirsch et al. 1993).

A viable alternative is suggested by the work of Torney-Purta and Amadeo (this volume), in which clusters of individuals with similar profiles are identified and described. The estimated population distributions across clusters in different jurisdictions provide useful comparative information.

With respect to the path of extending the reach, the most obvious strategy is to continue to add more jurisdictions to the roster of participants. However, this strategy has some potentially negative consequences. As the number of participating jurisdictions grows, it places an increasing burden on program staff, particularly if the additions involve new languages or nations with poor infrastructure. The question is whether staff could continue to achieve a broad consensus, preserve quality, and meet tight timelines. Failure to plan for the operational implications could lead ILSAs to become victims of their own success. To mitigate one aspect of the problem, the OECD specifies that only member countries participate in the design and item calibration. Other countries then pay for the opportunity to administer the assessment under supervised conditions, with the results reported on the common scale.

Fortunately, alternative strategies are available. Aspiring nations could ramp up to full participation by first using small, selected samples of students to sit for the assessment in order to gauge the appropriateness of the level for the full cohort. In some cases, it might be informative to have teachers take the assessment, although there likely would be political considerations involved in such a step. Intermediate levels of participation could also be organized through partnerships with regional

consortia. This could also create a channel for ILSA staff to share resources and expertise with the staff of the consortia. Indeed, building the requisite capacity in the developing world is a powerful, if indirect, way to extend the reach of the high-profile global ILSAs. The OECD is pursuing another direction, through the proposed development of a PISA-like instrument that could be administered by schools with the results reported on the PISA scales.

ILSAs also appear to be developing more sophisticated media strategies. With each passing year, "results-release events" are accorded more prominent coverage, and follow-on events build continuing interest in the outcomes and their implications. The problem is how to support the organization of such occasions in most participating jurisdictions, recognizing, as Ritzen (this volume) points out, that there will be political constraints in some settings. Equally important, there should be an ongoing effort to educate both policymakers and members of the media on the proper use and interpretation of ILSA results. This is not a trivial effort as many of the relevant issues involve technical issues that are not easily communicated to lay audiences.

Concluding Remarks

In this chapter I have suggested a framework for considering key issues that confront ILSA sponsors and contractors as they look to the future, and indicated how the topics presented relate to the framework. I have also taken the liberty of briefly addressing other topics. It should be borne in mind, however, that in addition to the speakers and discussants, this conference brought together nearly 100 individuals with interest, experience, and expertise in ILSAs. The comments following the presentations, as well as the conversations in the ample time between sessions, added immeasurably to the richness of the event.

My sense is that there was a general consensus that these global partnerships are a valuable resource for the international community and should continue to thrive. On the other hand, there is a contrarian perspective, not represented at the conference, which decries both the economic focus of human capital development and the growing influence of international assessments on national education policy (Spring 2011). Although these are minority views, they do remind us that equity should be given attention equal to that of efficiency as we consider different paths.

Not surprisingly, each speaker had a different focus and somewhat different recommendations on future directions. Over the next few years, these and other options will compete in the crucible of political, economic and technical realities. What will emerge? No one today can say. However, we should all bear in mind that ILSAs represent perhaps the only major international educational commitment for many countries, and sometimes their only source of nationwide information about the educational system. It is imperative that the sponsors and governing bodies strive to adapt the surveys to the evolving needs of an increasingly diverse set of countries while maintaining sufficiently strong links to the past to preserve trends.

Negotiating these sometimes conflicting desiderata will call on the best skills of both the measurement community and associated technical specialties—not to mention extraordinary political talents. Despite inevitable frustrations and setbacks, we should all keep our eyes on the prize of contributing to information-rich education policy decision making.

References

Author. 2008. *Psychosocial assessment of college readiness: A prospectus*. Princeton: New Constructs Center, Educational Testing Service.

Beaton, A.E., and N.L. Allen. 1992. Interpreting scales through scale anchoring. *Journal of Educational Statistics* 17:191–204.

Bracey, G.W. 2004. *Setting the record straight*. 2nd ed. Portsmouth: Heinemann.

Braun, H. 2008. McKinsey report: How the world's best performing school systems come out on top. *Journal of Educational Change* 9(3):317–320.

Braun, H., and A. Kanjee. 2006. Using assessment to improve education in developing nations. In *Educating All Children: A Global Agenda.* eds. J. E. Cohen, D. E. Bloom, and M. B. Malin. Cambridge, MA: American Academy of Arts and Sciences.

Campbell, D. T. 1957. Factors relevant to the validity of experiments in social settings. *Psychological Bulletin* 54:297–312.

Hargreaves, A. 2011, January 26. *Canada's Culture of Excellence in Education.* Toronto Star.

Kirsch, I., A. Jungeblut, L. Jenkins, and A. Kolstad. 1993. *Adult literacy in America: A first look at results of the National Adult Literacy Survey*. Washington, DC: National Center for Education Statistics, U.S. Department of Education.

Kirsch, I., H.I. Braun, K. Yamamoto, and A. Sum. 2007. *America's perfect storm: Three forces changing our nation's future*. Princeton: Policy Information Center, Educational Testing Service.

McKinsey & Co. 2007. *How the world's best school systems come out on top*. New York: McKinsey & Co.

McKinsey & Co. 2010. *How the world's most improved school systems keep getting better*. New York: McKinsey & Co.

National Center for Education Statistics. 2011. http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2011472

OECD. 2003. *Learners for life: Student approaches to learning, results from PISA 2000*. Paris: Organisation for Economic Co-operation and Development.

OECD. 2009. *Creating effective teaching and learning environments: First results from TALIS*. Paris: Organisation for Economic Co-operation and Development.

Paine, S.L., and A. Schleicher. 2011. *What the U.S. can learn from the world's most successful education reform efforts*. New York: McGraw-Hill Research Foundation.

Phillips, G.W. 2009. *The second derivative: International benchmarks in mathematics for U.S. states and school districts.* Washington, DC: American Institutes for Research.

Pont, B., D. Nusche, and H. Moorman. 2008. *Improving school practice. vol.1: Policy and Practice.* Paris: Organisation for Economic Co-operation and Development.

Spring, J. 2011. *The politics of American education*. New York: Routledge.

Stigler, J. W., P. Gonzales, T. Kawanaka, S. Knoll, and A. Serrano. 1999. *The TIMSS videotape classroom study: Methods and findings from an exploratory research project on eighth grade mathematics instruction in Germany, Japan, and the United States*. Washington, DC: National Center for Education Statistics.

Sum, A., I. Kirsch, and R. Taggart. 2002. *The twin challenges of mediocrity and inequality: Literacy in the U.S. from an international perspective.* Princeton: Policy Information Center, Educational Testing Service.