CICE Current Issues in Comparative Education

Volume 14(1) / Fall 2011 • ISSN 1523-1615 • http://www.tc.edu/cice

Evaluation, Assessment, and Testing

FEATURED ARTICLE

BOOK REVIEW

CURRENT ISSUES IN COMPARATIVE EDUCATION
Volume 14, Issue 1 (Fall 2011)

Issue Editors:
Ruaridh J. MacLeod, Sarah Flatto

CICE Current Issues in Comparative Education

Volume 14(1) / Fall 2011 • ISSN 1523-1615 • http://www.tc.edu/cice

Evaluation, Assessment, and Testing

FEATURED ARTICLE

BOOK REVIEW

CURRENT ISSUES IN COMPARATIVE EDUCATION
Volume 14, Issue 1 (Fall 2011)

Issue Editors:
Ruaridh J. MacLeod, Sarah Flatto

CICE Current Issues in Comparative Education

Volume 14(1) / Fall 2011 • ISSN 1523-1615 • http://www.tc.edu/cice

Evaluation, Assessment, and Testing

FEATURED ARTICLE

BOOK REVIEW

CURRENT ISSUES IN COMPARATIVE EDUCATION
Volume 14, Issue 1 (Fall 2011)

Issue Editors:
Ruaridh J. MacLeod, Sarah Flatto

How Much is Learning Measurement Worth? Assessment Costs in Low-Income Countries

Daniel A. Wagner[1]
International Literacy Institute, University of Pennsylvania

Andrew Babson
International Literacy Institute, University of Pennsylvania

Katie M. Murphy
International Literacy Institute, University of Pennsylvania

Timely and credible data on student learning has become a global issue in the ongoing effort to improve educational outcomes. With the potential to serve as a powerful diagnostic tool to gauge the overall health and well-being of an educational system, educational assessments have received increasing attention among specialists and the media. Though the stakes are high, relatively little is known about the cost-benefit ratio of various assessments compared to other educational expenditures. This paper presents an overview of four major types of assessments—national, regional, international and hybrid—and the costs that each has incurred within 13 distinct contexts, especially in low-income countries. The findings highlight broad variation in the total cost of assessment and the cost-per-learner. This underscores the importance of implementation strategies that appropriately consider scale, timeliness, and cost-efficiency as critical considerations for any assessment.

Throughout the global arena, government agencies, international organizations, donors and private sector partners increasingly emphasize the need for timely and credible data on student learning that may inform the design of effective mechanisms to improve educational outcomes. Considerable attention at these multiple levels, compounded by a heightened publicity in the media, has prompted a dramatic and global growth in the use of learning assessments (Kamens & McNeely, 2010).

At the same time, the reality of restricted educational budgets demands affordable and cost-effective options for assessments. Indeed, with the growth of Large Scale Educational Assessments (LSEAs) throughout the world, there has been a concomitant increase in attention to the fiscal burden of assessments in low-income countries (LICs). These costs have often been borne by external funders such as bilateral or donor agencies, resulting in the common perception that this burden of investments in knowledge is rather minimal when compared to the large amounts spent on education itself (Lockheed & Hanushek, 1988; Porter & Gamoran, 2002).

The perception that LSEAs are relatively low-cost has been supported by a limited number of studies showing that assessment costs represent a very small proportion of national education budgets.[2] Yet these studies do not appear to account for the increasingly limited amount of discretionary funds for such activities that may be available to ministers of education in low-income countries, with or without external support. Hence, for more than a decade, other critical perspectives have emerged that challenge the assertion that LSEAs are a relatively small-scale investment.[3]

The actual costs of LSEAs and other assessments are needed in order to determine cost-benefit analyses within low-income countries. A successful international assessment requires high-level skills in design, planning and management—skills that are in short supply globally—especially in LICs.[4] Ministries of education throughout the world are now confronting difficult decisions in regard to assessments. First, they must decide whether to participate in LSEAs, understanding the costs and complexity of large-scale assessments. Second, they must determine how to choose tests that are appropriate for students, languages and educational systems[5] given the wide variety of assessments available today.

Ensuring that policymakers and key stakeholders have accurate information on the actual costs of assessments is a critical step in identifying appropriate tools to inform and influence initiatives aimed at improving educational outcomes. This paper presents an overview of four major types of assessments: national, regional, international and hybrid, and the costs that each of these types of assessments have incurred during implementation in various contexts. The findings highlight a broad variation in the total cost of assessment and the total cost-per-learner within each of the four types of assessments.

Types of Educational Assessments

Large-scale educational assessments (LSEAs) have increasingly been used by national and international agencies beginning in the 1980s. Previously, only a small number of cross-national large-scale assessments had been conducted, mostly by the IEA (International Association for the Evaluation of Educational Achievement).[6] Technological and methodological advances in assessment, combined with the political pressure to improve educational systems, have spurred this trend, including in LICs (Kelleghan & Greaney, 2001). The 1990 Jomtien Conference demanded more accountability and systemic evaluation of education in LICs. Further, in 2000, the UNESCO Dakar Framework for Action called for the achievement of "measurable" learning outcomes, and that such progress should be "monitored systematically" (UNESCO, 2000, p. 21). LSEAs have increasingly become a key tool for meeting these demands.

Despite this momentum, the increasing complexity and expense of LSEAs have led to questions about the utility of conducting LSEAs in low-income countries. Although a number of agencies have carried out LSEAs in the OECD (Organization for Economic Co-operation and Development) countries, it was not until the 1990s that the capacity to participate in LSEAs (international and regional) became more available to LICs. The complexity of stakeholder

interests, as well as resource constraints, has limited growth of LSEAs in LICs. However, various donor agencies, such as the World Bank, have become increasingly important funders of LSEAs, making it more affordable and more likely for such assessments to be utilized even when national budgets are constrained.[7]

With a focus on learning assessments in low-income countries, the present discussion centers on four main types of assessments: national, regional, international and hybrid. Each of these is described below.

National assessments

National assessments (sometimes called national or public examinations) evaluate all students in a national educational system. Nearly all countries engage in some type of national assessment in order to ascertain whether desired and planned educational goals are achieved. The results can be used to modify curricula, train teachers, reorganize school access, or refashion numerous other aspects of a national educational system. The results also can be used for accountability purposes, to make resource allocation decisions, and to heighten public awareness of education issues.

Regional assessments

Regional assessments provide an opportunity to measure student learning across a group of countries, typically defined by a geographic region or by a shared national language. They have grown in popularity over the last 20 years, and as part of an effort to extend the use of LSEAs into developing countries, regional and international organizations have collaborated to create three major regional assessments: the *Latin American Laboratory for Assessment of Quality in Education* (LLECE), the *Southern and Eastern African Consortium for the Monitoring of Education Quality* (SACMEQ), and *Program for the Analysis of Educational Systems of the CONFEMEN* (francophone Africa) countries (PASEC).

International assessments

International assessments are designed to measure learning in multiple countries. Their aims include: (a) cross-national comparisons that target a variety of educational policy issues; (b) provision of 'league tables' that rank-order achievement scores by nation or region or other variables; and (c) within-country analyses that are then compared to how other countries operate at a sub-national level. These studies are undertaken by various international organizations and agencies, including: the IEA that conducts the *Progress in International Reading Literacy Study* (PIRLS), and the OECD that is responsible for the *Program for International Student Achievement* (PISA) studies. Each of these international assessments is now in use in dozens of countries, and is expanding well beyond the OECD country user base that formed the early core group of participation.

Hybrid assessments

In recent years, a new approach to assessment has sought to focus more directly on the needs of LIC assessment contexts. Initially, this approach was conceptualized under the acronym

"smaller, quicker, cheaper" (SQC) methods of literacy assessment (ILI/UNESCO, 1998; Wagner, 2003). The idea was to see whether LSEA methodologies could be reshaped into *hybrid*[8] methods that are just big enough, faster at capturing, analyzing and disseminating data, and cheaper in terms of personnel and cost outlays (Wagner, 2010, 2011). The *Early Grade Reading Assessment*, or EGRA, (Research Triangle Institute, 2009) contains a number of the above features, and is probably the best-known current example of a hybrid assessment in reading. The EGRA was initially designed with three main assessment goals: early reading (grades 1-3), local contexts (rather than comparability across contexts), and local linguistic and orthographic variation.

Cost-benefit Analyses in Educational Assessment

In the early 1990s, a limited number of studies examined the costs and benefits of LSEAs (Ilon, 1992, 1996; Koeffler, 1991; Loxley, 1992; Lockheed, 2008). The results supported the value of LSEAs for two main reasons: the fairly low overt costs (i.e., costs that are explicitly budgeted and accounted for) in relation to the overall education budget (Peyser & Costrell, 2004; Hoxby, 2002), and the high potential benefits of LSEAs to yield actionable results (Braun & Kanjee, 2006; Hanushek & Woesmann, 2005). Nonetheless, as pointed out by Lockheed (2008, p. 9), "national learning assessments in developing or transition countries rarely employ complex measurement instruments because such countries rarely have the requisite domestic capacity or can afford to purchase expertise from abroad." This point is echoed by Greaney and Kelleghan, (2008) and Wolff (2007), and further linked to potential wastage or failure down the road if sufficient investments are not made up front (Wolff, 2008).

Thus, while the assessment field itself—whether in high- or low-income countries—seems largely convinced of the importance of LSEAs, the total costs of assessments are becoming more clearly recognized as a serious obstacle for LICs. For example, Braun and Kanjee (2006) assert that, in countries with developing educational systems, "scarce resources are better devoted to assessments directed at improving learning and teaching, where the returns on investments are likely to be higher."[9] Research shows that the average costs of an LSEA appear small relative to national educational budgets (less than 1% generally per national budget, and as low as 0.3%). However, such low percentages may not reflect the percentage of the available discretionary budget (Coombs & Hallak, 1987, p. 50).

Calculating the costs

In order to make a cost-based decision about assessment choice, it is important to bear in mind both overt and hidden costs that come into play in any assessment (cf. Greaney & Kellaghan, 2008). Overt costs are those that are typically planned for in advance and that are included in the accounting mechanisms of the agency (or agencies) in charge of the LSEA. These would include staff costs of test management (such as test design and application) and training, as well as travel, supplies and equipment.

They can also vary by location, including: within-country costs (e.g., roll out and management of the assessment process within country); in-kind costs (e.g., non-cash contributions such as ministry staff, specialists, headmasters, and teachers); and international costs (e.g., international agency overheads, international experts, and travel).

Hidden costs are those that may escape the attention of authorities that put together fiscal plans for assessments. They include the following items.

- *Indirect (or overhead) costs*. These costs are absorbed by the agencies themselves in implementing the program. While often accounted for in wealthier countries, these costs sometimes escape the attention of ministries and other agencies in LICs. Obvious examples would include the cost of using infrastructure (e.g., buildings, networks, computer maintenance, and so forth). Less obvious, but significant, costs may be associated with seconded staff in the ministry and field workers who may be school inspectors or teachers.

- *Opportunity costs*. These costs are relative to what different strategy may have taken place in lieu of the particular choice that is made. For example, by not doing an assessment in a particular year, the ministry might have more resources to do the assessment in a subsequent year. Or, choice of one type of assessment may preclude opting for an additional or different choice.[10] However, the cost of *not* participating in an assessment—that is, foregoing the potential benefits (in terms of staff development, potential results, etc.) of participation in an assessment—must also be considered as another type of opportunity cost.

Cost Categories and Comparisons: Selected Assessments

The cost categories in assessments from the previous discussion may be seen in summary form in Appendix A. For purposes of comparison, a number of well-known assessment agencies were contacted for current information on expenditures (some in estimated form). The studies covered are listed in Appendix B. Data collected from each of the selected studies at a national level are represented in Table 1, which indicates the variability of known assessment costs, by assessment and national context across 13 recent assessments. Table 2 provides a summary of average percentages of total expenditures across the six main cost categories.[11]

As shown in Table 1, it is possible to make a number of observations. First, the student populations ranged from a modest 3,770 in EGRA-Liberia, to about 300,000 in SIMCE (Chile).[12] Second, it may be seen that the total (listed) overt costs of undertaking the assessment range from a low of about $122,000 in PISA (Uruguay) to a high of $2.8 million in SIMCE (Chile). Third, by considering these first two parameters, it is possible to calculate the 'cost-per-learner' (CPL) assessed, a useful way of looking at costs irrespective of size of the total enterprise. Results indicate that cost-per-learner ranges from about $8 in the Uruguay national assessment to about $51 in the SACMEQ III study in Swaziland to about $171 in PISA in Chile. The average for this sample of studies is about $42 per learner assessed. In addition (see Table 2), certain costs figured more prominently than others, such as test application (50%) and

institutional costs (23%), while processing and analysis (13%) and test preparation (11%) were substantially lower.[13]

The average CPL data show that, at the field level, these are not dramatically different when compared across types of tests. Some assessments are clearly more expensive, but it is interesting to note that the larger national and international studies confer economies of scale that reduce per-unit assessment costs. At present, the smaller EGRA studies are not less expensive at the field level. Further, some countries may have significantly more resources (financial, intellectual, infrastructural, etc.) in their evaluation departments upon which to draw. This will likely affect a number of cost variables, such as in-house versus external consulting fees and travel expenses. It must be understood that hybrid assessments are still in a research phase (with inherent costs of trial and error), such that their costs may be expected to drop substantially with the establishment of economies of scale. In addition, specific in-country needs and requirements (e.g., logistics in difficult terrain) may also play a major role in determining which types of assessment are chosen, and thus how much is ultimately spent on assessment.

Of course, much depends on whether cost estimates are correct and whether hidden costs are fully included. Not all teams collect and store cost data and, even if they do so, these data may not be complete or sufficiently detailed for comparative analyses. Inaccuracies and discrepancies are often the result of underfunding (Lockheed, 2008, p. 16). Thus, these data should be considered a preliminary view of cost comparisons, and more needs to be done with full and reliable auditing in place.

Cost parameters with low-income countries in mind

In low-income countries, educational decision makers will find themselves with more choices than available resources. The cost-benefit picture remains insufficient. Simply not enough reliable data have been collected on assessment costs for the variety of assessments currently in use. Moreover, the current scientific, technological and political dynamism in educational improvement strongly suggests that models of assessment will change in relation to testing advancements and increasing demand. The necessity for both clear testing choices and actionable indicators is likely to increase.

Recent assessment innovations (e.g., EGRA) suggest momentum toward models of assessment that both emphasize a needs-centered and 'just enough' approach to testing (Wagner, 2003). This means that innovations may help to grow the scale of test application, shrink upfront overt costs such as translation and test preparation, and reduce turnaround time. This way, government bodies can possess actionable data sooner and thus with less staff and overhead. Three key parameters summarize the cost issues of assessments that will need to be considered, especially in the context of resource-constrained LICs.

Scale

Ministries of education in LICs will need to consider which assessments would yield targeted and responsive educational data about a specific population (e.g., rural girls, ethno-linguistic groups), a group of schools, or concerning a particular subject at a particular grade level. LSEAs typically cannot respond flexibly to such requests due to the significant up-front preparation and pre-assessment exercises that constrain near-term changes, and lock in comparability parameters. Further, most LSEAs are not designed to provide classroom-level indicators but rather systemic indicators (Volante, 2006). By contrast, limited sample household-based surveys or EGRA style hybrid assessments can save money because they can reduce the number of individuals to be assessed in order to answer a more specific set of policy questions, and can be deployed and adjusted more frequently. Still, recent sampling innovations in LSEAs (such as PIRLS) suggest that such studies not only provide multi-level data, but also that the economies of scale can enable larger samples at marginal additional cost.[14] In other words, lower cost (in CPL) is a relative term.

Timeliness

Two types of timeliness are crucial to the possible benefits of assessments. First, there is the timeliness of the testing cycle from planning, rollout, and data collection to analysis and dissemination (and subsequent policy debates). Second, timeliness can also refer to the 'right time' of information availability and use. For example, if timely information about a group of schools is ready in advance of major school finance decisions, then those data can show real-time sensitivity. Or, a population of students may need assistance to reach grade-level competence in reading, and data may confirm, disconfirm, and/or guide the decision-making process. In addition, there is a need to consider the merits of early intervention in the learning trajectory of students, much as the arguments have been made in the medical field for early detection systems.[15] In sum, credible assessment data needs to be gathered as quickly as possible in order to effectively shape policymaking, yet it also needs to be available for application to decision-making at the right time. If 'time is money' (as the adage goes), then moving toward timeliness can also help to reduce overall costs of assessment and intervention.

Cost efficiency

As mentioned above, some assessments are relatively expensive in terms of up-front cost outlays, with requirements of expensive professional staff and consultants, and trained field enumerators. These and other costs can be seen in terms of either total costs or the CPL. Either way, budgetary limits on discretionary funds in LICs will require careful scrutiny as assessment choices are made. Given the paucity of credible data on costs in LICs today, it is difficult to derive an evidence-based decision pathway for multiple contexts. There is a clear need to more precisely determine which expenditures are likely to reveal particular policy outcomes. For example, will increasing expenditures for the training of enumerators yield better inter-rater reliability? Or, as in a recent effort in India, can volunteers become low-cost, reliable and sustainable enumerators with relatively little training at all (Banerji, 2006)? More research is needed to better clarify the cost merits of different assessments.

Conclusions

Costs are an inherent part of any social intervention. The assessment of learning and its policy consequences constitute a clear case in point. The key issue here is not that assessments are "expensive" or not. Rather, the issue is what a ministry (or funding agency) will receive in return for its investments.

Gathering data on the *comparative* costs of assessments is dificult. There are, however, some reference points now available that can be considered. Perhaps most important is the trade-off between time and money. Take, for example, a minister of education who may have up to five years to decide upon and implement policy. In this case, regional or international LSEAs such as SACMEQ or PASEC may provide some solid answers on key issues, and offer a sense of cross-national comparison. Given the current economies of scale in countries that repeat international assessments, the actual CPL of such LSEAs is not much different from that of the EGRA and hybrid assessments that have much smaller sample sizes.

On the other hand, if a minister has a shorter window of policymaking opportunity (such as the typical two to three-year mandate in office), and if the priority is helping programs, schools and regional districts attain their near-term learning achievement goals, even a small-scale sample-based assessment like EGRA looks much less expensive. While the CPL in EGRA appears similar to the larger international assessments at present, the future costs will likely drop as EGRA tools become more familiar, enumerator training improves,and technological advancements reduce the amount of time and human resources required to analyze and disseminate assessment data.

Finally, there are opportunity costs to consider. LSEAs are typically not administered until children reach grade 4 (or later), when children may be far behind in reading development; this can impose very high costs in remediation that early assessment could prevent. "Catching up" is expensive, difficult, and may lead to school failure—the most important cost that policy makers seek to avoid.

In sum, evaluating and learning from assessments is fundamental to credible change in educational systems across nations. But learning assessments entail costs that need to be evaluated and compared. Gone are the days when ministerial agencies can assign free seconded staff to the field, or when outside donor agencies will fully fund large scale assessments. We are in a time of fiscal constraints. Learning about education has to be balanced against *what is learned*, *for what purposes, and at what cost*. The evaluation of assessment costs is an issue that will need considerably greater attention in the field of international education.

Table 1

Test monetary costs (USD)	National Assessments			Regional Assessments			International Assessments					EGRA Assessments[i]	
	SIMCE 2004[a]	Honduras 2004[b]	Uruguay 2003[c]	PASEC 2010[d]	SACMEQ III Swaziland 2007[e]	SACMEQ III Tanzania 2007[f]	PISA Chile 2009[g]	PISA Mexico 2009[h]	PISA Panama 2009[i]	PISA Peru 2009[j]	PISA Uruguay 2003[k]	EGRA - Liberia 2008	EGRA - Nicaragua 2008
Test preparation	**258,236**	**174,275**	**21,528**	**34,164**	**12,561**	**12,666**	**26,448**	**100,301**	**61,475**	**47,956**	**12,357**	**29,345**	**10,882**
Creation and editing of test items	184,515			7,895		1,000	26,448	3,802	13,661				
Pilot testing	73,721			15,749	12,561	11,666		96,499	47,814			16,031	4,756
Training				10,520								13,314	6,126
Test application	**1,163,764**	**435,717**	**57,289**	**91,705**	**170,732**	**89,900**	**597,958**	**891,501**	**187,157**	**212,486**	**29,707**	**82,260**	**68,683**
Test design and editing	29,403			7,415		2,000	8,976		13,661	2,590		8,800	
Test printing	324,712			9,744	15,488	12,000		254,899	54,644	7,196		5,600	1,395
Printing of other materials	236,076				3,049	4,200		116,156	6,831				
Distribution to examiners	103,124			6,455	73,171	2,000		123,845	6,831				
Field testing	406,103			68,091	79,024	56,700	462,705	394,235	98,359	198,261		67,860	67,288
Control and supervision	64,346					13,000	126,277	2,366	6,831	4,439			
Processing and analysis	**382,239**	**130,721**	**26,272**	**12,624**	**454**	**33,300**		**167,782**	**128,414**		**22,838**	**13,533**	**5,734**
Coding and digital input	216,048			12,624		33,300		56,899	114,753			13,533	5,734
Marking open-ended questions	166,191				454			110,883	13,661				
Additional analyses													
Dissemination	**100,567**	**130,721**	**531**	**32,193**	**4,195**	**2,000**	**49,912**		**34,153**	**3,865**	**14,092**	**1,850**	
School communication	100,567				4,195	2,000			34,153	3,865		1,500	
Report production and distribution							49,912					350	
Public relations retainer													
Subtotal	**1,904,806**	**871,434**	**105,620**	**170,686**	**187,942**	**137,866**	**674,318**	**1,159,584**	**411,199**	**264,307**	**78,994**	**126,988**	**85,299**
Institutional costs	**938,766**			**12,481**	**24,878**	**25,500**	**179,233**	**490,203**	**94,261**	**20,473**		**103,520**	**87,157**
Personnel- in project budget	796,864			2,737	17,561	10,000	179,233	321,246	73,769	9,324		101,858	83,675
Personnel- contributed								107,286		11,149		1,403	2,500
Infrastructure- in project budget	35,369					5,000		2,743	6,831				
Infrastructure- contributed													
Equipment in - project budget	106,533			9,744	7,317	10,500		58,928	13,661			259	982
Equipment- contributed													
Test Fees							**49,863**	**118,599**			**43,197**		
Other	**20,028**			**2,043**			**72,494**		**13,661**	**2,000**		**10,619**	**6,958**
TOTAL	**2,863,600**	**871,434**	**105,620**	**185,210**	**212,820**	**163,366**	**975,908**	**1,768,386**	**519,121**	**286,780**	**122,191**	**241,127**	**179,414**
Total Students	300,000	45,657	12,993	5,400	4,155	3000 [m]	5700 [n]	45,079	42,000	7,967	5,797	3,770	5,760
Total Schools												240	120
Cost per student	**10**	**19**	**8**	**34**	**51**	**55**	**171**	**39**	**12**	**36**	**21**	**64**	**31**
Cost of educating a student	767	130	484		66			9,439	1,023	396	479		
Cost of testing as % of total budget for one grade	0.83	2.63							1.20838				
Cost of testing as % of total secondary education budget	0.17	0.33	0.07					0.001767	0.04419		0.08		

Table One: Detailed costs for national, regional, international and EGRA assessments.

[a] Source: Wolff, 2007, p. 6 (for 2004 SIMCE test). Wolff (2007) used local currencies for his figures on PISA Uruguay 2003 and all the national assessments above (namely SIMCE 2004, Honduras 2004 and Uruguay 2003). In order to facilitate comparisons across assessments in this table, we converted Wolff's figures to the average annual market rate for USD. Further, in his analysis of SIMCE 2004, Wolff used SIMCE 2002 figures, in Chilean Pesos (At the rate of 677.4916667 Chilean Peso to 1 USD).

[b] Source: Wolff, 2007, p. 13; 2004 17.68 Honduran Lempira to 1 USD

[c] Source: Wolff, 2007, p. 11; 2003 28.24279 Uruguayan Peso to 1 USD

[d] Source: PASEC 2010 technical report (personal communication, P. Varly, May 2009). Converted from Euros to USD, 2009 annual rate.

[e] Source: Personal communication, A. Mrutu, August 2009.

[f] Source: Personal communication, J. Shabalala, August 2009.

[g] Source: Personal communication, E. Lagos, September and October 2009.

[h] Source: Personal communication, M. A. Diaz, September 2009.

[i] Source: Personal communication, Z. Castillo, September 2009.
[j] Source: Personal communication, L. Molina, September 2009.
[k] Source: Wolff, 2007, p. 14; 28.24279 Uruguayan Peso to 1 USD (2003)
[l] Source: Personal communication, A. Gove, August 2009.
[m] Estimate, based on SACMEQ II sample of 2854
[n] Estimate, based on email of E. Lagos, October 2009

Table 2.

Cost category	Average	Lowest	Highest
Test preparation	11%	3% (PISA Chile, 2009)	20% (Uruguay, national assessment, 2003)
Test application	50%	24% (PISA Uruguay, 2003)	80% (SACMEQ III, Swaziland)
Processing and analysis	13%	1% (SACMEQ III, Swaziland)	25% (Uruguay, national assessment, 2003)
Dissemination	6%	1% (Uruguay national assessment, 2003)	17% (PASEC, 2010)
Institutional costs	23%	7% (PASEC 2010)	49% (Uruguay, national assessment, 2003)
Test fees	16%	5% (PISA Chile, 2009)	35% (PISA Uruguay, 2003)
Other	3%	1% (PISA Peru, 2009)	7% (PISA Chile, 2009)

Note. Above calculations based on data from 13 assessments (see Table 1 for costs included in each category and for each assessment).

Table two: Costs by category, as percentage of total assessment expenditures

Appendix A

Cost categories of the assessments used in selected studies. (Adapted from Wolff (2007)).

1. Test preparation
   a. Creation and editing of test items
   b. Pilot testing
   c. Training

2. Test application
   a. Test design and editing
   b. Test printing
   c. Printing of other materials
   d. Distribution to examiners
   e. Field testing
   f. Control and supervision

3. Processing and analysis
   a. Coding and digital input
   b. Marking open-ended questions
   c. Additional analysis

4. Dissemination
   a. Report to each school
   b. Report production and distribution
   c. Public relations retainer

5. Institutional costs
   a. Personnel- in project budget
   b. Personnel- contributed *(e.g., consultants)*
   c. Infrastructure- in project budget *(physical space for personnel)*
   d. Infrastructure- contributed
   e. Equipment- in project budget *(e.g., computers and related testing equipment)*
   f. Equipment- contributed
   g. Other (*e.g., telecommunications, electricity and office supplies)*
   h. Test fees

6. Cost breakdown

a. Cost of testing per student
b. Cost of educating a student (at test-specific grade level)
c. Cost of testing as % of total budget for one grade
d. Cost of testing as % of total secondary education budget

Appendix B

Cost studies of selected national, regional and cross-national assessments

- National assessments:
  - SIMCE/LLECE 2004
  - Uruguay national assessment 2002
  - Honduras national assessment 2002

- Regional assessments:
  - SACMEQ II
    - Swaziland 2006
    - Tanzania 2006
    - Zambia 2006
  - PASEC 2010

- International assessments:
  - PISA
    - PISA Chile 2009
    - PISA Mexico 2009
    - PISA Panama 2009
    - PISA Peru 2000
    - PISA Peru 2009
    - PISA Uruguay 2003
  - PIRLS

- Hybrid assessments:
  - EGRA
    - Liberia 2008
    - Nicaragua 2008

References

Banerji, R. & Wadhwa, W. (2006). Between outlays and outcomes. *Times of India*. Accessed on February 28, 2011 at http://images2.asercentre.org/Resources/Articles/Between__outlays_-_outcomes.pdf.

Braun, H. & Kanjee, A. (2006). Using assessment to improve education in developing nations. In Cohen, J. E., Bloom, D. E., & Malin, M. (Eds). *Improving education through assessment, innovation, and evaluation* (pp. 1-46). Cambridge, MA: American Academy of Arts and Sciences.

Coombs, P. H. & Hallak, J. (1972). *Managing educational costs.* New York: Oxford University Press.

Gilmore, A. (2005). *The impact of PIRLS (2001) and TIMSS (2003) in low- and middle-income countries: An evaluation of the value of World Bank support for international surveys of reading literacy (PIRLS) and mathematics and science (TIMSS).* New Zealand: IEA.

Greaney, V. & Kellaghan, T. (2008). *Assessing national achievement levels in education*. (National assessment of educational achievement, Vol. 1). World Bank: Washington.

Grin, F. (2005). The economics of language policy implementation: Identifying and measuring costs. In N. Alexander (Ed.), *Mother tongue-based bilingual education in Southern Africa: The dynamics of implementation.* Proceedings of a Symposium held at the University of Cape Town, 16-19 October 2003, Cape Town: Volkswagen Foundation & PRAESA.

Hanushek, E. & Woessmann, L. (2005). Does educational tracking affect performance and inequality? Differences-in-differences evidence across countries (NBER Working Paper 11124). Cambridge, MA: National Bureau of Economic Research.

Heugh, K. (2006). Cost implications of the provision of mother tongue and strong bilingual models of education in Africa. In H. Alidou, A. Boly, B. Brock-Utne, Y. Diallo, K. Heugh & H. Wolff (Eds.), *Optimizing learning and education in africa –the language factor: A stock- taking research on mother tongue and bilingual education in Sub-Saharan Africa* (pp. 56-84). Paris: IIEP/ADEA.

Hoxby, C. (2002). The cost of accountability. (NBER Working Paper 8855). Cambridge, MA: National Bureau of Economic Research.

ILI/UNESCO. (1998). *Literacy assessment for out-of-school youth and adults*. (ILI/UNESCO Technical Report from Expert Seminar, Paris, June 1998). Philadelphia: International Literacy Institute, University of Pennsylvania.

Ilon, L. (1992). A framework for costing tests in Third World countries (PHREE/92/65). Washington, DC: World Bank.

Ilon, L. (1996). Considerations for costing national assessments. In P. Murphy et al. (Eds.) *National assessments: Testing the system* (pp. 69-88). Washington, DC: World Bank.
Johnson, S. (1999). International Association for the Evaluation of Educational Achievement science assessment in developing countries. *Assessment in Education, 6* (1), 57–73.

Kamens, D. H. & McNeely, C. L. (2010). Globalization and the growth of international educational testing and national assessment. *Comparative Education Review*, *54*(1), 5-25.

Koeffler, S. (1991). *Assessment design*. Paper presented at the Seminar on Measurement/Assessment Issues. Educational Testing Services. Princeton, NJ.

Lockheed, M. (2004). The purpose of good quality education. Paper commissioned for the *Education For All Global Monitoring Report 2005, The quality imperative.* Paris: UNESCO.
Lockheed, M. & Hanushek, E. (1988). Improving educational efficiency in developing countries: What do we know? *Compare, 18*(1), 21-38.

Loxley, W. (1992). Managing international survey research. *Prospects*, *22*(3), 289-296.
Ndaruhutse, S. (2008). *Grade repetition in primary schools in Sub-Saharan Africa: An evidence base for change*. London: CfBT Education Trust.

Peyser, J. & Costrell, R. (2004). Exploring the costs of accountability. *Education Next, 4*, pp. 22-29.

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 Academies Press.

Postlethwaite, T. N. (2004). What do international assessment studies tell us about the quality of school systems? Background paper prepared for the *Education For All Global Monitoring Report 2005, The quality imperative.* Paris: UNESCO.

Ravela, P., Arregui, P., Valverde, G., Wolfe, R., Ferrer, G., Martínez, F., Aylwin, M. & Wolff, L. (2008). The educational assessment Latin America needs (Working Paper Series No. 40). Washington, DC: PREAL.

Research Triangle Institute (RTI). (2009). *Early Grade Reading Assessment Toolkit.* Washington, DC: RTI International.

Siniscalco, M. T. (2006). What are the national costs for a cross-national study? In K. Ross & I. J. Genevois (Eds.), *Cross-national studies of the quality of education: Planning their design and managing their impact* (pp. 185-209). Paris: UNESCO IIEP.

Szucs, T., Belisari, A. & Mantovani, L. (1997). Is preventive medical care worth the cost? *Biologicals, 25*, 247-252.

Topol, B., Olson, J., & Roeber, E. (2010). *The Cost of New Higher Quality Assessments: A Comprehensive Analysis of the Potential Costs for Future State Assessments*. Stanford, CA: Stanford University, Stanford Center for Opportunity Policy in Education.

UNESCO. (2000). *Dakar framework for action*. Education For All: Meeting our collective commitments. Dakar/Paris: UNESCO.

Volante, L. (2006). An alternative for large-scale assessment in Canada. *Journal of Learning and Teaching, 4*(1), 1-14.

Wagner, D. A. (2003). Smaller, quicker, cheaper: Alternative strategies for literacy assessment in the UN Literacy Decade. *International Journal of Educational Research,* 39, 3, 293-309.

Wagner, D. A. (2010). Quality of education, comparability, and assessment choice in developing countries. *COMPARE: A Journal of Comparative and International Educatio*n, 40, 6, 741-760.

Wagner, D.A. (2011). *Smaller, quicker, cheaper: Improving learning indicators for developing countries*. (In press). Washington/Paris: Fast Track Initiative & UNESCO-IIEP.

Wolff, L. (1998). *Educational assessments in Latin America: Current progress and future challenges*. Working Paper. Washington, DC: PREAL.

Wolff. L. (2007). The costs of student assessments in Latin America (Working Paper Series No. 38). Washington, DC: PREAL.

Wolff. L. (2008). *Costs and financial implications for implementing national student assessments: Lessons for South Africa.* Unpublished Presentation, November 1.

[1] Authors' addresses: Daniel A. Wagner, International Literacy Institute, University of Pennsylvania, wagner@literacy.upenn.edu; Andrew Babson, University of Michigan, ababson@umich.edu; and Katie Maeve Murphy, University of Pennsylvania, kmaeve@gmail.com. Mailing address: Dan Wagner, Professor of Education, Graduate School of Education, University of Pennsylvania, 3700 Walnut Street, Philadelphia, PA 19104-6216. Parts of this paper are drawn from a larger manuscript dealing with learning assessments in developing countries (Wagner, 2011).

[2] See Hoxby (2002) for a review. Also see Wolff (2008), p. 14, who states that, "testing in Latin America, as well as the USA, is not a significant financial burden -- constituting 0.3 percent or lower of the total budget of the level of study tested."

[3] See, for example, Johnson (1999, p. 70): [T]he cost of active participation [in IEA studies] is high, too high for many developing countries to bear. Where developing countries have been able to finance their participation, one might wonder whether the expense of that involvement could possibly be fully justified, given what has been learned, and given alternative uses for the funds consumed. How useful is it for Thailand, South Africa and Colombia to find themselves at or near the bottom of the international rank order in science, while Korea and Japan appear at the top and European countries are scattered throughout?" Also, see Postlethwaite, 2004 (p. 17), who stated that the total burden of assessment "varies between an annual international expenditure of 200,000 US dollars for SACMEQ through about 3.6 million US dollars for PISA to about 7 million US dollars for IEA. This is without counting what the countries themselves have to pay for their own staff members working on the projects and the cost of data collections."

[4] See Lockheed, 2008, p. 10, on LICs. Topol et al. (2010) provide a recent review of the US efforts the costs of more complex assessments in the US, where it is claimed, in part, that improved technology can reduce costs of increased R&D. But since LICs are, for the time being, hampered by technological constraints, the increased costs of R&D will likley end up as further bottom line expenditures.

[5] As Braun and Kanjee (2006) note, "in educational systems that lack basic resources, decisions to fund national assessments are extremely difficult to make" (p. 24).

[6] International Association for the Evaluation of Educational Achievement (IEA). See Chromy, 2002, p. 84 for a listing of major studies; also Lockheed, 2008, p. 6.

[7] According to a survey of national policy makers (Gilmore, 2005, p. 45), World Bank funding has been a key determinant of decision-making in LSEA adoption for low- and middle-income countries.

[8] "Hybrid" refers to drawing together discrete elements of various LSEAs, national curricular assessments and tests that were initially designed as cognitive assessments of reading and other basic skills. See Wagner (2011) for an in-depth review.

[9] See also: Siniscalco, 2006; Ravela et al., 2008; Wolff, 1998.

[10] For example, such a choice occurred in South Africa when it was decided not to participate in the TIMSS, citing the overall cost in time and resources (Greaney & Kelleghan, 2008, p.75). Also on South Africa, see Braun and Kanjee (2006, p. 19).

[11] These data were acquired as part of a jointly sponsored project by the Fast Track Initiative and UNESCO-IIEP (Wagner, 2011); and we thank the various agencies and their representatives for providing these data, some of which are estimates, as indicated in Table 3. Percentages are rounded to nearest whole number.

[12] Sample sizes of international assessments compiled across countries can yield much larger population totals, and numbers of participation countries continue to increase. For example, PISA (2006) had more than 400,000 students participating from 57 countries.

[13] It should be noted that not all the data were complete for each category, or reflective of full actual costs. For example, the only available PASEC data were those projected costs for the 2010 assessments; only three sources provided *test fees* data; and several sources provided no data for the *processing and analysis* or *dissemination* categories. Further, noting the ranges above, some categories demonstrated more variability than others. For example, *processing and analysis* includes average expenditures from .02% to 24.8%, while apart from three assessments (Honduras national assessment 2004, PASEC 2010 and PISA Uruguay 2003), dissemination expenditures had a mean of 5.9%. In addition, analysis would also need to account for the hidden costs or even unspecified costs discussed above – for example, costs in the *other* category for PISA Chile 2009 was over 7%.

[14] Wolff (2008, p. 19) states: "…[L]arge samples can be expanded to censal testing at a low marginal cost, since the fixed costs of developing items and pilot testing can be amortized over a larger population."

[15]Preventive medicine highlights the need for good and timely information. Timely information can make the difference between life and death or the spread of an epidemic or its curtailment. Proactive measures cost less and help avert the worst. Preventive medicine is encouraged not only to avert illness, but also to reduce costs of diagnosing and treating that illness (Szucs, 1997). Similar arguments can be made in the education arena. For instance, absenteeism and drop-out are well-known problems in LICs, incurring huge financial and social costs. Two measures have been highly successful against these difficulties: decreased grade repetition (Ndaruhutse, 2008) and increased bilingual education (Grin, 2005; Heugh, 2006). Assessments can assist in both detecting and "diagnosing" schooling difficulties earlier – from the cognitive to the socio-behavioral – thus heading off costly student problems such as dropout. In other words, even if SQC-style diagnostic tools cannot easily determine the 'best' remediation plan of action (which may be varied and complex), the early-detection aspect will nearly inevitably be cost-effective in the long run.

Assessing Student Engagement in China: Responding to Local and Global Discourse on Raising Educational Quality

Heidi Ross
Indiana University

Yuhao Cen
Indiana University

Zejun Zhou
Indiana University

China's heated education policy climate in 2010 indicated an increasing national concern for improving educational quality and educational quality assessment. Despite glowing portraits of Chinese education painted by international observers, the Chinese public has expressed consistent dissatisfaction with educational quality. The inter-related research projects described in this article were launched with a desire to deepen comparative discussion of educational quality and to respond to China's drive to improve and assess educational quality across all levels of schooling. This paper will introduce how educational quality is framed in key policy reform documents shaping Chinese education over the next decade. This will provide the backdrop for findings from two research projects that represent an effort to re-focus quality and quality assessment debates on high school and college students and their educational experiences. Derived and adapted from over a decade of robust research on student engagement in the U.S., the research projects include surveys on student engagement, a key factor in effective educational practice largely missing from Chinese quality assessment frameworks and toolkits.

Not long before Shanghai's PISA results caught the world's attention in 2010, Chinese educational authorities launched plans to support the country's goal to "build a moderately prosperous society in all respects by 2020" 1 (Hu, 2007). Shortly after, China experienced a year of educational media attention and policy-making frenzy. A key document, titled Outline of China's National Plan for Medium and Long-term Education Reform and Development 2010-2020 (hereafter the Blueprint) debuted in July 2010 after three years of planning, multiple revisions, expert and public involvement. The Blueprint outlines goals for all stages and aspects of education over the next decade. In January 2011, the State Council, the "cabinet" of the central government of the People's Republic of China, issued a follow-up action plan for the Blueprint (hereafter the Action Plan) detailing key implementation projects.2

These two documents have received unprecedented public scrutiny and provide the context for our examination of how Chinese policy makers and educators are re-conceptualizing educational quality. Serious public concern for quality assessment features prominently in new policies in response to a suffocating environment of test score-equivalent-quality and outcome-

 Current Issues in Comparative Education, Vol. 14(1): 24-37

centered assessments. As a result of reforms characterized by rapid educational expansion, devolution of funding, diversification of institutional missions, and the quest for world class schools, a nationwide debate has centered on how to define, measure, and achieve educational effectiveness and innovation. Despite glowing portraits of Chinese education painted by international observers, Ministry of Education (MOE) officers, school administrators, and researchers have expressed consistent dissatisfaction with existing mechanisms. These are primarily standardized tests, research activity, and infrastructure measures for assessing the quality of student, faculty, and institutional outcomes. The inter-related research projects described in this article were launched with a desire to deepen comparative discussion of educational quality and to respond to China's drive to improve and assess educational quality across all levels of schooling.

The research projects' use of student engagement surveys represents an effort to re-focus on the educational experiences of students in relation to education quality and new quality assessments. The phrase "student engagement" denotes the amount of time and effort students put into their studies and other educational activities in high school and college, and what they think of these experiences (Fredricks, Blumenfeld, & Paris, 2004; Kuh, 2003). Derived and adapted from robust research on student engagement at Indiana University, the National Survey of Student Engagement-China (NSSE-C) and the High School Survey of Student Engagement-China (HSSSE-C) were designed to measure student engagement at undergraduate and high school levels respectively, a key factor in effective educational practice largely missing from Chinese quality assessment frameworks. The two are distinct in their emphasis and methodology, but together provide insight on how debates about quality are largely consistent across secondary and post-secondary schooling. The surveys represent the first evaluation instruments to be used in China that focus on the concept of student engagement. A key feature of this project is the simultaneous development of surveys for both high school and college students. Tracking the development of student engagement across the secondary and post-secondary years is crucial to understanding educational quality in the Chinese context. High-stakes tests (and students' ability to pay) determine college matriculation and also shape college experiences. A focus on student engagement allows researchers to explore factors that have impact on college access and success. It also allows policy makers and practitioners to address and act upon these factors. HSSSE-C data allow us to probe whether students' activities in high schools are consistent with the normative performance demands reported by college students in NSSE-C. Complementing the high-stakes performance tests that characterize the Chinese secondary school experience, HSSSE-C also allows us to begin to identify specific educational processes that are linked to outcomes that Chinese standardized and entrance examination tests measure.

In this article, we first analyze education quality improvement and quality assessment discourse in the new policies and their mandates for Chinese high schools and colleges. We then trace the trajectories of the NSSC-C and HSSSE-C projects and research findings in relation to quality improvement and assessment. Our discussion of the NSSE-C data draws on research already published in China, while our analysis of HSSSE-C data draws on its first 2007 pilot study in

Shanghai. We conclude by summarizing the significance of our student-centered approach, survey instruments, and specific conclusions for China's new reform era.

An Analysis of Chinese National Policy on *Zhiliang* Assessment and Evaluation for High Schools & Higher Education

Akin to their counterparts in the U.S. and worldwide, Chinese policy documents are frequently scrutinized to evaluate changing policy climates. Examination of *tifa*, the Chinese expression for "how policies are framed and formulated," provides a useful starting point for comprehending what education policies intend and how their implementation is envisioned. Our brief examination of the *Blueprint* and corresponding *Action Plan* highlights two salient themes: systematic improvement of educational quality and innovation in education quality assessment/evaluation. This analysis serves as a backdrop for our introduction of the two student engagement projects, the HSSSE-C at the high school level and the NSSE-C at the undergraduate level.

The *Blueprint*'s preamble summarizes China's major educational challenges, namely that, "teaching contents and methods are relatively outdated, schoolwork burdens on primary and middle school students are too heavy, the promotion of quality education is arrested, our students are weak in their adaptability to society, and innovative, practical and versatile professionals are in acute shortage" (Blueprint, 2010). Quality improvement emerges as a catchphrase for resolving these perceived weaknesses,[3] as evidenced by two of the *Blueprint*'s five guiding principles, "reform and innovation" and "improving quality of education" (Blueprint, 2010). One crucial innovation in the principle of "reform and innovation" is education quality assessment/evaluation reform. The proposed reform mandates for high schools that "a scientific teaching quality evaluation system shall be in place, and academic proficiency tests and comprehensive evaluation of student quality should be instituted throughout senior middle school education" (Blueprint, 2010). For higher education, the *Blueprint* advocates "a project to ensure undergraduate teaching quality and to transform college education shall be undertaken comprehensively. Supervision over teaching shall be tightened up, and teaching quality guaranteed institutionally. College teaching evaluation shall be improved" (Blueprint, 2010). Of particular relevance to student-focused institutional reforms, the *Blueprint* reiterates the significance of quality assessment/evaluation as a complex process demanding diverse assessment approaches that involve multiple stake holders:

> We will improve the evaluation of teaching. We will set up scientific and diverse benchmarks for such evaluation, according to teaching goals and concepts on talents or professionals. *Teaching quality shall be evaluated with the participation of government, schools, parents and communities.* We will keep records of students and improve the assessment of comprehensive quality. *Diverse evaluation approaches that help promote student development shall be explored to encourage students to be optimistic and independent and become useful persons* [emphasis added by authors] (Blueprint, 2010).

In addition, the *Blueprint* highlights quality assessment/evaluation of education processes in key reform experiments. For higher education this involves improving undergraduate education quality and teaching quality; for high schools "an educational quality monitoring and evaluating system shall be perfected and the findings of the evaluation should be publicized at regular intervals" (Blueprint, 2010). The *Action Plan* adopts similar priorities. One of its ten special reform experiments is establishing an education quality monitoring and assessment mechanism for elementary, middle and high schools. As for higher education, the *Action Plan* aims to improve quality and implement an innovation-oriented academic evaluation system. Despite the fact that the *Blueprint* and *Action Plan* aim to diversify assessments and broaden participation, student voice and experience are largely missing and educational outcomes are prioritized over processes. As noted in the extended quotation above, students are mentioned merely in passing and only with regard to increasing participation in assessment/evaluation.

The National Survey on Student Engagement (NSSE-China): Focusing Reform on Improving Undergraduate Learning

At the postsecondary level, quality improvement in 'talent training' (*rencai peiyang*), or teaching and learning, stands out in the *Blueprint*:

> Establish the central status of talent training in higher educational institutions. Cultivate specialized talents and innovative talents that are persistent, ethical, knowledgeable and competent. Increase input in teaching and learning. Prioritize teaching in faculty evaluation…Enhance the quality assurance system in teaching and learning. Improve assessment of teaching and learning in higher education. (Blueprint, 2010)

Here, assessment of teaching and learning in higher education primarily refers to the 2002-2008 National Undergraduate Teaching and Learning Evaluation (*Quanguo benke jiaoxue gongzuo shuiping pinggu, abbr. Pinggu*), a state-initiated and sponsored evaluation program. In addition to government effort in assessing quality, a number of centers and institutes publish annual university ranking reports. These rankings, regarded as assessments of institutional quality, have drawn enormous attention from the public as well as scholars and policy makers.

Distinct from these efforts of quality assessment, the NSSE-C student engagement survey provides an alternative approach to addressing and assessing quality in higher education by adopting a student-centered perspective. It was modeled after the National Survey of Student Engagement (NSSE), which obtains information on an annual basis from North American four-year colleges and universities about student participation in programs that are provided for their learning and personal development or engagement. Administered by the Center for Postsecondary Research at Indiana University since 2000, NSSE has attracted more than 1,400 four-year institutions in the United States and in Canada to participate in the annual survey (NSSE, 2010). The NSSE instrument offers item-level data and summary institutional performance scores based on five benchmarks of student engagement (NSSE, 2000) compared with peer institutions. These benchmarks include the level of academic challenge, active and

collaborative learning, student-faculty interaction, enriching educational experiences, and supportive campus environment. As institutions are using NSSE as an assessment tool, they also seek to convert results from the engagement surveys into actions that improve student experiences and educational effectiveness. Each benchmark represents a domain area that is conveyable and actionable on campuses.

Chinese policy makers and educational researchers explicitly sought international models of higher education as they debated the structures and processes that lead to world class educational quality. Heidi Ross wondered whether an appropriately contextualized survey such as the NSSE might provide a useful springboard for promoting cross-national dialogue on educational quality for three key reasons. First, student engagement surveys gather student responses about their college experiences and data about educational *processes* instead of *inputs* for teaching and learning. These inputs include infrastructure, expenditures on undergraduate education, teaching materials, and faculty numbers holding advanced degrees, which are prioritized in existing assessment systems. Second, the student engagement survey elicits student voices by inviting students to share, deliberate on and learn from their perspectives on the quality of education received in colleges and universities. In contrast, existing assessment systems of national evaluations or rankings entail minimum participation of college students, the very participants who are described as the center of higher education. Last but not least, results from the student engagement survey identify practices in higher education that are actionable for institutional diagnosis and improvement. The student engagement survey not only serves as an assessment tool for accreditation and accountability, but also provides institutions with information they can use to enhance educational quality.

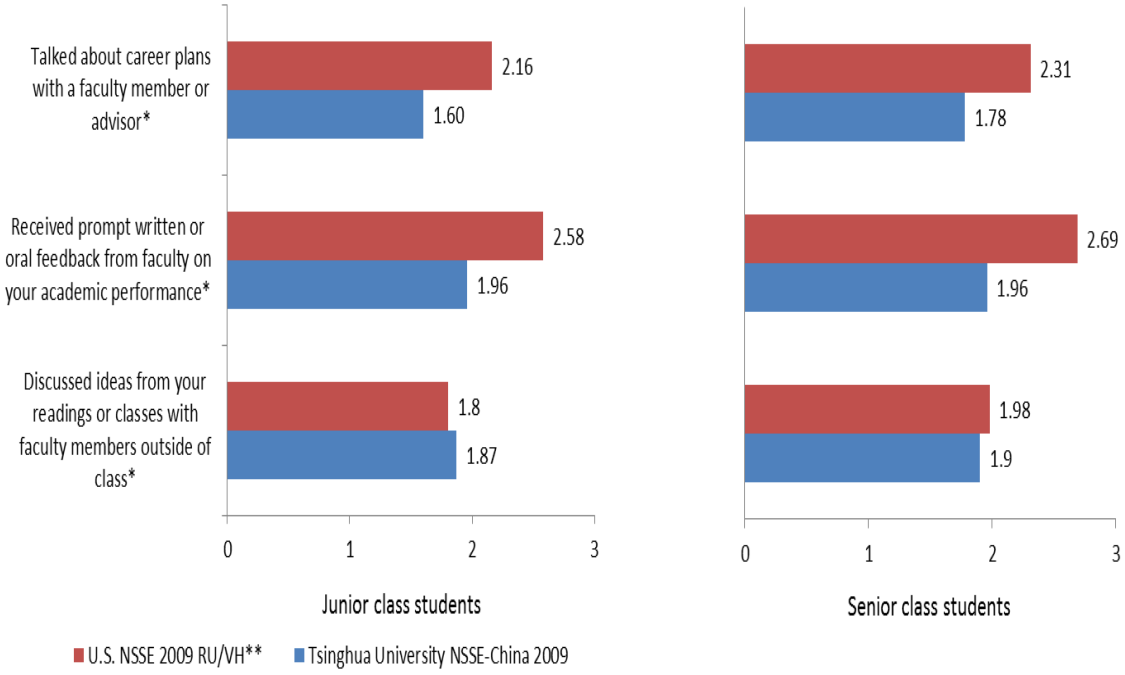
Initiated by Ross in collaboration with Tsinghua University in China, NSSE-C has developed into an influential project in China. Based on NSSE, the NSSE-C instrument was translated into the Chinese language and adapted to the Chinese context[4] by a team of doctoral students and a visiting professor from Tsinghua University in the fall of 2007. The instrument was pre-tested in China with pilot surveys in six institutions in Beijing in the winter of 2007, and further tested with cognitive interviews in five institutions of various types and in different regions in China, during the summer of 2008.

The first full survey administration in 2009 was joined by 27 voluntarily participating institutions throughout China. In April 2009, the NSSE-C research team held a national workshop at Tsinghua University on project goals and survey administration. Participating institutional researchers gathered at Tsinghua University again in December 2009 to discuss experiences in analyzing, reporting and utilizing the survey data. Two key questions structured discussions: (1) Was the goal of assessment *for improvement* being adequately addressed by participating institutions? And (2) Were the survey data merely an addition to respective institutional data banks, or utilized by institutional researchers to inform policies and practices? One institutional case we introduce here suggests that the NSSE-C project has begun to enrich teaching and learning quality assessment discussions and practices.

Based on the NSSE-C 2009 data collected from undergraduate students at Tsinghua University, Dr. Yan Luo and her colleagues published a report on undergraduate education quality. The report triggered a series of initiatives. Three examples of such activities were university-wide discussion and policy drafting in student-faculty interaction, professional development for undergraduate teachers, and directed attention to student learning as an alternative view of quality education. Institutional pressures to assess and improve undergraduate education quality came from both global competition for talent and domestic demand for accountability. In the past few years, overseas universities and especially those in Hong Kong have begun to lure high school graduates away from the best universities in mainland China with either fellowship provision or reputation in quality teacher and learning. As an elite member of China's *Project 985* [5] – comprised of institutions that aim to become world-class universities – Tsinghua University secured a fiscal allocation of 1.8 billion RMB *yuan* for this project alone. Evidence for quality undergraduate teaching and learning is seen as crucial to demonstrating accountability. By comparing itself with high ranking institutions in the United States on the measures of undergraduate college experiences and effective educational practices, Tsinghua was responding to external and internal pressures for quality assessment. In this context, NSSE-C was perceived as an appropriate and timely cross-cultural assessment tool that took undergraduate education seriously. Beyond quality assessment, NSSE-C data also provided direct evidence to support policy and practice reforms at Tsinghua, leading to quality improvement in undergraduate teaching and learning.

Data and policy recommendations shared here include the benchmark of student-faculty interaction. Researchers found that Tsinghua undergraduates reported much less student-faculty interaction compared with their peers in the United States. Item by item comparison (Figure 1) was used to demonstrate that Tsinghua undergraduates scored significantly lower than their U.S. peers in talking with faculty about career plans and in receiving faculty feedback on academic performance. Of the survey respondents, 27.1 percent reported that they had *never* received prompt feedback from faculty on their academic performance, and 44.3 percent *never* had a discussion regarding career plans with any faculty member or advisor. Comparative figures from the U.S. NSSE 2009 survey were about 7 percent and 20 percent respectively.

Figure 1 Comparing Tsinghua Undergraduate and U.S. Peer Student-faculty Interaction Data source: Tsinghua data from Luo et al. (2010); U.S. data from NSSE (2009a, 2009b)



*Grand mean of the item scale: 4="Very often", 3="Often", 2="Sometimes", 1="Never".
**Carnegie classification: Research universities with very high research activities

These findings provided a solid foundation for policy recommendation on improving the student-faculty relationship at Tsinghua University through promoting undergraduate research with faculty and changing faculty office hour policies. Beyond spurring reform at the institutional level, Tsinghua's published results have also been widely read and commented upon by academic and non-academic readers, and have subsequently influenced how universities and the wider public might come to perceive the quality of higher education as student-centered and student-valued.

Like Tsinghua University, other institutions that participated in NSSE-C are faced with fierce competition for students and strong pressure for accountability. China's transition from elite to mass higher education in the last decade has transformed the college landscape. It is now characterized by a more diverse student population, an increasingly explicit stratified hierarchy of higher education institutions, and a common perception that education is a private investment rather than a public good. The number of students studying in China's tertiary sector today is greater than that of the U.S., as nearly 25 percent of the age cohort is paying its way through an increasingly expensive system of public, private, and hybrid institutions in

order to compete in the labor force. This generation of students and their parents experience more choices and more risks. They are asking tough questions of college administrators and teachers, demanding affordable and marketable education that is both relevant and individually-tailored. Competing for students, faculty, funding, and prestige, college administrators seek alternative measures to assess whether their programs address the needs and desires of their clients. NSSE-C provided such a measure.

Looking back, NSSE-C and related institutional actions prefigured the *Blueprint's* call to "enhance the quality assurance system in teaching and learning and improve assessment of teaching and learning in higher education" (Blueprint, 2010). This call is not new rhetoric being imposed from the top echelons. Universities craving educationally sound methods to stay competitive as Chinese higher education has become increasingly entangled in the dense web of global higher education have been motivated to effectively assess and improve undergraduate education with a student-centered approach, or "the central status of talent training" (Blueprint, 2010).

The High School Survey of Student Engagement China (HSSSE-China): Refocusing Reform on Student Voice in Educational Experience

Chinese high schools are keenly aware of how they are inextricably linked to yet distinct from colleges, and this dual focus is captured in our conclusions regarding the findings of the HSSSE-C. Shanghai students' performance on PISA 2009 took the world by surprise, although it should not have (OECD, 2010). Individuals such as Robert Compton and President Obama called it "our generation's Sputnik moment" (Dillon, 2010). Sensational news headlines affirmed China's image as an educational superpower that educated students through a carefully articulated curriculum with high expectations and high incentives for success.

Student experiences and educational processes need to be assessed to interpret Shanghai's "stunning" scores and the significance they hold in relation to students. Yet, as noted above, assessing educational quality as defined by the *Blueprint* and the *Action Plan* generally leaves out student experiences; students are arguably considered objects of education reform rather than masters of their own education. How education and learning is perceived by students themselves (not to mention their teachers) remains largely unknown and empirically under-examined, and studies of student engagement can begin to fill in some of the missing pieces.

Debuted in 2004, the U.S.-based High School Survey of Student Engagement (HSSSE) administered by the Center for Evaluation and Education Policy (CEEP) at Indiana University is the largest of its kind in the United States. Since 2006, more than 350,000 students in over 40 states have taken the survey (Yazzie-Mintz, 2010). HSSSE investigates the attitudes, perceptions, and beliefs of students about their work, their school learning environment, and their interaction with the school community. The survey is structured in three dimensions (compared with NSSE's five benchmarks), including engagement of the mind, the heart, and engagement in the life of the school.[6] The survey examines a set of relations that shape student experience in high schools, namely, relations "between the student and school community, the student and

school adults, the student and peers, the student and instruction, and the student and curriculum." Distinctive in its student-centered and process-focused approach, HSSSE, like NSSE, was created to complement if not counter student achievement outcomes assessed through test scores, graduation rates, and adequate yearly progress.

Since September 2007, HSSSE researchers have collaborated with their Chinese counterparts to create HSSSE-C, addressing quality assessment concerns of Chinese high schools that are similar to those outlined by their U.S. counterparts. The HSSSE-C project involves a collaboration of researchers and graduate research assistants at CEEP and East China Normal University. In 2007 the translation and cultural adaption of the survey was completed, and the first pilot study was conducted in a Shanghai high school with 119 participants. In 2008 the survey was revised by both parties and a second pilot study was completed in 16 high schools, located in 15 provinces and municipalities across China, involving approximately 8,000 participants.

Like NSSE-C, HSSSE-C has been adapted to retain a reliable student engagement-focused core while being culturally sensitive to the Chinese context. The first pilot indicated that most Chinese high school students understood the instrument well and accurately responded to questions. We have also found potential ambiguities and inapplicable items. The revision pertains to twenty-seven changes designed to allow for a more precise and relevant report of student lives.[7] HSSSE-C's validity is manifested by the fact that its findings highly resonate with the existing literature on Chinese high school student life.

Quality Education Initiatives/Failures in Chinese High Schools and HSSSE's Potential

Literature on Chinese high school education abounds, but lacks depth and nuance. With some astute exceptions (OECD, 2010) most of it critically examines high stakes testing, namely the National College Entrance Examinations (NCEE), and its ensuing consequences for student learning. The culture of high school is portrayed in most accounts as conscribed by a severe exam-centered ethos that generates two often cited criticisms, namely: psychological pressure and rote learning. In 2009, only 24.2 percent of high school graduates entered higher education institutions in China. Achieving a high NCEE score and matriculating to a good college or university are considered an essential route to social mobility in China. Coupled with China's one-child family policy, college degree inflation since the tremendous expansion of higher education in the late 1990s[8] has exacerbated the competition among high school students seeking admission into prestigious Chinese universities. Twelve hours of schoolwork a day is commonplace for high school students, not to mention extra weekend tutorials arranged by parents. Most students do not express a strong intrinsic motivation for learning. Many are under tremendous pressure, and most pressure centers on test anxiety and fear of failure, and sleep deprivation is reported as a widespread problem (Liu, Uchiyama, Okawa, & Kurita, 2000). Many Chinese researchers make the point that failure to bring the NCEE in alignment with both national reform policy and inquiry-based pedagogical and curricular reforms will severely hamper educational effectiveness and innovation.

To counter these phenomena, the MOE and other agencies have repeatedly called for *Suzhi* [9] education since the 1990s, aiming at transforming examination-centered education into quality education (State Council & Chinese Communist Party Central Committee, 1999; State Education Commission, 1997). *Suzhi* education emphasizes fostering an innovative spirit and practical ability by, for instance, reforming curricula to encourage a holistic approach to education (Dello-Iacovo, 2009). The effectiveness of the new mandate generally has remained at a low level (Zhao, 2007). It is in this reformative atmosphere that HSSSE-C has been developed to add student voices to China's secondary education assessment toolkit.

Preliminary Findings: Disengaged Minds and Hearts

Data from HSSSE-C's pilot survey administered in Shanghai in 2007 underscores challenges to quality teaching and learning reported in the existing literature on Chinese high schools. In open-ended responses, participating students questioned the meaning of high schools. They expressed frustration with rote learning, dull instruction, and anxiety about the anticipated fierce competition of the college entrance examination.

In response to the question "Why do you go to school?"[10] out of thirteen respondents, only one student stated, "It is fun." Five students reasoned that they go to school because it is a necessary means of social mobility. Four students said that "It is my parents' wish." When asked about whether they were bored in classes, and why, fifteen participants responded.[11] They listed reasons closely connected with a high school ethos centered on college examination preparation. Four participants described boring lessons; four participants described lack of sleep; three participants described low interest in study; one complained about too much homework. Only two students indicated personal difficulties. Typical complaints include, "The class is too quiet," "Not much response between teachers and students," and "I have much homework to do at night. It's very late whenever I finish it."

In the last open-ended question, in which participants were asked to provide additional comments, one respondent presented a disheartening portrayal of high school as, "*counterproductive to our well-being as teenagers. We are given almost no free time, not to mention enough time to rest. What we study is useless. We study for college entrance exams only, nothing else. My life as a high school student is disinteresting, what I do is to repeat this dull life every day.*" This quotation and similar responses, attest to the daunting challenges with which Chinese educators, policy makers, students and parents have been wrestling. These students expressed neither optimism nor appreciation for a school environment in which, from their perspective, independent learning barely thrives. Using student engagement as a window on the experience of students studying in one of China's most cosmopolitan, well-funded, educationally adaptive and intensely competitive environments raises important questions about the purpose and value of high school. As Zhang Minxuan, head of Shanghai PISA testified, Shanghai's PISA performance needs to be interpreted with confidence in the direction of Shanghai's educational reform, reflexivity in the realization of issues, gaps, new ideas and methods in educational practices in light of worldwide trends in student assessment - particularly the issue of study overload.

Conclusion

In response to the demands for quality education and global competitiveness, another wave of educational reform is rising in China. This wave is global in its visibility and centers on education quality assessment and improvement as indicated in the *Blueprint* and the *Action Plan.* The intent of the policy is quite clear in that more expansive and rigorous assessments involving various actors in education have been initiated. In addition, international assessments, such as PISA at the high school level and university rankings at the higher education level, have been sought out and applied in China. These existing assessment practices are outcome and/or reputation-driven and have overlooked the student experience and educational processes. The research projects summarized here shine a spotlight on student experience and direct attention in institutionally actionable ways to the subjects of education and educational processes. Student engagement as one focus of quality assessment respects students as masters of their own education and injects their experiences into the dialogue of educational effectiveness reform policy.

Successful adaptation of an American assessment tool to address key issues in educational development and reform in China illustrates that student-centered and process-driven assessment can cross boundaries and mutually inform highly diverse institutions and cultures. On the other hand, adaptation shows that China is actively responding to global trends in defining and redefining educational quality. Chinese reformers and educators consider exam-centered education as the key impediment to creating "innovative, practical and versatile professionals" (Blueprint, 2010). Tracking the development of student engagement across the secondary and post-secondary years will be crucial to understanding and assessing Chinese educational quality and the reforms designed to achieve it. To date, NSSE-C has drawn significant attention from policy makers, scholars, and student services personnel at both national and institutional levels, and has catalyzed a broader definition of educational quality and through changing quality assessment methods. Although HSSSE-C data from diverse cities and regions of China await final analysis, together with NSSE-C the survey tool enriches China's education quality assessment protocols. It provides the means for appreciating, documenting, and enhancing students' educative experiences to inform an educational system perceived to be both constrained by outmoded convention, whilst simultaneously ahead of its international competitors in several measures of significance.

References

Dello-Iacovo, B. (2009). Curriculum reform and 'Quality Education' in China: An overview. *International Journal of Educational Development, 29*(3), 241-249.

Dillon, S. (2010). Top test scores from Shanghai stun educators. *New York Times* Retrieved April 29, 2011, from http://www.nytimes.com/2010/12/07/education/07education.html

Fredricks, J. A., Blumenfeld, P. C., & Paris, A. H. (2004). School engagement: Potential of the concept, state of the evidence. *Review of Educational Research, 74*(1), 59-109.

Hu, J. (2007). Chairman's Report at the 17th National Congress of the Chinese Communist Party. *Xinhua News Agency.* Retrieved April 29, 2011, from http://news.xinhuanet.com/newscenter/2007-10/24/content_6938568_3.htm

Kuh, G. D. (2003). The National Survey of Student Engagement: Conceptual framework and overview of psychometric properties.

Liu, X., Uchiyama, M., Okawa, M., & Kurita, H. (2000). Prevalence and correlates of self-reported sleep problems among Chinese adolescents. *Sleep: Journal of Sleep Research & Sleep Medicine, 23*(1), 27-34.

Luo, Y., Shi, J., & Tu, D. (2010). Annual report of Tsinghua undergraduate education survey 2009: Comparing with U.S. top research universities. *Tsinghua Journal of Education, 30*(5), 1-13.

Ministry of Education, China. (2010). Outline of China's National Plan for Medium and Long-term Education Reform and Development (2010-2020) (English version translated by Australian Education International). Retrieved from http://www.aei.gov.au/News/newsarchive/2010/Documents/China_Education_Reform_pdf.pdf

National Survey of Student Engagement. (2000). National benchmarks of effective educational practice. Bloomington, IN: Indiana University Center for Postsecondary Research.

National Survey of Student Engagement. (2009a). NSSE 2009 Grand Means Means and Standard Deviations by Carnegie Classification: First-year Students. Bloomington, IN: Indiana University Center for Postsecondary Research.

National Survey of Student Engagement. (2009b). NSSE 2009 Grand Means Means and Standard Deviations by Carnegie Classification: Senior Students. Bloomington, IN: Indiana University Center for Postsecondary Research.

National Survey of Student Engagement. (2010). Major differences: Examining student engagement by field of study—annual results 2010. Bloomington, IN: Indiana University Center for Postsecondary Research.

OECD. (2010). Shanghai and Hong Kong: Two distinct examples of education reform in China. In Organisation for Economic Co-operation and Development (Ed.), *Strong performers and successful reformers in education: Lessons from PISA for the United States* (pp. 83–115). Retrieved from http://www.oecd.org/dataoecd/34/45/46581016.pdf

State Council, & Chinese Communist Party Central Committee. (1999). The Chinese Communist Party Central Committees and State Council's Decision on further promoting educational reform and Suzhi education. Retrieved April 11, 2011, from http://www.edu.cn/20051123/3162313.shtml

State Education Commission. (1997). On promoting Suzhi education in primary and secondary education. Retrieved April 11, 2011, from http://www.edu.cn/20051018/3156152.shtml

Yazzie-Mintz, E. (2010). Charting the path from engagement to achievement: A report on the 2009 High School Survey of Student Engagement. Retrieved from http://www.indiana.edu/~ceep/hssse/images/HSSSE_2010_Report.pdf

Zhao, Y. (2007). China and the whole child. *Educational Leadership, 64*(8), 4.

[1] This 2020 goal was put forward in 2007 at the 17th Congress of the Chinese Communist Party.
[2] The Action Plan's official title is *Notice of the State Council on the Experimental Sites for the National Educational Reform*.
[3] China's Blueprint references "education quality" fifty-one times, twenty-eight of which relate to strengthening education quality. Assessment/evaluation is mentioned thirty-nine times, among which ten references are pertinent to education quality assessment/evaluation. The Action Plan mentions education quality improvement six times among seventeen quality-relevant references.
[4] Adaptations fall into four categories: language-driven adaptation (e.g. the word "presentation" was paraphrased as "oral report" as there was not exact counterpart of the word in the Chinese language), adaptation to ensure local coverage of a concept (e.g. "student-faculty interaction" was substituted with "student-teacher interaction", as the concept of teacher is broader than faculty and more relevant in the context of Chinese higher education), adaptation to ensure questions are understood as intended (e.g. multiple versions of phrasing were tested in the field to make sure that survey respondents' understanding was consistent with what was intended by us researchers), and social, system-driven adaptation (e.g. the item asking about spiritual practices was removed as they were minimum at Chinese colleges and universities).

[5] Institutions in Project 985 are most prestigious universities in China. The Project was launched by the Ministry of Education soon after the then-President Jiang Zemin's speech in May 1998 that called for a number of world-class universities. The Project was named after the year (98) and the month (5) of Jiang's speech. Project 985 has over 40 institutions altogether, and the announced fund for these 985 institutions in total was approximately 30 billion RMB Yuan.

[6] It should be noted that initially, the HSSSE derived from the NSSE in 2004. Since 2005, it has been administered and developed by the CEEP.

[7] For example, examination scores in China largely determine college matriculation, thus the item to evaluate whether student parents converse about college application strategies in the HSSSE is irrelevant should Chinese students score below designated lines. Another example is questions of minimal relevance such as work for pay and AP classes in the HSSSE. Work for pay under sixteen is illegal in China. AP classes are rarely offered in Chinese high schools.

[8] Since the expansion, gross enrollment rate to higher education roared, but more competition comes if one wishes to enter prestigious universities.

[9] Scholarship lacks consensus on the definition and translation of the 'Suzhi' education. Common translation is quality education, quality-orientated education and diathesis education

[10] 119 students took the survey, but only 13 responded to this open-ended question.

[11] 15 students answered this open-ended question among 119 survey takers.

CICE Current Issues in Comparative Education

Volume 14(1) / Fall 2011 • ISSN 1523-1615 • http://www.tc.edu/cice

Evaluation, Assessment, and Testing

FEATURED ARTICLE

BOOK REVIEW

CURRENT ISSUES IN COMPARATIVE EDUCATION
Volume 14, Issue 1 (Fall 2011)

Issue Editors:
Ruaridh J. MacLeod, Sarah Flatto

On the Right Track: Measuring Early Childhood Development Program Quality Internationally

Maria Cristina Limlingan
Tufts University

Two of the main obstacles for drawing educational comparisons consist in determining what are considered 'high quality' initiatives, and finding a common tool that can adapt to differences in both structure and content, as well as to the cultural and demographic characteristics of the population it wishes to serve. This paper focuses on addressing such obstacles by inquiring whether an instrument such as the Early Childhood Environment Rating Scale (ECERS) can be used to compare the early childhood development initiatives in countries like Chile and Bangladesh. Using Pena's (2007) model that considers linguistic, functional, cultural and metric equivalence, we examine the implementation of the ECERS in these two settings, and identify the factors of significance in the instrument's successful adaption to a different context.

Comparisons between early childhood development (ECD) programs offer many potential benefits but are often difficult to execute. Two of the main challenges when drawing comparisons are (1) determining which ECD programs are considered high quality initiatives; and (2) finding a common tool that can adapt to differences in the structure and content of educational systems, as well as the cultural and demographic characteristics of the population it wishes to serve. Ensuring the quality of the program is the primary challenge for the reason that much of the success of an ECD program depends on the quality and approach (Magnuson, Ruhm and Wadofogel, 2007; Myers, 2004). Previous ECD interventions that have produced positive outcomes for children have been designated 'high quality' programs, but little has been written to describe exactly what high quality means. Quality is often relegated to indicators such as teacher-student ratio, teacher education level and teacher experience (in terms of years). These structural factors are often used as indicators of quality because data are obtained easily and studies show that there is a relationship between these factors and childhood outcomes (Peisner-Feinberg & Burchinal, 2001). However, though these structural factors are undeniably useful, they are unable to provide a comprehensive picture of other elements of importance within early childhood settings. Process elements such as teacher-student interaction, learning opportunities and the kinds of activities available are important to look at because these provide a better picture of what is happening, and allow agents to see which areas require increased focus in a child's immediate environment, the very setting which is the most influential and meaningful for the child.

The second challenge is to make comparisons on ECD programs in different cultural contexts. Cross cultural methods enable researchers to test, modify and extend current theories of child development by providing insight into factors in child development that can either be universal or local (Pena, 2007). Applying cross cultural methods allows agents from different regions to gather comparative data in order that it serve as a reference point for the examination of their

respective education systems. Comparisons are also a good way to summarize data that can be easily understood, and it can be used to construct a more persuasive argument during the decision-making process. The availability of information about different early childhood practices has been steadily increasing, however the kind of data collected is usually a mere consolidation of structural elements present in a country's ECD initiative, with little information about how the relative levels of ECD process quality elements compare (Tietze, Bairrao, Leal & Rossbach, 1998).

This paper focuses on addressing these obstacles of determining ECD quality in the international context by inquiring if an instrument such as the Early Childhood Environment Rating Scale (ECERS) can be used to compare the ECD initiatives in countries like Chile and Bangladesh. It will then examine the implementation of the ECERS in these two settings to determine what was considered when the instrument was adapted to a different context.

ECERS and Considerations for Application

Even if educational, socio-economic and cultural differences exist in various regions, certain elements have been recognized as necessary for a child's positive development because of their recurring presence within successful interventions and the literature on the subject. These elements include safe and healthful care, developmentally appropriate stimulation, positive interaction with adults, encouragement of individual emotional growth and the promotion of positive relationships with other children (Tietze et al., 1998). The ECERS covers many of these elements through its seven subscales in the following areas: personal care routines of children, furnishing and display for children, language-reasoning experiences, fine and gross motor activities, creative activities, social development and adult needs. Scores are obtained using a 7-point scale, ranging from inadequate (1) to excellent (7), with each item providing a description of the salient features which need to be observed (Harms, Clifford & Cryer, 1998). Scoring is based on a 2-3 hour classroom observation, and includes a teacher's interview conducted after the observation process. Researchers and practitioners are also encouraged to undergo intensive training to ensure reliability. The ECERS is part of a series of assessments that cover infant-toddler settings (Infant Toddler Environment Rating Scale or ITERS), day care (Family Day Care Environment Rating Scale or FDCERS) and school age children (School Age Children Environment Rating Scale or SACERS). What is unique about the ECERS is its attempt to measure the quality of classroom features enabling the occurrence of pre-identified key processes, whilst simultaneously assessing the processes themselves in the current ECD environment (Villalon et al., 2002). During the creation of the ECERS, Harms, Clifford and Cryer (1980) debated whether to include interpersonal relationships in the scale. The authors relate, however, that they found it impossible to ignore interpersonal behavior and deal adequately with the environment at the same time. The inclusion of items that deal with children's interactions allows teachers and researchers to adopt a more comprehensive means of assessing ECD settings.

As a result of its extensive usage across different countries, changes were made on the ECERS. The adjustments were based on information from a content analysis of the relationship of

ECERS to other global quality instruments, an examination of early childhood program documents, data from studies using the ECERS in preschool and child care settings, and feedback from ECERS users (Harm, Clifford and Cryer, 1998). The result was the Early Childhood Environment Rating Scale Revised Edition (ECERS-R). This version retained basic similarities to the ECERS to ensure continuity, but eliminated some questions to avoid redundancy, provided more detailed descriptions and added items such as health and safety practices, television and computer use and a greater use of interaction questions (Harms, Clifford and Cryer, 1998).

In the United States, results from the ECERS and ECERS –R have been used extensively in research that has examined associations between preschool quality and child development, and dozens of investigations have demonstrated an association between higher scores on ECERS observations and a child's developmental outcomes (Peisner-Feinberg & Burchinal, 1997; Peisner- Feinberg et al., 2001). Results from the ECERS and ECERS – R have also been used to monitor the quality of program(s) and provide guidance for improving quality, and there is ample evidence suggesting that an assortment of program investments, technical assistance, and professional development efforts can be used to improve such scores (Bryant et al., 1994, Howes, Phillips and Whitebook, 1992).

Aside from being used in various capacities in the United States, the ECERS has also been used in other countries where it has been adapted to measure program improvement efforts. Early childhood may exhibit similar elements as being necessary for successful development, but the outcomes in children's development vary and relate in different ways to the many different measures of quality in the ECERS (Sylva et al., 2006). The difficulty in using instruments created in another country is that the standard measures of quality are based on expertise relevant to one region that may not be appropriate for another context. This poses a significant problem if researchers are unable to ensure that the instrument being used retains its relevance and ability to measure in a fair manner. Pena (2007) offers four important features for establishing a study's validity that need to be considered when conducting research across different cultural groups. The first and most commonly known measure, according to Pena (2007), is linguistic equivalence. This refers to translating both instructions and the instrument, and checking to make sure that the words are appropriate for the context in which they are used. Functional equivalence is concerned with ensuring that the instructions and instruments will elicit the same target behavior. Cultural equivalence looks at how respondents will interpret a given direction or test item, and determines if there are possible underlying cultural interpretations that may affect the way an individual responds to the instrument and instruction. Finally, metric equivalence deals with addressing the change in the level of difficulty that might occur (Pena, 2007).

In order to see the effectiveness of the ECERS and the ECERS-R in the international context, we examine the implementation of the instrument in Chile and Bangladesh and look at how the measures were able to address the issues of linguistic, functional, cultural and metric equivalence. Chile and Bangladesh were selected based on their use of the ECERS-R and the

availability of information on how these two countries were able to adapt the instrument to meet their needs.

Chile

Early childhood education has been present in Chile for a long time and increasing its utilization has been one of the priorities of the government (Herrera, Mathiesen, Merino & Recart, 2005). One reason for the uneven coverage may be the fact that even if children from birth to age six can attend ECD programs, attendance is not compulsory (Herrera et al., 2005). Currently, one out of four children from low income families attend ECD programs, compared to one out of two medium or high income children (Herrera et al., 2005). In order to address these issues, the government hoped to first evaluate the existing quality of public preschools by examining how well they met the needs of children and their families.

To carry these goals out, Villalon et al. (2002) conducted a study that compared the different types of preschool (private non subsidized, private subsidized, city council preschool, national program, and those sponsored by non-governmental organizations) that were offered in terms of quality. Using a Spanish version of the ECERS, Villalon et al. (2002) dropped the cultural awareness item which assessed the provision of materials and activities related to diverse context because of its low mean. Ratings were also obtained from 33 experienced pre-school teachers using a questionnaire that asked them to rate quality criterion on a three point scale. In the resulting data, both the *provision of special needs* item and the *naps* item scored very low on the scale. According to Villalon et al., (2002), the reason for the low score may relate to the fact that it is not common practice to have naps within a half day curriculum or to integrate children with disabilities in mainstream early childhood programs. Items such as sand and water, space to be alone, furnishings for relaxation and comfort were rated as relatively important by experts, compared to the remaining 33 items that were rated as very important. Prior to data collection, preschool teachers were trained on the use of the scale until they reached an agreement of 95%. Data collection was done in the middle of the school year in June and July, meaning that classes were settled within a daily routine.

Villalon et al. (2002) found that the average ECERS scores fell into the minimal quality category for the seven subscales with scores that ranged from a low of 3.09 for social development and a high of 4.58 for personal care. Significant differences were found among the two regions (that is, metropolitan and rural areas) in which the study was conducted. Despite differences among six different types of preschool, *personal care routines* and *fine and gross motor skill* areas had the highest average score across the board, while *creativity* and *social development* had the lowest scores.

Bangladesh

Early childhood education programs have grown in popularity with governments in developing countries as a way to prepare the children from high-risk families for school (Myers, 1992). Aboud (2006) confirms that this is the case in Bangladesh, where the early home environment alone is unable to adequately prepare children for school due to the prevalence of

factors such as high malnutrition and low parental education. While it is common for mothers to stay at home, they perceive their role as protecting their child from illness and injury rather than encouraging play and conversation (Aboud, 2006). Children have little exposure to books, media and other educational toys. The purpose of the study by Aboud (2006) was to assess the curriculum typically used in Bangladesh with the objective of developing, through use of different activities, skills related to the process of learning, positive learning attitude and individual learning styles.

Since the study was conducted in a rural setting, a modified version of the ECERS for the South India context called the Tamil Nadu or TECERS (Isley, 2001) was used, together with the ECERS-R. Contextual adaptations were made with the ECERS-R in order to define terms quantitatively. For example, "enough blocks" meant 20 blocks allotted per child, and "some books" meant 10 books. The physical settings used in the TECERS subscale are relevant to a rural setting that has to deal with various indoor and outdoor hazards such as availability of water at toilet and open defecation or urinating. Since the ECD program was a half day program, items concerning meals and naps were excluded. Nine items received the lowest score because there were no televisions, videos, soft toys or cozy areas, and little attempt to protect privacy. Two new subscales were included to address requests arising from parents in the population, and these related to literacy and math activities such as attending to environmental print, emphasizing sounds in words, writing letters and numbers, and counting and matching objects (Aboud, 2006). Twelve research assistants with university degrees were trained for 5 days to conduct testing, and the ECERS-R measures were practiced at nearby schools. Data were collected from October to mid-November, the end of the Bangladeshi school year.

Aboud's (2006) study indicates that the results of ECERS-R ranged from 1.8-3.7 on a 7-point scale, while the converted scores from TECERS ranged from 5.2-5.9 on a 7-point scale. The highest ratings were from the areas of mathematics and literacy, with the lowest scores relating to the areas of activities and program structure, largely due to the scarcity of available fine motor materials.

Comparison between Chile and Bangladesh

In spite of their different contexts, the Chile and Bangladesh studies were able to adapt the ECERS and ECERS-R effectively to suit their needs. The primary goal for linguistic equivalence is to make certain that the words and linguistic meaning used in the instruments and instructions are the same for both versions (Grisay, 2003; Sireci & Berberoglu, 2000). For Chile, the English edition of the ECERS-R was used together with a Spanish translation, Spanish being the language more widely spoken in the country. Simply translating instruments, however, may be insufficient to guard against potential biases and validity threats, and it is important, therefore, to scrutinize the instructions and the choice of words used. In the case of the Chile project, it might have been beneficial to have an expert teacher review the Spanish version of the ECERS before the other teachers were asked to review the tool. In this way, the expert teacher could check the content for differences in word usage between the Spanish used in Spain and that used in a Latin American country such as Chile, differences which may

significantly alter the meaning of the items. Future researchers might also want to consider utilizing the method of back-translation to protect against biases in language. This process consists of a translator who first translates the instrument or instructions from the source language to the target language, then involves another translator who independently translates the target version back to the source language (Pena, 2007). Once both translations are complete, the two versions are compared to identify differences and resolve them. Together with the translation/back-translation, decentering may occur wherein the instrument with translated items may have shifted away from the wording of the source instrument to represent the concept in a manner familiar within the target language. The content resulting from the dual process of back translation and decentering would represent the final version of a tool that is functionally equivalent and linguistically different, yet wherein both versions elicit linguistically similar responses.

Functional equivalence was accounted for by training those who would be using the ECERS to score items consistently. This process was done in Bangladesh, where the researcher and a local Bangladeshi research colleague conducted a five day training course with research assistants who had university degrees. Ensuring the reliability of the research assistants also included having the trainers accompany them during their initial classroom observations and at least one other time during the 6-week data collection. It was noted in the Chile example that the preschool teachers who conducted the classroom observations had previously been trained to use the ECERS, and had reached an agreement of at least 95%.

Metric equivalence is a crucial factor, especially when making comparisons between different instruments. In the Bangladesh study, the researchers decided on using the TECERS because the rural setting of the study made the items on the TECERS more relevant, especially the items relating to the physical environment and personal care and hygiene. However, after using the TECERS in 6 classrooms, the researchers realized that due to the way the TECERS instrument was constructed – with items having a restricted scoring range from 0 – 2 – there was insufficient variability to perform correlations with the ECERS-R. Thus when the TECERS was converted to the ECERS-R rating, the preschool classrooms achieved a 'high quality' designation – that is, exhibiting scores between 5 to 6 on a 7 point scale – which did not seem to correspond to what the researchers had actually observed. Because the TECERS was not metrically equivalent to the ECERS-R the use of the TECERS in the study was discontinued.

Even if items meet the criteria for linguistic, functional and metric equivalence, researchers must also be careful, when considering cultural equivalence, to look and see if items may have salience for different groups due to the distinct cultural and historical ways in which concepts are interpreted by respondents. In the case of the study in Chile, the researchers decided to present a questionnaire with the ECERS items to 33 experienced preschool graduate teachers including those from universities, the ministry of education, directors of preschool centers and preschool practitioners for them to review. The result of this consultation was that certain items such as *nap time*, *children with special needs* and *use of books and videos* were either taken out or modified beforehand, in accordance with the teacher's input, so that the scale would give each

preschool a fair chance with the scores for the ratings. It was interesting to note that the scoring of items as being only *relatively important* amongst the Chilean experts was consistent with the low scores these items obtained in many of the observed Chilean preschool centers, with around 50-78% scoring a 1 in the 7-point scale.

When looking at the results, it is also necessary to understand the cultural context and the outcomes that are important to each specific country. With regard to Chile, similarities found in the ECERS profiles of the different types of preschool assessed demonstrate the influence of shared educational values and orientation in Chile's early childhood education system. Despite significant differences in quality level, some areas were either consistently higher or lower across the entire range of institutions involved, regardless of whether they were private non-subsidized, private subsidized, city council preschool, national program, and non-government organization-sponsored preschool centers. From a practical perspective, what can be taken from the Chile study is the fact that even if there is variation in quality within the type of program, good quality classrooms were identified for each type and this could serve as a model or reference point for raising standards at the classroom level. The challenge then would be decreasing variability within the type of programs and trying to achieve a higher standard.

The Bangladesh project added scales to the ECERS-R such as *literacy*, *math* and *interpersonal interaction* to reflect the importance parents and educators attached to these academic indicators. The fact that the schools scored lower on scales relating to *activities* and *program structure* was largely due to the small variety of challenging materials, and the lack of teacher assistance with individual child progress via hands-on activities and scaffolding. The low results were also consistent with the cultural emphasis on memorization of math phrases and stories rather than on reasoning and vocabulary, a tendency which appears to attenuate the effect of both interest and comprehension amongst the children (Aboud, 2006). Although teachers may be more comfortable teaching by demonstration, these may not be the best methods for enabling children to learn math and language. From these results, certain recommendations were adopted to increase the amount of stimulation children received through materials, activities and instruction (Moore, Akhter, and Aboud, 2008).

In a follow up study by Moore, Akhter & Aboud (2008), curriculum changes in this half day program gave more prominence to language and literacy. Targeted measures, such as daily story reading – with several new stories introduced each week instead of each month – and teaching instructors how to read and talk about stories in an engaging manner (rather than simply requiring memorization) were introduced. The format of the learning was also changed so that it emphasized working in small groups or pairs rather than continuously in small groups. The results of such changes were an increase in ECERS-R scores in *activities* and *program structure* subscales from 3.5 to a score of 4.7 and 6.5 respectively. The increase in scores was also associated with some observable gains in child outcomes (Moore, Akhter and Aboud, 2008). What is even more promising was that the cost to make these changes was estimated at $1.50 per child, per year (Moore, Akhter and Aboud, 2008). In this way, Bangladesh was able to pinpoint and act upon specific areas requiring improvement within their curriculum.

Conclusion

Comparative international research represents an attempt to understand local problems from a global perspective in a world becoming more interconnected and interdependent at all levels (Villalon, Suzuki, Hererra & Mathiesen, 2002). A common instrument is a good way of enriching our knowledge of what is important in different countries. Providing a list of quality measures helps us both to see a more comprehensive picture as well as to make informed decisions about what is important, and what should be modified, added or taken out. For example, mainstreaming special education was not emphasized in Chile, but the government may want to examine this in the future to see if it may be appropriate to their national setting. The rural areas of Bangladesh lacked technology such as television and computers, and this may be an area that they would like to consider developing when planning later on.

Having instruments like the ECERS that measure the quality of early childhood development can serve as a foundation for other countries, especially those that do not have existing measures in place. As demonstrated in Chile and Bangladesh, it is possible to use an instrument such as the ECERS to measure ECD quality internationally. It is important to remember that quality for ECD initiatives must be contextualized in ways that are relevant to the values of a group, and those who plan to use such instruments must be able to take into account linguistic, functional, cultural and metric equivalence. For Chile, the ECERS enabled them to identify areas that were of local and national importance and to compare quality across program types. As a result of this, they may look at ECD programs that are better performing for each type and find ways to improve the quality of their programs. In Bangladesh, they examined a commonly used curriculum in the rural areas. Additional scales were used to adapt to the rural surroundings and the needs of the parents and educators. In the end, they were able to pinpoint the areas that scored the lowest and to make the subsequent decision to work on them to improve child outcomes within a follow-up study.

As the world becomes smaller, it will become more important to find the most effective ways to organize and share information. Cross cultural comparisons using a common instrument, so long as it is composed and utilized in the right way, provides a good method to facilitate discussions which allow us to learn from one another. It is only by doing so that we can hope to avoid making the same mistakes as in the past, and to build a better future for our children.

References

Aboud, F. E. (2006). Evaluation of an early childhood preschool program in rural Bangladesh. *Early Childhood Research Quarterly*, *21*(1), 46-60.

Moore, A.C, Akhter, S., Aboud, F.E (2008). Evaluating an improved quality preschool program in rural Bangladesh. *International Journal of Educational Development*, 28, 118–131.

Bryant, D., Burchinal, M., Lau, L.B and Sparling, J.J (1994) Family and classroom correlates of

Head Start Children's Developmental Outcomes. *Early Childhood Research Quarterly*, 9, 289-309.

Early Childhood Environment Rating Scale (ECERS) Website (2007). Accessed from http://www.fpg.unc.edu/~ecers/ on October 30, 2007

Grisay, A. (2003). Translation procedures in OECD/PISA 2000 international assessment. *Language Testing*, 20, 225–240.

Harms, T., Clifford, R.M.and Cryer D. (1980). *Early childhood environment rating scale*. New York: Teachers College Press.

Harms, T., Clifford, R. M., & Cryer, D. (1998). *Early childhood environment rating scale*. revised edition. New York: Teachers College Press.

Herrera, M. O., Mathiesen, M. E., Merino, J. M., & Recart, I. (2005). Learning contexts for young children in Chile: Process quality assessment in preschool centres. *International Journal of Early Years Education*, *13*(1), 13-27.

Howes, C., Phillips, D., & Whitebook, M. (1992). Thresholds of quality: Implications for the social development of children in center-based child care. *Child Development,* 63, 449-460.

Isely, E. (2001). *Tamil Nadu early childhood rating scale*. Chennai, India: M.S. Swaminathan Research Foundation

Myers, R. G., & 20 Consultative Group on Early Childhood Care and Development(1992). *The Twelve who survive : Strengthening programmes of early childhood development in the third world*. High/Scope Press.

Myers, R.G (2004) . *In search for quality programmes of early childhood care and education*. Background paper for Education for All, Global Monitoring Report 2005. Paris: UNESCO. www.unesco.org/education/gmr_download/references_2005.pdf; retrieved December 13, 2007

Magnuson, K., A., Ruhm, C., & Waldfogel, J. (2007). Does prekindergarten improve school preparation and performance? *Economics of Education Review*, *26*(1), 33-51.

Peisner-Feinberg, E., Burchinal, M. (1997). Concurrent relations between child care quality and child outcomes: The study of cost, quality, and outcomes in child care centers. *Merrill-Palmer Quarterly*, 43 , 451-477.

Peisner-Feinberg, E. S., Burchinal, M. R., Clifford, R. M., Culkin, M. L., Howes, C., Kagan, S. L., et al. (2001). The relation of preschool child-care quality to children's cognitive and Social developmental trajectories through second grade. *Child development*, *72*(5), 1534-1553.

Pena, E. D. (2007). Lost in translation: Methodological considerations in cross-cultural research. *Child development*, *78*(4), 1255-1264.
Sireci, S. G., & Berberoglu, G. (2000). Using bilingual respondents to evaluate translated-adapted items. *Applied Measurement in Education*, 13, 229 – 248.

Sylva, K., Siraj-Blatchford, I., Taggart, B., Sammons, P., Melhuish, E., Elliot, K., & Totsika, V., (2006). Capturing quality in early childhood through environmental rating scales. Early *Childhood Research Quarterly*, *21*, 76-92.

Tietze, W., Bairrao, J., Leal, T. B., & Rossbach, H. (1998). Assessing quality characteristics of centerbased early childhood environments in Germany and Portugal: A cross-national study. *European Journal of Psychology of Education*, *13*(2), 283-298.

Villalon, M., Suzuki, E., Herrera, M. O., & Mathiesen, M. E. (2002). Quality of Chilean earlychildhood education from an international perspective. *International Journal of Early Years Education*, *10*(1), 49-59.

CICE Current Issues in Comparative Education

Volume 14(1) / Fall 2011 • ISSN 1523-1615 • http://www.tc.edu/cice

Evaluation, Assessment, and Testing

FEATURED ARTICLE

BOOK REVIEW

CURRENT ISSUES IN COMPARATIVE EDUCATION
Volume 14, Issue 1 (Fall 2011)

Issue Editors:
Ruaridh J. MacLeod, Sarah Flatto

BOOK REVIEW

Hadjiyanni, M. (2008). *Contesting the Past, Constructing the Future: A Comparative Study of the Cyprus Conflict in Secondary History Education*. Saarbrucken: VDM Verlag Dr. Muller Aktiengesellschaft & Co. KG. 152 pp. ISBN 978-3-639-01973-5

– *Reviewed by Demetrios Spyridakis, Teachers College, Columbia University*

In this pioneering study, Marina Hadjiyanni carefully analyzes the biases inherent within history textbooks assigned in the secondary schools of Cyprus. A small Eurasian country situated at the northeastern end of the Mediterranean basin, Cyprus has become the focus of international attention since the Turkish military invasion of the northern portion of the island in 1974. The Turkish invasion came in response to the actions of a Greek nationalist regime that staged a *coup d'etat* in Cyprus with the ultimate goal of annexing the island to Greece. The invasion itself has sparked considerable controversy and poisoned Greek-Turkish relations ever since. Though the United Nations, the European Court of Human Rights, and the international community have denounced Turkey's military aggression and refuse to recognize a separate Turkish-Cypriot state in the north, the consequences of the invasion continue to resonate over thirty-five years later.

With the 1974 invasion so deeply imbedded in the psyche and historical consciousness of Greek and Turkish Cypriots alike, it is hardly surprising that secondary level history textbooks from both communities present radically different interpretations of the island's division. After carefully investigating a number of such textbooks, Hadjiyanni affirms that the historical narratives of both communities lack objectivity. From the onset, Hadjiyanni clearly states two important aims of her study: (1) to corroborate her argument that both Greek and Turkish Cypriot textbooks are biased by using textual evidence, and (2) to provide perspectives on the biases from history teachers and educational researchers who actively propose solutions to this problem.

Hadjiyanni's overarching theme is both sensible and well-argued: that Greek and Turkish Cypriot educators should abandon efforts to create a nationalistic barrier, which creates an "us" versus "them" dichotomy between the two communities within Cyprus. Instead, they should actively work toward historical peacebuilding in which mutual tolerance, reconciliation, and an appreciation for differing perspectives might facilitate the emergence of a united Cyprus. Given the overwhelming scale of the Cypriot problem, Hadjiyanni's suggestion that peace might be facilitated through historiography may at first strike the reader as idealistic and impractical. But few would disagree with the contention that nationalistic tendencies have their roots in historical understanding. It can be argued that this situation provides educators with a significant role in helping

shape a nation's identity. After all, as Samuel Butler noted, "It has been said that though God cannot alter the past, historians can."

The book is organized into six chapters and includes a list of five appendices at the end. The first chapter provides a pithy introduction, stating the study's research questions and themes. Here the reader learns of the novelty of the study, as Hadjiyanni correctly asserts that, prior to the appearance of her book, "no specific research was conducted on how the historical period of 1955-1974, which led to the de-facto separation of the island, has been depicted in the secondary school history textbooks on each educational context" (10). In the second chapter, Hadjiyanni surveys the literature addressing various academic curricula authored by credible educators who analyze how history can be taught for peacebuilding, as opposed to promoting nationalism. Together, these authors provide one common insight: the teaching of history can either exacerbate a terrain of struggle in the classroom or grant a more objective avenue through which students might unite. They suggest this can be achieved by building a spirit of academic collaboration which, in leading students to understand the past experiences of humanity, creates the conditions by which they can help bring about positive social change in the future.

In chapters three, four, and five, Hadjiyanni analyzes both a Turkish-Cypriot and Greek-Cypriot textbook produced by the Ministry of Education for each community. The Turkish-Cypriot textbook was written for children aged from fourteen through fifteen, and the Greek-Cypriot book for students aged from seventeen through eighteen years of age (36). Hadjiyanni documents examples of bias from both textbooks. The conclusions she draws from her analyses are that both the Greek-Cypriot and Turkish-Cypriot textbooks conceal acts of wrongdoing committed by their own side, and that both groups "seem to reinforce nationalistic and intolerant attitudes among Greek-Cypriot and Turkish-Cypriot students" (88).

After examining the two textbooks individually, Hadjiyanni describes the process of interviews she conducted in which she would (a) question the participants about their perspectives on the textbooks, and (b) request suggestions for rectifying the problem of bias in textbooks. Conducting interviews of 45-50 minutes each (41), she conversed with the following five participants: a Turkish-Cypriot history instructor who argued that the Turkish textbooks are biased and communicate a "notion of revenge" (90); a Turkish researcher, holding a Ph.D in History, who was forced to resign from a textbook revision committee which she claimed consisted of "very strict Turkish-Cypriot nationalists" (92); a Greek-Cypriot educator trained in philology and history, who contended that the "role of Turkish-Cypriots during 1955-1964 is completely excluded," and that "Greek-Cypriot teachers usually avoided integrating this issue in the history classroom" (92); a Greek-Cypriot researcher and instructor of pre-primary teachers/trainees in history who asserted that Greek textbooks provide "chauvinistic representations" (93); and a Danish history teacher at a private school in Greek southern Cyprus, who identified extant

educational propaganda inasmuch as students are "not given the opportunity to question and critically think about historical narratives" (94). The common solution provided by the participants involves the revision of textbooks so that they can "discuss alternative historical perspectives and interpretations" (98). Finally, in chapter 6, Hadjiyanni summarizes her research findings.

Notwithstanding the book's numerous strengths which have been addressed *supra*, a few glaring weaknesses deserve mention. First, in Hadjiyanni's literature review appearing in chapter two, she includes a brief section titled "The Cyprus Problem in Historical Context." Paradoxically, Hadjiyanni's narrative contains no discussion of the significant historical events that transpired prior to the Turkish invasion of Cyprus in 1974. In fact, her historical discussion begins with the invasion itself. Thereafter, she provides commentary on the current disputes that exist over the island, namely that the "Turkish Cypriot leadership demands the resolution of conflict through the creation of a bi-communal, bi-zonal federation which recognizes political equality" (22), and that the Greek-Cypriots "demand a unitary, sovereign state with indivisibility of territory and single citizenship, which identifies the Turkish Cypriots as the minority" (22). Given the complexity of the Cyprus dispute and the widely divergent positions concerning its origin, the lack of general information is sure to leave unfamiliar readers puzzled. Only when analyzing the Greek-Cypriot and Turkish-Cypriot textbooks in subsequent chapters does Hadjiyanni include some critical historical discussion, though it is scattered throughout the text. There, however, her historical discussions function more as personal commentaries and are often abruptly abandoned for other topics.

Second, in the historical context section of chapter two, Hadjiyanni simplifies her coverage of the Annan Plan to a discussion of the failure of UN General Secretary Kofi Annan to convince the people of Cyprus to approve his 2004 proposal, which Greek-Cypriots perceived to be unfair. Missing are the key features of the Annan Plan which the Greek-Cypriots found objectionable: for example, the failure to demilitarize the island, the lack of guarantees against the unilateral intervention of foreign powers, and the absence of a property recovery system for displaced Greek-Cypriot refugees. Unfortunately, a lack of appropriate historical context pervades the book and significantly detracts from Hadjiyanni's aim to educate readers about the biases found in history textbooks. Readers without the proper factual background and objective historical data will find it difficult to identify biases present in history textbooks and, by extension, to evaluate the merits of Hadjiyanni's assessments of those biases.

Finally, a review of the eighty-six sources consulted by Hadjiyanni reveals very few pertaining to the period from 1955 to 1974, a watershed in Cypriot history which she repeatedly emphasizes is her primary research focus. Since her work deals largely with the subject of history, her citation of only five secondary historical sources discussing the "Cyprus Problem" appears inadequate. Of these five sources, only two are books, one written by Volkan (1979) and dealing primarily with psychoanalysis, and the other

authored by the Cyprus Ministry of Education (2003). The Cyprus Ministry of Education, however, mainly comprises specialists in the field of curriculum development, rather than historians. Books written by experts in history would have been more appropriate sources to consult. Moreover, the other three cited sources are concise essays by Papadakis and Volkan appearing in a volume edited by Calotychos (1998), and a short article by Zervakis (2002). Also of concern is the absence of citations relating to primary sources.

A few contradictions within the text are also problematic. Hadjiyanni identifies a Turkish-Cypriot textbook's characterization of Nikos Sampson – the forcibly installed successor to Makarios as President – as the murderer of several British and Turkish individuals an item of propaganda for the purpose of justifying the subsequent Turkish invasion of Cyprus in 1974 (45). Curiously, however, Hadjiyanni then criticizes a Greek-Cypriot textbook for not properly documenting acts of persecution aimed toward the Turkish-Cypriots, especially by the leadership: "The Greek-Cypriot narrative does not provide any evidence that indicates the burden of responsibility of the Greek-Cypriot leadership for causing the sufferings of the Turkish-Cypriot minority" (82). The reader is left to wonder why Hadjiyanni would criticize the Turkish-Cypriot textbook's documentation of persecution as propaganda, and then censure the Greek-Cypriot textbook for failing to document acts of the same persecution. Hadjiyanni presents a similar inconsistency undermining her peacebuilding history argument when she reprimands the Turkish-Cypriot textbook authors for their "characterization of Turkish-Cypriots as the brothers of Greek-Cypriots" (52). Hadjiyanni opposes this characterization since the textbook "establishes the representation of Turkey as the peacemaker" (52). Hadjiyanni argues that Greek-Cypriots and Turkish-Cypriots should live harmoniously in a multicultural world united as one Cypriot nation, and "move beyond the homogenous national identity," (121) yet she skeptically dismisses a fraternal reference made by Turkish-Cypriots.

While Hadjiyanni's work is not without its flaws, the positive aspects far outweigh the negative. Hadjiyanni possesses a vast amount of knowledge about the state of affairs of Cypriot historical education and offers important insights in her analysis of nationalistic and peace building approaches to the teaching of history. She is sensitive to the challenges of teaching history in a more objective manner, and should be lauded for her efforts. Her work demonstrates how historical narratives can be skewed to justify a community's every action, presenting a significant obstacle in the way of peace. This book will surely appeal to scholars of both history and education. It remains a definitive study of the biases plaguing textbooks which shape the historical consciousness of a generation of Cypriots who have inherited one of the last divided capitals in Europe.