IERI Monograph Series

Issues and Methodologies in Large-Scale Assessments

VOLUME 5

October 2012

A joint publication between the International Association for the Evaluation of Educational Achievement (IEA) and Educational Testing Service (ETS)

ISBN 978-0-88685-412-6

Copies of this publication can be obtained from:

IERInstitute
IEA Data Processing and Research Center
Mexikoring 37
22297 Hamburg,
Germany

IERInstitute
Educational Testing Service
Mail Stop 13-E
Princeton, NJ 08541,
United States

By email: ierinstitute@iea-dpc.de
Free downloads: www.ierinstitute.org

Copyeditors: Paula Wagemaker, Editorial Services, Oturehua, Central Otago, New Zealand
with David Robitaille, Ruth R. Greenwood, and Tom Loveless
Design and production by Becky Bliss Design and Production, Wellington, New Zealand

IERI Monograph Series
Issues and Methodologies in Large-Scale Assessments

Volume 5 **2012**

TABLE OF CONTENTS

Introduction

Dirk Hastedt (Editor)
IEA Data Processing and Research Center, Hamburg, Germany

Matthias von Davier (Editor)
Educational Testing Service, Princeton, New Jersey, United States

The IEA-ETS[1] Research Institute (IERI) undertakes activities focused on three broad areas of work: research studies related to the development and implementation of large-scale assessments of educational outcomes, professional development and training, and dissemination of research findings and information gathered through large-scale assessments. Part of IEA and ETS's collaborative work focuses on improving the science of large-scale assessments. The IERI series *Issues and Methodologies in Large-Scale Assessments* is the institute's vehicle for disseminating research findings and helping to improve understanding of the science of large-scale assessments.

This fifth volume of our periodical includes four papers and a technical note. The four papers all make use of IEA international study data—three of them use IEA Progress in Reading Literacy Study (PIRLS) data and the fourth uses IEA Trends in International Mathematics and Science (TIMSS) data. The first two papers are methodological in nature. They deal with actual issues in large-scale assessment studies. The latter two deal more with the content of these studies and what we can learn from them. Interestingly, there is also a link in terms of content between the first and the third papers because both focus on SES measures, the first more from the contextual side, and the third more from a country perspective. The technical note discusses an important analytic matter of considerable relevance for researchers, namely, sampling weights and resampling procedures.

The first paper, "Measuring Family Socioeconomic Status: An Illustration Using Data from PIRLS 2006" by Daniel H. Caro and Diego Cortés, shows how to develop an indicator of socioeconomic status (SES) in international large-scale assessments (ILSA). A good number of analyses of ILSA data make use of SES indicators, mostly in order to eliminate the effect of SES in their models and thereby achieve "pure" results that are not influenced by students' SES. For this purpose, it is essential to have a good measure of the students' SES. This paper should help researchers reflect on SES measures and develop an SES measure that is best suited to their respective analyses.

1 International Association for the Evaluation of Educational Assessment-Educational Testing Service.

"Estimating Linking Error in PIRLS" by Michael O. Martin, Ina V. S. Mullis, Pierre Foy, Bradley Brossman, and Gabrielle M. Stanco analyzes the effect of linking errors across study cycles. When the history of ILSA began, the studies were one-time assessments, but today all major ILSAs also assess trends in achievement. This development has created new challenges, such as how to quantify errors when comparing data from two different cycles. The study examines the linking error between PIRLS 2001 and PIRLS 2006, a practice that hopefully all ILSAs providing trend measures will adopt in order to ensure reliable comparisons across cycles.

"Exploring the Measurement Profiles of Socioeconomic Background Indicators and Their Differences in Reading Achievement: A Two-Level Latent Class Analysis" by Kajsa Yang Hansen and Ingrid Munck also analyses SES information. The authors, using Swedish data from PIRLS 2006, split up different components of SES and then categorized students in terms of those components. From there, they analyzed the reading achievement of the students in the different groups and drew conclusions about potential explanatory mechanisms. This paper is interesting not only in terms of presenting an application of an innovative analytical approach but also in terms of the substantive conclusions drawn. The analytical approach the authors used might inspire other researchers when conducting their analyses.

The fourth paper, "Leadership, Learning-Centered School Conditions and Mathematics Achievement: What Can the United States Learn from Top Performers on TIMSS?" by Nianbo Dong and Xiu Chen Cravens, demonstrates how ILSA can be used to contrast national data from a specified country with the data from other countries. The authors compared school-level information from TIMSS 2007 and its relationship to achievement outcome measures in the United States to corresponding data from Taipei, Korea, Singapore, Hong Kong SAR, and Japan, some of the highest achieving countries in the study. The authors found both similarities and differences in achievement across the different cultural contexts. This article shows not only the possibilities but also the limitations of ILSAs.

The last contribution, "Rescaling Sampling Weights and Selecting Mini-Samples from Large-Scale Assessment Databases," by Eugenio J. Gonzalez, explains how to deal with large databases and how to subsample students from such databases. LSA databases are getting bigger and bigger, and despite increased computing power, researchers are finding many analytic procedures a challenge, and even more so given the models used for these analyses are becoming ever more complex. Also, it is essential that researchers know how to draw samples whenever they conduct analyses which require equal contribution from subsets of the participants or when they want to test models in several subsamples. The author of this technical note not only explains the principal techniques but also provides source code for SPSS to assist researchers in their analyses.

We are extremely pleased with the selection of high-quality papers presented in this fifth issue. We hope that you will find them interesting and inspiring for your own research. We also hope that you will consider supporting this periodical by submitting your own methodological research on international large-scale assessments to IERI.

ABOUT IEA

The International Association for the Evaluation of Educational Achievement (IEA) is an independent, non-profit, international cooperative of national research institutions and governmental research agencies. Through its comparative research and assessment projects, IEA aims to:

- Provide international benchmarks that can assist policymakers to identify the comparative strengths and weaknesses of their education systems;
- Provide high-quality data that will increase policymakers' understanding of key school-based and non-school-based factors that influence teaching and learning;
- Provide high-quality data that will serve as a resource for identifying areas of concern and action, and for preparing and evaluating educational reforms;
- Develop and improve the capacity of educational systems to engage in national strategies for educational monitoring and improvement; and
- Contribute to development of the worldwide community of researchers in educational evaluation.

Additional information about IEA is available at www.iea.nl and www.iea-dpc.de.

ABOUT ETS

Educational Testing Service (ETS) is a non-profit institution whose mission is to advance quality and equity in education by providing fair and valid assessments, research, and related services for all people worldwide. In serving individuals, educational institutions, and government agencies around the world, ETS customizes solutions to meet the need for teacher professional development products and services, classroom and end-of-course assessments, and research-based teaching and learning tools. Founded in 1947, ETS today develops, administers, and scores more than 24 million tests annually in more than 180 countries, at over 9,000 locations worldwide.

Additional information about ETS is available at www.ets.org.

Measuring family socioeconomic status: An illustration using data from PIRLS 2006

Daniel H. Caro and Diego Cortés
IEA Data Processing and Research Center, Hamburg, Germany

Many analyses of educational outcomes include a single socioeconomic status (SES) index as a predictor or for statistical control. In large-scale assessment research, these analyses have primarily examined the influence of SES on academic achievement in schools and the influence of other individual, family, and school variables when controlling for SES. The findings have helped elucidate the mechanisms underlying the influence of SES and identify possible avenues for reducing it. The measurement of SES is critically important with respect to the findings and implications of these analyses. This paper demonstrates a step-by-step procedure for calculating a family SES index using data from the Progress in International Reading Literacy Study (PIRLS) 2006. Variables reflecting parental education, parental occupational status, and family wealth were reduced into a single SES index using principal component analysis (PCA). The conceptual definition of SES, its reliability and crossnational comparability are discussed, and recommendations for further research and survey developers are offered. The presented procedure can be extended to other large-scale studies of educational achievement. Its illustration is the main contribution of this paper.

IERI Monograph Series: Issues and Methodologies in Large-Scale Assessments Volume 5

INTRODUCTION

Researchers have studied the influence of family background on academic performance mainly in school settings (Sirin, 2005; White, 1982). Studies indicate that the performance of students from socioeconomically disadvantaged backgrounds tends to be worse than that of their peers from more affluent families. Gaps between and across students from varying socioeconomic backgrounds tend to widen as students get older (Caro, McDonald, & Willms, 2009; Condron, 2007; Kerckhoff, 1993; Oakes, 1985) and have lasting consequences on the educational attainment and labor force outcomes of students as adults (Alexander, Entwisle, & Olson, 2007; Kerckhoff, Raudenbush, & Glennie, 2001; Raudenbush & Kasim, 1998; Rumberger, 2010).

National and international student assessment studies offer a significant opportunity to study socioeconomic gaps in academic achievement between and within countries, and thereby to increase our understanding of how socioeconomic inequality reproduces across generations. Although some analyses have included separate variables representing concepts such as human capital, social capital, cultural capital, and economic capital, each postulated by a well-established theoretical model (e.g., Bourdieu, 1977, 1986; Coleman, 1988), others have employed a single variable to represent family socioeconomic status (SES).

Family SES in this latter group of studies is typically defined as the relative position of an individual or family within a hierarchical social structure, based on their access to, or control over, wealth, prestige, and power (Mueller & Parcel, 1981). SES indexes are traditionally operationalized through measures characterizing parental educational levels, parental occupational prestige, and family wealth (Buchmann, 2002; Gottfried, 1985; Hauser, 1994; Schulz, 2005; Yang & Gustafsson, 2004). There is, however, no strong consensus on the conceptual meaning of SES, which limits its use for testing theories and making policy recommendations (Bornstein & Bradley, 2003; Deaton, 2002). Despite these limitations, many studies using the SES approach have made valuable contributions to educational research.

Practical Contributions of SES Studies

The many studies that have attested to the usefulness of including an SES index in analyses of student learning outcomes generally include an SES index for two main purposes (Buchmann, 2002):

1. To gain an understanding of the extent to which and the mechanisms by which family SES is associated with academic achievement; and
2. To evaluate the influence of individual, family, school, and community aspects while controlling for SES.

In regard to the first goal, Willms (2006a) proposed using socioeconomic gradients to examine the association between SES and academic achievement. Socioeconomic gradients display, in single gradient lines, the strength of the association between SES and academic achievement, the association within schools, and the association across schools. Also, the functional form of socioeconomic gradients (i.e., linear or

curvilinear) indicates whether the SES gap widens, narrows, or remains stable as SES levels increase.

Analyzing socioeconomic gradients helps us characterize the following:

- Achievement inequalities related to SES;
- How these inequalities relate to differences in family SES;
- Which aspect of the inequalities can be attributed to the SES composition of different schools; and
- The influence that the socioeconomic intake of schools has on achievement.

Analysis findings can also allow us to provide guidelines on the potential influence of different policy interventions (Willms, 2006a).

Many studies, including the Programme for International Student Assessment (PISA), conducted by the Organisation for Economic Co-operation and Development (OECD), have adopted the socioeconomic gradient framework for studying educational inequality within and across countries (e.g., Caro & Mirazchiyski, 2012; OECD, 2003, 2004, 2007, 2010; Willms, 2002, 2003, 2006b; Willms, Smith, Zhang, & Tramonte, 2006; Willms & Somers, 2001). Also, in a recent article, Caro and Lenkeit (2012) showed how researchers can use international assessment data to extend this framework to include the analysis of sociological theories. Using data from the Progress in International Reading Literacy Study (PIRLS) 2006, carried out by the International Association for the Evaluation of Educational Achievement (IEA), the authors used two-level and three-level hierarchical linear models to evaluate 10 hypotheses about educational inequality within and across countries.

Researchers have examined other relevant aspects of the association between SES and academic achievement. These include studies of whether and how:

- This association changes over time (Heath & Clifford, 1990; Willms & Raudenbush, 1989);
- Is mediated and moderated by risk and protective factors (Chao & Willms, 2002; Guo & Harris, 2000; Yeung, Linver, & Brooks-Gunn, 2002);
- Is consistent across subject areas (Ma, 2000);
- Changes over the course of schooling (Caro & Lehmann, 2009; Caro, McDonald et al., 2009); and
- Helps to explain school grouping or tracking decisions (Caro, Lenkeit, Lehmann, & Schwippert, 2009; Condron, 2007; Maaz, Trautwein, Lüdtke, & Baumert, 2008; Schnabel, Alfeld, Eccles, Köller, & Baumert, 2002).

Overall, these studies have provided us with a better understanding of the mechanisms behind the influence of SES on academic achievement.

The second goal of analyses including an SES index has been to control for SES differences when evaluating the influence of other variables on academic achievement. For example, school effectiveness research controls for school SES in order to capture

the effect of schools on academic achievement. Findings from school effectiveness research have helped identify school practices that affect student achievement positively irrespective of the socioeconomic composition of schools (Rutter & Maughan, 2002; Rutter, Maughan, Mortimore, & Ouston, 1979).

Other applications include studying associations with migration background, social capital, cultural capital, and other family variables net of the effect of SES. Also, PISA (OECD, 2003, 2004, 2007) sets out league tables that rank countries according to overall student performance while controlling for SES. Such rankings take into account the varied socioeconomic conditions in which education systems operate and thus provide a fairer picture of relative performance.

Research Purpose

Our main aim in this paper is to demonstrate the calculation and validation of an SES index, using data from PIRLS 2006. Unlike in PISA, an SES index is not available in PIRLS and other studies managed by IEA. Another aim is to stimulate discussion on the possibilities for and the importance of calculating and providing an SES index in IEA datasets. Apart from the implications that our work has for IEA studies, we hope it will provide researchers with a detailed procedure that they can extend to other national and international assessment studies.

Although we apply the traditional procedure for calculating SES, our approach is somewhat unusual on several counts. First, we employ coefficients other than the traditional Cronbach's alpha so as to offer a broader assessment of reliability. Second, we transform the job class classification scheme in PIRLS into an ordinal and internationally comparable occupational status scale. Third, we provide guidelines for SES assessment by placing the measurement of SES within literature on reflective and formative models. And, fourth, we offer researchers and survey developers recommendations relating to SES validation and improvement.

REFLECTIVE AND FORMATIVE MODELS

The literature distinguishes between formative and reflective models for construct measurement (Bollen, 1989; Bollen & Lennox, 1991). These models differ mainly in their underlying assumptions of the causal relationship between the construct and its manifest indicators. In particular, three broad theoretical considerations are indicative of this type of model (Bollen & Lennox, 1991; Borsboom, Mellenbergh, & van Heerden, 2003; Edwards & Bagozzi, 2000; Jarvis, MacKenzie, & Podsakoff, 2003). These are:

1. The nature of the construct (i.e., reflected or formed);
2. The direction of causality between the indicators and the construct (i.e., from the construct to the indicators, or vice versa); and
3. The characteristics of the indicators used to measure the construct (i.e., interchangeable or not).

The reflective view dominates psychological research, while the formative view is common in economics and sociology.

The construct in a reflective model reflects existing theories and exists independently of its manifest indicators. Causality flows from the construct into the observed indicators, but not vice versa, and the nature of the model suggests that adding or dropping one constituent indicator does not alter the definition of the construct. Personality and attitude scales are typical examples of reflective models.

In contrast, a formative scale is viewed as a weighted combination of the observed items, and the construct depends entirely on its operationalization. Causality flows from the observed indicators to the construct. The nature of this approach is such that the constituent components are not interchangeable. In other words, the definition of the construct is affected by the number and type of items.

A well-known example of a formative indicator is the Human Development Index (HDI). This index does not exist as an independent entity but is a composite measure directly determined by health, education, and income variables (United Nations Development Program, 2006). The measurement of SES is also more appropriate under a formative model than under a reflective model (Diamantopoulos, Riefler, & Roth, 2008). Rather than being grounded in solid theory, the SES construct is formed and depends largely on its operationalization. The manifest indicators are given in the literature (i.e., parental educational levels, parental occupational status, and family wealth) and are not interchangeable.

The rules for assessing formative indicators are not well established in the literature, and traditional methods are not appropriate. Reliability analysis assumes high intercorrelations among constituent components, but this is not necessarily the case for formative indicators, where the resulting indicator can be a composite of uncorrelated variables. Validity analysis assumes an underlying theoretical model, but the formative model is fully contingent on the operationalization procedure. A common practice for assessing formative indicators is to examine their effect on a benchmark variable (Bollen & Lennox, 1991).

SES MEASUREMENT

Data

We sourced our data from the PIRLS 2006 survey conducted by IEA (see Mullis, Martin, Kennedy, & Foy, 2007). PIRLS 2006, the second iteration of this survey (the third is presently underway), assessed the reading literacy of fourth graders in 45 education systems. Socioeconomic data were collected through student and home questionnaires (Foy & Kennedy, 2008). The home questionnaire surveyed parents about their educational attainment, jobs, and self-perceived financial situation. The student questionnaire included questions about home possessions. These questions were not available in the home questionnaire. The data relating to home possessions and self-perceived financial situation reflect family wealth. We calculated the parental education and parental occupation indicators of SES by recoding the original data, as follows.

SES Items

Parental education (momed/daded)

Parents were asked to state their highest level of education completed. Table 1 presents the original response options and the recoding scheme for the derived mother's and father's education variables. Original response options followed the International Standard Classification of Education (ISCED) (UNESCO, 1999) and were collapsed into five comparable categories for the mother (momed) and the father (daded).

Table 1: Parental education: coding of original and derived variable

Original variable		Derived variable	
Categories		*Code categories*	*Code original*
Some <ISCED Level 1 or 2> or did not go to school	1	Finished some primary or lower-secondary or did not go to school	1 1
<ISCED Level 2>	2	Finished lower-secondary	2 2
<ISCED Level 3>	3	Finished upper-secondary	3 3
<ISCED Level 4>	4	Finished post-secondary but not university	4 4 and 5
<ISCED Level 5B>	5	Finished university or higher	5 6 and 7
<ISCED Level 5A, first degree>	6	Missing	. 8 and .
Beyond <ISCED Level 5A, first degree>	7		
Not applicable	8		

Parental occupational status (momsei/dadsei)

The question on occupation in PIRLS aligns with a class-based or categorical rationale to social stratification wherein occupations/members of society are divided into a limited number of discrete categories/classes (see, for example, Goldthorpe, 1980; Goldthorpe, Payne, & Llewellyn, 1978). Accordingly, parents were asked to mark, from a set of options, the kind of work that constituted their main job (see Table 2). The categorical approach assumes that stratification and mobility processes are multidimensional in nature.

In contrast, the SES concept presupposes a hierarchical order that allows an unlimited number of SES groups to be captured in a single status dimension with a continuous scale (Ganzeboom, de Graaf, & Treiman, 1992; Mueller & Parcel, 1981). The continuous approach captures variability within occupation categories, is more amenable to multivariate analysis than the categorical approach, and does not neglect significant information from other occupational dimensions (Ganzeboom et al., 1992). Due to these advantages, and in order to maintain consistency with the SES hierarchical nature, we transformed the class-type job categories in PIRLS into a continuous scale that reflected occupational status.

Ganzeboom et al. (1992) proposed using a scoring scheme to construct their International Socioeconomic Index of Occupational Status (ISEI). This index has been used extensively in studies of social mobility. Table 2 converts job categories from the home questionnaire into the ISEI scores. We took the original job types and descriptions,

as stated in the questionnaires, and used them to find matching categories with the index developed by Ganzeboom and his colleagues. When multiple categories matched the descriptions, we assigned the average score to the job category, thus creating an ISEI variable for mothers (momsei) and fathers (dadsei).

Table 2: Parental occupational status (ISE scores)

Original categories		Score
1	*Has never worked outside the home for pay*	22
2	*Small business owner (< 25 employees)* Includes owners of small business such as retail shops, services, restaurants	57
3	*Clerk* Includes office clerks, secretaries, typists, data entry operators, customer service clerks	49
4	*Service or sales worker* Includes travel attendants, restaurant service workers, personal care workers, protective service workers, salespersons	45
5	*Skilled agricultural or fishery worker* Includes farmers, forestry workers, fishery workers, hunters and trappers	31
6	*Craft or trade worker* Includes builders, carpenters, plumbers, electricians, etc.; also metal workers, machine mechanics, handicraft workers	37
7	*Plant or machine operator* Includes plant and machine operators, assembly-line operators, motor-vehicle drivers	33
8	*General laborers* Includes domestic helpers and cleaners; building caretakers; messengers, porters and doorkeepers; farm, fishery, agricultural, and construction workers	24
9	*Corporate manager or senior official* Includes corporate managers such as managers of large companies (25 or more employees) or managers of departments within large companies; legislators or senior government officials; senior officials of special-interest organizations; military officers	67
10	*Professional* Includes scientists, mathematicians, computer scientists, architects, engineers, life science and health professionals, teachers, legal professionals, social scientists, writers and artists, religious professionals	73
11	*Technician or associate professional* Includes science, engineering, and computer associates and technicians; life science and health technicians and assistants; teacher aides; finance and sales associate professionals; business service agents; administrative assistants	52

Home possessions (hompos)

Students in all countries answered questions about the possessions they had available at home, using binary responses (yes/no) to do so. Possessions listed in the survey were a computer, a study desk, a daily newspaper, their own room, and own mobile phone. We fitted the one-parameter Rasch (1960) model, specifically suited for the analysis of binary responses, to the home possessions items data. We then used the conditional maximum likelihood (CML) method to estimate item parameters, and the expected a posteriori (EAP) method (Bock & Aitken, 1981) to estimate home possessions individual scores.

Financial status (finan)

In PIRLS, parents rated the financial wellbeing of their family according to five categories: (1) not at all well-off, (2) not very well-off, (3) average, (4) somewhat well-off, and (5) very well-off. Responses provide a subjective measure of family wealth.

Missing Data

As is common in survey studies, our estimation of SES was hampered by a high rate of missing data (May, 2006). Data on parental education, parental occupational status, and home possessions were not collected in the United States, while France and South Africa lacked financial status variables. We therefore excluded these education systems from the estimation of SES. We also excluded education systems with more than 30% of missing student data in three or more SES items: Scotland (51%), England (50%), Israel (45%), Spain (41%), New Zealand (40%), Qatar (39%), Kuwait (38%), the Netherlands (34%), and Iceland (Grade 5) (33%). The remaining 35 education systems comprised the analytic sample.

However, missing data were still present in the SES items of the analytic sample. If the data were missing completely at random (MCAR), analysis relying on complete cases would have yielded unbiased estimates. But unreported analyses suggested that the data were not MCAR. Instead, it appeared that students with missing data on SES items differed systematically from the rest of the students on observed characteristics. For example, we observed that students with missing data tended not to perform as well as the rest of the student cohort on the PIRLS reading test. Therefore, performing complete-cases analysis or listwise deletion would have led to biased estimates; for example, the average reading performance would have been overestimated.

To account for this source of bias, we decided to replace the missing values in SES items with simulated or plausible values calculated with multiple imputation. Importantly, the purpose of multiple imputation is not to predict missing values as closely as possible to the true values, but to account for missing data uncertainty in SES items and thereby yield statistically valid inferences.

Multiple imputation methods produce unbiased estimates if the data are missing at random (MAR) and even when they are not MCAR. The MAR assumption means that the propensity for missing values on SES items is related to other observed variables, but not to unobserved variables or values of the SES items themselves (Little & Rubin, 2002).

In our case, the MAR assumption implied two possibilities. The first was that the propensity for missing values in each of the SES items could be explained through reference to other variables in the PIRLS dataset, for example, family and student characteristics. The second was that the missing SES data would not be higher for specific values of SES items. Thus, for example, the probability of missing parental education information would not be greater for parents with educational attainment at or above the university degree level after we had controlled for observables. If the

MAR assumption held, multiple imputation techniques would produce valid inferences at the analysis stage. (In most analyses, MAR is regarded as a reasonable assumption, especially if a rich set of covariates is available and can be included in the imputation model, as was the situation in our case.)

We employed data augmentation (DA) for the multiple imputation model. DA is an iterative Markov chain Monte Carlo (MCMC) procedure that assumes a multivariate normal distribution for the data (Gelman, Carlin, Stern, & Rubin, 2004; StataCorp, 2009). The imputation model produced five complete versions of a dataset that included the SES items, reading performance, gender, age, reading attitudes, self-concept, early literacy skills, parental attitudes toward reading, migration background, the sampling weight, and the school cluster indicator, among other variables. The sampling variables and potential reading performance outcome helped us preserve the structural characteristics of the data and so obtain valid inferences in the analysis stage (Rubin, 1996). We then applied Rubin's (1987) rule to the imputed datasets to produce point estimates of item weights, reliability coefficients, and correlation coefficients, as reported in Tables 3 and 4 on pages 20 and 22, respectively.

Principal Component Analysis

We calculated the SES index by applying principal component analysis (PCA) to the six SES constituent items: mother's education, father's education, mother's occupational status, father's occupational status, home possessions, and financial status. The PCA model is consistent with a formative measurement in that the direction of causality goes from the items to the SES index and not vice versa (Diamantopoulos et al., 2008). Put more formally, the SES score for each student *i* is a weighted average of the SES items:

$$SES_i = \alpha_1 momed_i + \alpha_2 daded_i + \alpha_3 momsei_i + \alpha_4 dadsei_i + \alpha_5 hompos_i + \alpha_6 finan_i \quad (1)$$

We calculated item weights in each imputed dataset by applying PCA to the complete sample of countries, and then averaged the resulting weights across imputed datasets to produce point estimates of each item weight. We then used the average weights, α_s, and the item-imputed datasets in equation (1) to calculate five SES scores for each student. As a result, five complete datasets of SES scores were produced. We recommend that model-based estimates using SES should use these five imputed datasets and Rubin's (1987) rule to obtain standard errors that take into account missing data uncertainty.

Figure 1 shows the distribution of the first SES score by education system and for the total international sample. The figure also includes a normal distribution for comparative purposes.

SES VALIDATION

Criteria for assessing the reliability and validity of formative indicators are not well established in the literature, and traditional methods are not appropriate (Diamantopoulos et al., 2008). As we explained earlier, reliability analysis assumes internal consistency among constituent components, but this is not necessarily the case for formative indicators where the derived variable can be a composite of uncorrelated variables. Validity analysis assumes an underlying theoretical model, but the formative model is not necessarily grounded in theory; it could be fully contingent on the operationalization procedure.

Despite the SES concept not being comparable to the established theories of cultural capital or social capital postulated by Bourdieu (1977) and Coleman (1988), it has proved very useful for understanding educational inequalities. From this perspective, it could be said to refer to a less developed stage of theorization. As such, we considered it likely that the SES model would satisfy certain conditions anticipated by the theory and previous empirical evidence:

1. The SES items of parental education, parental occupational status, home possessions, and financial status would show some degree of intercorrelation or internal consistency if they all indicated the relative position of a family within a hierarchical social structure (Mueller & Parcel, 1981);
2. The item weights would not vary substantially across countries if the SES items consistently reflected SES across cultures; and
3. SES would correlate positively with student achievement, as has been extensively reported in the literature (Sirin, 2005; White, 1982).

We used these criteria for our validation of SES (Bollen & Lennox, 1991; Diamantopoulos et al., 2008).

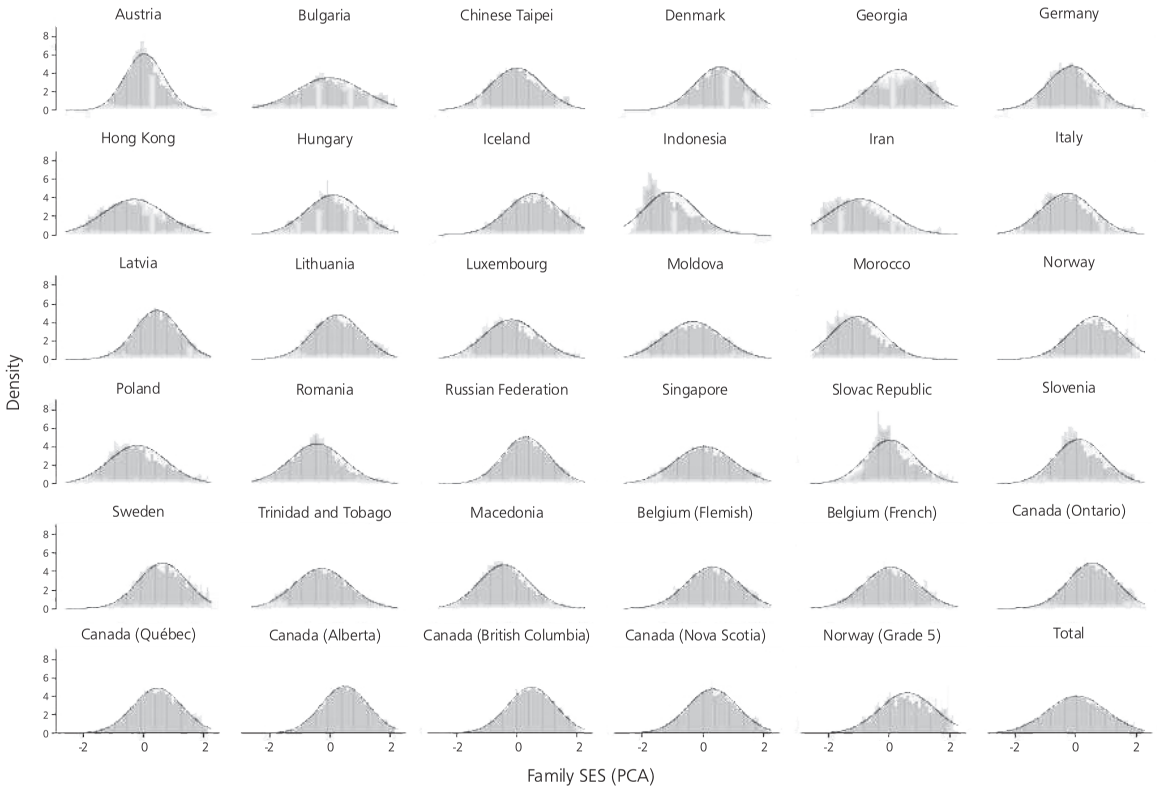
Reliability

We used Cronbach's alpha coefficient, the beta coefficient (minimum lower bound), and the greater lower bound (GLB) coefficient to examine the consistency of the SES items. Most reliability analyses rely exclusively on alpha, probably because it is easy to estimate with available statistical software. However, alpha does not comply with certain theoretical conditions. It tends, for example, to underestimate true reliability or it provides a lower bound estimate of reliability.

Psychometricians have criticized alpha's overuse and have proposed complementary measures of reliability, such as the beta and GLB coefficients (Bentler, 2009; Sijtsma, 2009). These coefficients are also not exempt from statistical flaws, but they provide a broader assessment of reliability by giving a minimum and maximum value of reliability under lower bound conditions (Revelle, 2012).

The beta coefficient is based on split-half reliability and yields a minimum lower bound of the reliability estimate. We could intuitively expect the coefficient to be derived by finding the split halves least related to the SES item data and by using the inter-

Figure 1: Distribution of SES by education system and for the total international sample



Source: Grapghs by 'COUNTRY ID'.

Table 3: Reliability analysis of SES

Country	Beta	Alpha	Glb
Austria	0.17	0.62	0.73
Belgium (Flemish)	0.07	0.70	0.80
Belgium (French)	0.12	0.68	0.79
Bulgaria	0.65	0.86	0.91
Canada (Alberta)	0.22	0.64	0.74
Canada (British Columbia)	0.35	0.65	0.75
Canada (Novia Scotia)	0.20	0.67	0.76
Canada (Ontario)	0.10	0.65	0.75
Canada (Québec)	0.32	0.69	0.78
Chinese Taipei	0.31	0.73	0.80
Denmark	0.41	0.72	0.81
Georgia	0.39	0.72	0.81
Germany	0.20	0.71	0.81
Hong Kong SAR	0.36	0.79	0.85
Hungary	0.40	0.80	0.86
Iceland	0.04	0.68	0.79
Indonesia	0.41	0.74	0.82
Iran	0.50	0.81	0.86
Italy	0.31	0.75	0.83
Latvia	0.48	0.69	0.78
Lithuania	0.48	0.75	0.83
Luxembourg	0.19	0.72	0.81
Macedonia	0.30	0.70	0.82
Moldova	0.39	0.69	0.79
Morocco	0.50	0.67	0.75
Norway	0.14	0.68	0.79
Norway (Grade 5)	0.28	0.72	0.82
Poland	0.54	0.78	0.84
Romania	0.62	0.82	0.88
Russian Federation	0.46	0.72	0.81
Singapore	0.38	0.76	0.83
Slovak Republic	0.53	0.79	0.86
Slovenia	0.13	0.75	0.84
Sweden	0.23	0.69	0.79
Trinidad and Tobago	0.42	0.71	0.78
Total	0.48	0.76	0.84
Median	0.35	0.72	0.81
Min	0.04	0.62	0.73
Max	0.65	0.86	0.91
CV	0.48	0.08	0.05

group intercorrelation to estimate the total variance accounted for by a general factor (Revelle, 2012). Rather than being a particular estimator, the GLB coefficient is more of a theoretical concept that can be approximated with several methods (Sočan, 2000). We used an iterative procedure that allowed us to produce the highest lower-bound reliability measure still consistent with the data (see also R Development Core Team, 2011; Revelle, 2011; Sočan, 2000).

We used the functions *omega* and *guttman* developed by Revelle (2011) in the *Psych package* for the R project to calculate the beta and GLB coefficients. Table 3 reports the derived coefficients for each education system. As with alpha, beta, and GLB, the coefficients ranged from 0 to 1, with the higher values indicating greater reliability.

The median, minimum, and maximum values across education systems can be found at the bottom of the table, together with the coefficient of variation (CV), that is, the standard deviation divided by the mean.

The alpha coefficient of 0.76 for the total sample indicated satisfactory reliability (see Table 3). Alpha values ranged from 0.62 in Austria to 0.86 in Bulgaria, with a median of 0.72 across education systems and a coefficient of variation of 0.08. The beta and GLB coefficients offered complementary information pertaining to overall reliability and dispersion.

The beta coefficient for the total sample was 0.48, while the specific education system beta coefficients ranged from 0.04 in Iceland to 0.65 in Bulgaria, with a median value of 0.35 (see Table 3). The beta coefficients varied substantially more than the alpha, producing a coefficient of variation of 0.48. The greater variability of beta reflects the half-split nature of this coefficient and was driven mostly by the home possession weight. In fact, the correlation of the home possession weight and the beta coefficient across education systems was 0.95.

In education systems where the correlations between the home possessions item and the other SES items were stronger, the home possessions weight and beta coefficient were also higher. In systems where these correlations were weaker, the beta coefficient and the home possessions weight were lower, at times approaching zero. For example, in Bulgaria, the country with the highest beta coefficient, the correlations of home possessions with financial status and with mother's education were 0.34 and 0.43, respectively, and the home possession weight amounted to 0.32. In Iceland, however, the intercorrelations with home possessions and other items were almost zero, leading to a beta coefficient and home possession weight of about zero (see Table 3 above and also Table 4 below).

GLB coefficients ranged from 0.73 in Austria to 0.91 in Bulgaria, with a median of 0.81 and a coefficient of variation of 0.05 across education systems (see Table 3). The GLB coefficients exhibited the lowest variability according to the coefficient of variation, an outcome that probably reflected the greater stability and accuracy of this estimator (Sočan, 2000). We anticipate that other reliability coefficients in the literature will yield coefficient values that range between 0.48 (beta) and 0.84 (GLB) for the complete international sample (see Table 3).

Validation

We carried out this process separately for the SES items and the total SES scores. We evaluated item validation by considering the variation of item weights across education systems, and evaluated score validation by examining the correlation with reading performance, here the benchmark variable (Bollen & Lennox, 1991; Diamantopoulos et al., 2008).

Individual Items

Table 4 reports weights for the SES items by education system. The median for all education systems as well as the minimum, maximum, and coefficient of variation are also reported. Here, we could expect that weights would all be positive and relatively stable across education systems if they served to indicate SES equally in different societies.

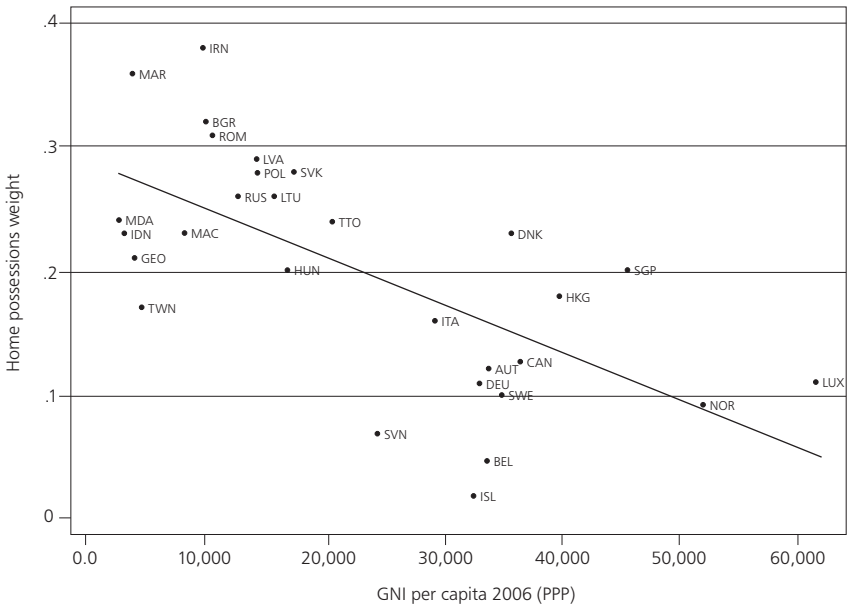
Table 4: SES item weights and correlation with reading achievement

Country	Item weights						
	Parental education		*Parental occupational status*		*Home possessions*	*Financial status*	*Rho (SES, reading)*
	Father	Mother	Father	Mother			
Austria	0.48	0.50	0.50	0.50	0.12	0.08	0.36
Belgium (Flemish)	0.50	0.50	0.47	0.46	0.03	0.24	0.41
Belgium (French)	0.48	0.50	0.48	0.48	0.07	0.22	0.42
Bulgaria	0.46	0.46	0.43	0.45	0.32	0.29	0.42
Canada (Alberta)	0.47	0.49	0.46	0.42	0.13	0.35	0.28
Canada (British Columbia)	0.49	0.49	0.47	0.44	0.14	0.29	0.26
Canada (Novia Scotia)	0.47	0.47	0.46	0.45	0.12	0.36	0.31
Canada (Ontario)	0.47	0.50	0.46	0.45	0.06	0.34	0.24
Canada (Québec)	0.48	0.47	0.45	0.45	0.19	0.32	0.33
Chinese Taipei	0.49	0.49	0.48	0.44	0.17	0.25	0.39
Denmark	0.46	0.47	0.46	0.46	0.23	0.29	0.35
Georgia	0.51	0.50	0.47	0.43	0.21	0.20	0.35
Germany	0.51	0.49	0.48	0.44	0.11	0.24	0.42
Hong Kong SAR	0.47	0.47	0.46	0.42	0.18	0.36	0.15
Hungary	0.47	0.48	0.46	0.47	0.20	0.27	0.49
Iceland	0.48	0.50	0.46	0.49	0.02	0.25	0.31
Indonesia	0.49	0.50	0.47	0.41	0.23	0.27	0.34
Iran	0.48	0.49	0.43	0.37	0.38	0.26	0.51
Italy	0.47	0.49	0.47	0.45	0.16	0.29	0.30
Latvia	0.43	0.47	0.46	0.46	0.28	0.31	0.32
Lithuania	0.47	0.48	0.45	0.47	0.26	0.24	0.40
Luxembourg	0.50	0.49	0.48	0.45	0.11	0.26	0.40
Macedonia	0.50	0.51	0.45	0.48	0.23	0.05	0.46
Moldova	0.50	0.51	0.42	0.46	0.24	0.19	0.30
Morocco	0.52	0.51	0.42	0.29	0.36	0.30	0.23
Norway	0.49	0.51	0.48	0.48	0.08	0.20	0.36
Norway (Grade 5)	0.48	0.50	0.48	0.48	0.17	0.16	0.39
Poland	0.47	0.47	0.45	0.43	0.28	0.31	0.40
Romania	0.46	0.46	0.45	0.44	0.31	0.28	0.48
Russian Federation	0.47	0.48	0.46	0.45	0.26	0.26	0.35
Singapore	0.49	0.49	0.47	0.44	0.20	0.27	0.44
Slovak Republic	0.46	0.46	0.47	0.46	0.28	0.27	0.49
Slovenia	0.46	0.47	0.47	0.47	0.07	0.35	0.40
Sweden	0.49	0.49	0.46	0.48	0.11	0.25	0.36
Trinidad and Tobago	0.48	0.48	0.46	0.44	0.24	0.28	0.41
Total	0.48	0.49	0.44	0.45	0.26	0.26	0.45
Median	0.48	0.49	0.46	0.45	0.19	0.27	0.36
Min	0.43	0.46	0.42	0.29	0.02	0.05	0.15
Max	0.52	0.51	0.50	0.50	0.38	0.36	0.51
CV	0.04	0.03	0.04	0.08	0.50	0.26	0.22

Parental education exerted the greatest load on the SES index, with median weights of 0.48 for father's education and 0.49 for mother's education, followed by parental occupational status, with median weights of 0.46 for father's occupational status and 0.45 for mother's occupational status. The lowest weights corresponded to financial status and home possessions, with median values of 0.27 and 0.19, respectively. The weights of these latter items also exhibited the greatest variation across education systems, with coefficients of variation of 0.26 and 0.50, respectively (see Table 4).

Variation of the home possessions weight was particularly high. We found that it related to the country's income per capita. Figure 2 depicts the association between the home possessions weights and the country's gross national income per capita (PPP) in 2006, as reported by the World Bank (2007). The correlation between income per capita and home possessions weights amounted to -0.66, indicating the greater importance that home possessions has with respect to SES in lower income per capita societies. Iran, for example, had a relatively low income per capita and a home possessions weight of 0.38, whereas Luxembourg had a much higher income per capita; there, the weight was 0.11 (see Figure 2).

Figure 2: Home possessions weights and national income per capita



Note:

Country abbreviations: AUT = Austria, BEL = Belgium, BGR = Bulgaria, CAN = Canada, DEU = Germany, DNK = Denmark, GEO = Georgia, HKG = Hong Kong SAR, HUN = Hungary, IDN = Indonesia, IRN = Iran, ISL = Iceland, ITA = Italy, LVA = Latvia, LTU = Lithuania, LUX = Luxembourg, MAC = Macedonia, MAR = Morocco, MDA = Moldova, NOR = Norway, POL = Poland, ROM = Romania, RUS = Russian Federation, SGP = Singapore, SVK = Slovak Republic, SVN = Slovenia, SWE = Sweden, TTO = Trinidad and Tobago, TWN = Chinese Taipei.

This result can also be partly explained by the lower variability in the home possessions item in wealthier societies. In unreported analyses, we found that the mean of the home possessions item increased and the variance reduced significantly as national income per capita levels increased. The home possessions item is thus a less important indicator of SES in wealthier societies because most families tend to have all the possessions surveyed, whereas in poorer countries these possessions still serve to distinguish between higher and lower SES families.

The correlations between the remaining SES items and gross national income per capita were lower and positive for all other SES items: 0.08 for financial status, 0.04 for both father's education and mother's education, 0.55 for father's occupational status, and 0.35 for mother's occupational status.

Correlation with Reading Achievement

We expected to find a positive correlation between SES and academic achievement. Meta-analyses in the United States indicate a correlation of around 0.30 (Sirin, 2005; White, 1982). Table 4 above shows the correlation between SES and reading achievement in PIRLS 2006 by education system. Correlations ranged from 0.15 in Hong Kong to 0.51 in Iran, with a median of 0.36 for the international sample. These numbers agree with data in studies from the United States, which show a medium level of association between SES and academic achievement. However, a direct comparison was not possible because of a lack of socioeconomic data in the United States.

As with other association statistics, international comparisons of correlation coefficients are restricted by crossnational comparability limitations and therefore need to be interpreted with considerable caution. For example, the correlation between SES and reading achievement yielded the same value of 0.42 in Germany and Bulgaria, but the weight of home possessions was stronger in Bulgaria than in Germany, with values of 0.32 and 0.11, respectively (see Table 4). Because the international SES index uses the complete sample home possessions weight of 0.26, it tended to underestimate the importance of home possessions in Bulgaria and overestimate it in Germany.

Comparability of SES across nations can also be affected by unobserved quality differences in SES. For example, parental education levels capture quantitative differences in education but may not capture differences in the quality of education. And two students in two different countries with comparable SES levels may actually have very different SES if the quality of education attained by their parents differs by country.

DISCUSSION

The procedure that we employed to measure and validate SES used PIRLS data for illustrative purposes. We consider that the procedure can serve as a practical guide for educational researchers seeking to construct single SES indexes and study educational inequalities related to family background when using data from national and international student assessment studies.

The proposed SES index is a composite of variables reflecting parental education, parental occupational status, and family wealth, and it follows traditional operationalizations of the SES concept (Buchmann, 2002; Gottfried, 1985; Hauser, 1994; Mueller & Parcel, 1981). Specifically, the SES index is a composite of six SES items: mother's education, father's education, mother's occupational status, father's occupational status, home possessions, and financial status, with the last two items expected to capture the family wealth component.

Other operationalizations of SES use the maximum educational level and occupational status of either parent to reduce the amount of missing information in SES (e.g., OECD, 2009). However, we used multiple imputation techniques and contextual data to produce five complete versions of the SES item data. The SES model based on six separate items yielded greater reliability than the one with four items using the maximum strategy (alpha = 0.76 vs. 0.63), produced a greater number of possible item combinations and therefore also of SES scores, and achieved a greater balance between concepts and variables, with each SES concept (i.e., education, occupation, and wealth) being measured by two items.

The presented reliability analysis provided a broader evaluation of internal consistency than the traditional alpha coefficient. The alpha coefficient of 0.76 indicated the overall satisfactory internal consistency for the total sample, and the beta and GLB coefficients indicated that the reliability of the lower bound coefficients lay between 0.48 and 0.84. The reliability analysis also suggested that relying on a single coefficient can be misleading. For example, if we had exclusively based our conclusions on alpha, we would have found no substantial difference in terms of SES data consistency between Québec's education system (alpha = 0.69) and Iceland's (alpha = 0.68). However, we would have neglected critical differences captured by the beta coefficient (beta = 0.32 and 0.04 in Québec and in Iceland, respectively). Differences in reliability captured by beta, but not by alpha coefficients are likely explained by the split-half nature of the beta coefficient.

The beta coefficient was severely affected by the home possessions weight or the intercorrelation of home possessions and other SES items. In education systems where intercorrelations are low, the beta coefficient will also tend to be low, and vice versa, which is why the correlation of the home possessions weight and the beta coefficient across the education systems was almost one. In Iceland, intercorrelations with the home possessions items were almost zero, which meant the beta coefficient was almost zero as well, whereas in Québec, home possessions still indicated SES with a weight of 0.19. Thus, in addition to the alpha coefficient, the beta coefficient suggests that SES reliability in Iceland was lower than in Québec, and particularly so if we assume that items should weigh more or less equally in the calculation of SES.

In general, the lower bound reliability coefficients indicated overall satisfactory internal consistency for the total sample (beta = 0.48, alpha = 0.76, GLB = 0.84). However, the SES items were not equally indicative of SES across nations. For example, we found a substantial positive correlation between the parental occupational status items

and national income per capita. That the weight of the mother's and the father's occupational status items was higher in wealthier nations likely reflects the greater income and more varied occupations of fathers and mothers in wealthier societies. Mothers, especially, tend to be more educated and employed in different sectors of the economy in wealthier than in poorer societies.

Although parental occupational status was a better indicator of SES in the wealthier societies, it still served to indicate SES in poorer societies. In contrast, the home possessions component did not seem to reflect SES in all nations. The home possessions weight was the least stable across nations and was strongly and negatively correlated with the national income per capita. It seems that the home possessions items surveyed in the study were important indicators of social mobility in the poorer societies, but that their importance decreased as income per capita increased, eventually becoming negligible for the wealthiest countries participating in PIRLS.

Limitations and Implications for Future Research and Survey Development

Probably the most important limitation is the theoretical basis of the proposed SES index. There is no consensus on the conceptual meaning of SES, which prevents researchers from understanding the mechanisms underlying the association between SES and student outcomes. And although most studies regard SES as a formative indicator, they often also use different variables to measure it, thereby violating the operationalization condition of formative measurement and implicitly suggesting that SES does reflect a well-established concept.

We think it likely that researchers intuitively invoke well-developed concepts of human capital, cultural capital, and economic capital to justify the calculation of a single SES index, but a comprehensive SES theory has yet to be elaborated. It seems, therefore, that further work is needed on the theoretical foundations of SES. A related conjecture is that once a formal definition of SES is developed, construct measurement and validity could switch from the current formative model to a reflective model. Confirmatory factor analysis and crosscultural invariance could then be applied for validity assessment, for example. In this current study, we validated the SES index by analyzing internal consistency, the correlation with reading achievement, and item weights across education systems, but these analyses were insufficient for proper validity assessment.

Another limitation for SES validation is the substantial variability of the home possessions weight across education systems. Further research should study the possible reasons as to why the lowest weight was that for home possessions in wealthier societies. One plausible explanation supported by the data is that students in wealthier societies have greater access to and low variability in the home possessions surveyed (i.e., computer, study desk, daily newspaper, own room, and own mobile phone), whereas these items still distinguish higher-SES and lower-SES families in poorer countries. But another plausible explanation is that, irrespective of item selection, home possessions play a less important role in SES for wealthier societies.

Research identifying home possessions items that reflect SES in both poorer and wealthier societies would contribute to survey development and improve the crossnational validity of the SES index. In this regard, May (2006) provides an interesting methodological proposal using national-specific items in TIMSS 1995 and also available for PIRLS. Brese and Mirazchiyski (2010), however, found a negligible association between the national-specific home possessions items and academic achievement. Also, a review of the empirical evidence and theories in the international literature could help us determine if this pattern can be explained mostly by data restrictions or by substantive reasons. Whatever the case, research in this area could contribute importantly to survey development and the study of educational inequalities.

Still another limitation is the presence of missing data in the SES items. We employed multiple imputation to counteract this potential source of bias, but the quality of the derived SES index might still be affected by the substantial amount of missing data. It is also possible that multiple imputation would not yield unbiased estimates if the MAR assumption were not satisfied, that is, if the data were not missing at random (NMAR). This situation would occur if the missing data mechanism depended not only on the observed information but also on unobserved variables. And although the MAR assumption is usually realistic when a large set of variables is available, as with PIRLS, in some cases it can be violated. This situation is evident, for example, in longitudinal studies where lower-SES participants are more likely to leave the study earlier. Another example relates to collecting annual income data, wherein families are less likely to report their income once a certain income level is surpassed. In general, though, it is almost impossible to determine the missing data mechanism, the MAR assumption is not testable, and the possibility of a NMAR mechanism can never be completely ruled out.

These situations are why we recommend further research that uses sensitivity analysis to address the possibility that the data are NMAR—where the "missingness" depends on unobservables (Resseguier, Giorgi, & Paoletti, 2011). Sensitivity analysis identifies groups according to missing-data patterns and then evaluates whether critical estimates vary across groups. The results help establish a range of conditions under which derived estimates are unbiased as well as conditions under which estimates are biased. Another alternative is to use techniques that directly allow using data without the need to impute missing responses, or directly incorporate the mechanisms for missing data into the model (Rose, von Davier, & Xu, 2010).

Yet another limitation relates to how parental occupational data were collected during PIRLS. A multiple-choice question was used to ask parents about their occupations. However, research shows that the categorical approach to social stratification has several disadvantages in relation to a continuous or hierarchical scheme achieved with an open-ended question (Ganzeboom et al., 1992). We referenced the ISEI developed by Ganzeboom et al. (1992) to assign ISE scores to the 11 job categories in PIRLS. However, the reduced number of categories in PIRLS limited the variability and validity that the ISEI offers. A broader set of options grounded in theory would therefore be

needed to establish a more accurate scoring of occupations. Survey developers should therefore consider existing theories in this area, their proposed survey questions, and the missing data rates in previous studies (Buchmann, 2002).

Researchers and survey developers should also carefully consider the consequences of collecting socioeconomic data from parents and students. Theoretically, parents are the ultimate authority for reporting SES, while students, especially in the early grades, tend to lack a precise knowledge of socioeconomic home resources and are more likely to provide less reliable SES information (Kreuter, Eckman, Maaz, & Watermann, 2010; Sirin, 2005). In practice, however, the implementation of parent questionnaires or the application of specific items is not always possible (Schulz, 2005). Therefore, it is important to know how unreliable student responses are with respect to the different SES concepts.

Research shows that unreliability of student responses is lower for home possessions, as utilized in PIRLS 2006, than for parental occupations and for parental educational levels, in that order (Brese & Mirazchiyski, 2010; Schulz, 2005; Yang & Gustafsson, 2004). We therefore expect that our SES index will be less affected by unreliable responses than other measures that draw on student reports of parental education or parental occupations. Research also shows that the reliability of socioeconomic information provided by students varies by country and increases with age, family SES, and academic achievement levels (Brese & Mirazchiyski, 2010; Kreuter et al., 2010). Socioeconomic information provided by parents can also be somewhat biased by social desirability and high levels of non-response (Schulz, 2005). In any case, survey developers and researchers should be aware of these caveats when collecting and analyzing SES information.

A final limitation regarding the application of this approach to other IEA studies is that studies unevenly collect socioeconomic data. For example, PIRLS and the International Civic and Citizenship Education Study (ICCS) gather more information on socioeconomic background than does TIMSS, where no parental occupation or financial status data are available (Buchmann, 2002; Foy & Kennedy, 2008). Therefore, our presented SES estimation procedure cannot be directly generalized to other IEA studies. If the goal is to adopt the SES approach, survey developers should consider defining a set of socioeconomic variables that can be collected evenly across studies. However, further research and discussion are required before advocating this approach. Our aim in this paper has been to contribute to this direction.

Acknowledgement: The authors are very grateful to two anonymous reviewers as well as Dirk Hastedt, Robert Whitwell, and Aurora Cortés for helpful comments and suggestions on previous drafts of this paper.

References

Alexander, K. L., Entwisle, D. R., & Olson, L. S. (2007). Lasting consequences of the summer learning gap. *American Sociological Review*, *72*, 167–180.

Bentler, P. M. (2009). Alpha, dimension-free, and model-based internal consistency reliability. *Psychometrika*, *74*(1), 137–143.

Bock, R. D., & Aitken, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, *46*(4), 443–459.

Bollen, K. A. (1989). *Structural equations with latent variable*s (Wiley Series in Probability and Mathematical Statistics). New York, NY: Wiley.

Bollen, K. A., & Lennox, R. (1991). Conventional wisdom on measurement: A structural equation perspective. *Psychological Bulletin*, *110*(2), 305–314.

Bornstein, M. H., & Bradley, R. H. (Eds.). (2003). *Socioeconomic status, parenting, and child development*. Mahwah, NJ: Lawrence Erlbaum Associates.

Borsboom, D., Mellenbergh, G., & van Heerden, J. (2003). The theoretical status of latent variables. *Psychological Review*, *110*(2), 203–219.

Bourdieu, P. (1977). Cultural reproduction and social reproduction. In J. Karabel & A. H. Halsey (Eds.), *Power and ideology in education* (pp. 487–511). New York, NY: Oxford University Press.

Bourdieu, P. (1986). The forms of capital. In J. G. Richardson (Ed.), *Handbook of theory and research for the sociology of education* (pp. 241–258). New York, NY: Greenwood Press.

Brese, B., & Mirazchiyski, P. (2010, July). *Measuring students' family background in large-scale education studies*. Paper presented at the fourth IEA International Research Conference in Gothenburg, Sweden.

Buchmann, C. (2002). Measuring family background in international studies of education: Conceptual issues and methodological challenges. In A. C. Porter & A. Gamoran (Eds.), *Methodological advances in cross-national surveys of educational achievement* (pp. 150–197). Washington, DC: National Academy Press.

Caro, D. H., & Lehmann, R. (2009). Achievement inequalities in Hamburg schools: How do they change as students get older? *School Effectiveness and School Improvement*, *20*(4), 407–431.

Caro, D. H., & Lenkeit, J. (2012). An analytical approach to study educational inequalities: Ten hypothesis tests in PIRLS 2006. *International Journal of Research and Method in Education*, *35*(1), 3–30.

Caro, D. H., Lenkeit, J., Lehmann, R., & Schwippert, K. (2009). The role of academic achievement growth in school track recommendations. *Studies in Educational Evaluation*, *35*(4), 183–192.

Caro, D. H., McDonald, T., & Willms, J. D. (2009). Socio-economic status and academic achievement trajectories from childhood to adolescence. *Canadian Journal of Education*, *32*(3), 558–590.

Caro, D. H., & Mirazchiyski, P. (2012). Socioeconomic gradients in Eastern European countries: Evidence from PIRLS 2006. *European Educational Research Journal*, *11*(1), 96–110.

Chao, R. K., & Willms, J. D. (2002). The effects of parenting practices on children's outcomes. In J. D. Willms (Ed.), *Vulnerable children: Findings from Canada's national longitudinal survey of children and youth* (pp. 149–166). Edmonton, Alberta, Canada: University of Alberta Press.

Coleman, J. S. (1988). Social capital in the creation of human capital. *American Journal of Sociology*, *94*(Supplement), S95–S120.

Condron, J. (2007). Stratification and educational sorting: Explaining ascriptive inequalities in early childhood reading group placement. *Social Problems*, *54*(1), 139–160.

Deaton, A. (2002). Policy implications of the gradient of health and wealth. *Health Affairs*, *21*(2), 13–30.

Diamantopoulos, A., Riefler, R., & Roth, K. P. (2008). Advancing formative measurement models. *Journal of Business Research*, *61*(12), 1203–1218.

Edwards, J. R., & Bagozzi, R. P. (2000). On the nature and direction of relationships between constructs and measures. *Psychological Methods*, *5*(2), 155–174.

Foy, P., & Kennedy, A. M. (Eds.). (2008). *PIRLS 2006 user guide Supplement 3 for the international database*. Chestnut Hill, MA: Boston College.

Ganzeboom, H. B. G., de Graaf, P. M., & Treiman, D. J. (1992). A standard international socio-economic index of occupational status. S*ocial Science Research*, *21*, 1–56.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2004). *Bayesian data analysis* (2nd ed.). London, UK: Chapman & Hall.

Goldthorpe, J. H. (1980). *Social mobility and class structure in modern Britain*. Oxford, UK: Clarendon Press.

Goldthorpe, J. H., Payne, C., & Llewellyn, C. (1978). *Trends in class mobility. Sociology*, *12*(3), 441–468.

Gottfried, A. (1985). Measures of socioeconomic status in child development research: Data and recommendations. *Merrill-Palmer Quarterly*, *31*(1), 85–92.

Guo, G., & Harris, K. (2000). The mechanisms mediating the effects of poverty on children's intellectual development. *Demography*, *37*(4), 431–447.

Hauser, R. M. (1994). Measuring socioeconomic status in studies of child development. *Child Development*, *65*(6), 1541–1545.

Heath, A. F., & Clifford, P. (1990). Class inequalities in education in the twentieth century. *Journal of the Royal Statistical Society: Series A*, *153*(1), 1–16.

Jarvis C. B., Mackenzie, S. B., & Podsakoff, P. M. (2003). A critical review of construct indicators and measurement model misspecification in marketing and consumer research. *Journal of Consumer Research*, *30*(3), 199–218.

Kerckhoff, A. C. (1993). *Diverging pathways: Social structure and career deflections*. Cambridge, UK/New York, NY: Cambridge University Press.

Kerckhoff, A., Raudenbush, S., & Glennie, E. (2001). Education, cognitive skill, and labor force outcomes. *Sociology of Education*, *74*(1), 1–24.

Kreuter, F., Eckman, S., Maaz, K., & Watermann, R. (2010). Children's reports of parents' education level: Does it matter whom you ask and what you ask about? *Survey Research Methods*, *4*(3), 127–138.

Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data* (2nd ed.). Hoboken, NJ: Wiley.

Ma, X. (2000). Socioeconomic gaps in academic achievement within schools: Are they consistent across subject areas? *Educational Research and Evaluation*, *6*(4), 337–355.

Maaz, K., Trautwein, U., Lüdtke, O., & Baumert, J. (2008). Educational transitions and differential learning environments: How explicit between-school tracking contributes to social inequality in educational outcomes. *Child Development Perspectives*, *2*(2), 99–106.

May, H. (2006). A multilevel Bayesian item response theory method for scaling. *Journal of Educational and Behavioral Statistics*, *31*(1), 63–79.

Mueller, C. W., & Parcel, T. L. (1981). Measures of socioeconomic status: Alternatives and recommendations. *Child Development*, *52*(1), 13–30.

Mullis, I. V. S., Martin, M. O., Kennedy, A. M., & Foy, P. (2007). *PIRLS 2006 international report: IEA's Progress in International Reading Literacy Study in primary schools in 40 countries*. Chestnut Hill, MA: Boston College.

Oakes, J. (1985). *Keeping track: How schools structure inequality*. New Haven, CT: Yale University Press.

Organisation for Economic Co-operation and Development (OECD). (2003). *Literacy skills for the world of tomorrow: Further results from PISA 2000*. Paris, France: Author.

Organisation for Economic Co-operation and Development (OECD). (2004). *Learning for tomorrow's world: First results from PISA 2003*. Paris, France: Author.

Organisation for Economic Co-operation and Development (OECD). (2007). *PISA 2006: Science competencies for tomorrow's world*. Paris, France: Author.

Organisation for Economic Co-operation and Development (OECD). (2009). *PISA 2006 technical report*. Paris, France: Author.

Organisation for Economic Co-operation and Development (OECD). (2010). *PISA 2009 results: Overcoming social background. Equity in learning opportunities and outcomes* (Vol. II). Paris, France: Author. Available online at http://dx.doi.org/10.1787/9789264091504-en

R Development Core Team. (2011). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Available online at http://www.R-project.org/

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen, Denmark: Nielsen and Lydiche.

Raudenbush, S. W., & Kasim, R. M. (1998). Cognitive skill and economic inequality: Findings from the National Adult Literacy Study. *Harvard Educational Review*, *68*(1), 33–79.

Resseguier, N., Giorgi R., & Paoletti, X. (2011). Sensitivity analysis when data are missing not-at-random. *Epidemiology*, *22*(2), 282.

Revelle, W. (2011). *Procedures for personality and psychological research*. Evanston, IL: Northwestern University. Available online at http://personality-project.org/r/psych.manual.pdf.1.0-97

Revelle, W. (2012, Spring). *An introduction to psychometric theory with applications in R*. Evanston, IL: Northwestern University. Available online at http://personality-project.org/r/book/

Rose N., von Davier, M., & Xu, X. (2010). *Modeling nonignorable missing data with item response theory (IRT) (ETS Research Report No. RR-10-11)*. Princeton, NJ: Educational Testing Service.

Rubin, D. B. (1987). *Multiple imputation for non-response in surveys*. New York, NY: John Wiley & Sons.

Rubin, D. B. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association*, *91*(434), 473–489.

Rumberger, R. W. (2010). Education and the reproduction of economic inequality in the United States: An empirical investigation. *Economics of Education Review*, *29*(2), 246–254.

Rutter, M., & Maughan, B. (2002). School effectiveness findings 1979–2002. *Journal of School Psychology*, *40*(6), 451–475.

Rutter, M., Maughan, B., Mortimore, P., & Ouston, J. (1979). *Fifteen thousand hours: Secondary schools and their effects on children*. Cambridge, MA: Harvard University Press.

Schnabel, K., Alfeld, C., Eccles, J., Köller, O., & Baumert, J. (2002). Parental influence on students' educational choices in the United States and Germany: Different ramifications—same effect? *Journal of Vocational Behavior*, *60*(2), 178–198.

Schulz, W. (2005, April). *Measuring the socio-economic background of students and its effect on achievement in PISA 2000 and PISA 2003*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Sijtsma, K. (2009). On the use, misuse, and the very limited usefulness of Cronbach's Alpha. *Psychometrika*, *74*(1), 107–120.

Sirin, S. R. (2005). Socioeconomic status and academic achievement: A meta-analytic review of research. Review of Educational Research, 75(3), 417–453.

Sočan, G. (2000). Assessment of reliability when test items are not essentially τ-equivalent. *Developments in Survey Methodology*, *15*, 23–35. Available online at http://mrvar.fdv.uni-lj.si/pub/mz/mz15/socan.pdf

StataCorp. (2009). *Stata 11 multiple-imputation reference manual*. College Station, TX: Stata Press.

The World Bank. (2007). *GNI per capita, PPP (current international $)*. Washington, DC: Author. Retrieved from http://data.worldbank.org/indicator/NY.GNP.PCAP.PP.CD?page=1

United Nations Development Program (UNDP). (2006). *Human development report*. New York, NY: Palgrave.

United Nations Educational, Scientific and Cultural Organization (UNESCO). (1999). *International Standard Classification of Education–ISCED*. Paris, France: Author.

White, K. R. (1982). The relation between socioeconomic status and academic achievement. *Psychological Bulletin*, *91*(3), 461–481.

Willms, J. D. (Ed.). (2002). V*ulnerable children: Findings from Canada's National Longitudinal Survey of Children and Youth*. Edmonton, Alberta, Canada: University of Alberta Press.

Willms, J. D. (2003). *Ten hypotheses about socioeconomic gradients and community differences in children's developmental outcomes*. Ottawa, Ontario, Canada: Applied Research Branch of Human Resources Development Canada.

Willms, J. D. (2006a). *Learning divides: Ten policy questions about the performance and equity of schools and schooling systems*. Montreal, Québec, Canada: UNESCO Institute for Statistics.

Willms, J. D. (2006b). Variation in socioeconomic gradients among cantons in French- and Italian-speaking Switzerland: Findings from the OECD PISA. *Educational Research and Evaluation*, *12*(2), 129–154.

Willms, J. D., & Raudenbush, S. W. (1989). A longitudinal hierarchical linear model for estimating school effects and their stability. *Journal of Educational Measurement*, *26*(3), 209–232.

Willms, J. D., Smith, T. M., Zhang, Y., & Tramonte, L. (2006). Raising and levelling the learning bar in Central and Eastern Europe. *Prospects*, *36*(4), 411–418.

Willms, J. D., & Somers, M.-A. (2001). Family, classroom, and school effects on children's educational outcomes in Latin America. *Journal of School Effectiveness and School Improvement*, *12*(4), 409–445.

Yang, Y., & Gustafsson, J.-E. (2004). Measuring socioeconomic status at individual and collective levels. *Educational Research and Evaluation*, *10*(3), 259–288.

Yeung, W. J., Linver, M. R., & Brooks-Gunn, J. (2002). How money matters for young children's development: Parental investment and family processes. *Child Development*, *73*, 1861–1879.

Estimating linking error in PIRLS

Michael O. Martin, Ina V. S. Mullis, Pierre Foy, Bradley Brossman, and Gabrielle M. Stanco

Boston College, Chestnut Hill, Massachusetts, United States

The Trends in Mathematics and Science Study (TIMSS) and Progress in Reading Literacy Study (PIRLS), as well as other large-scale assessments, measure changes in student achievement over time by linking one assessment to the next. Linking error is conceptualized as the result of changing the pool of items used to measure achievement as well as shifts in the measurement properties of the common items from one assessment cycle to the next. The estimation of the scale-linking transformation is, as with any statistical approximation, susceptible to estimation error. This study describes the method used to estimate linking error for the TIMSS and PIRLS assessments and examines the magnitude of linking error between the PIRLS 2001 and 2006 assessments. As anticipated, linking error was small and had little impact on significance tests of achievement differences between the two assessments, most likely because almost half the items were common to both PIRLS assessments.

IERI Monograph Series: Issues and Methodologies in Large-Scale Assessments Volume 5

INTRODUCTION

Policymakers have become increasingly interested in student achievement trends that provide information about changing patterns of student achievement and enable them to monitor the results of educational reforms over time. To measure changes in student achievement, trend assessments operate on a regular cycle, administering a pool of achievement items to comparable samples of students every three to five years. IEA's Trends in International Mathematics and Science Study (TIMSS) has been measuring trends for some 70 participating countries every four years since 1995 (i.e., 1999, 2003, 2007, and 2011). IEA's Progress in International Reading Literacy Study (PIRLS) is conducted every five years, with assessments taking place in 2001, 2006, and 2011.

Because of the necessity for public disclosure and the need to ensure the current relevance of each new assessment, a selection of items from the assessment is released after each assessment cycle and replaced by newly developed items in the next. To maintain comparability across cycles, however, it is also necessary to have a substantial number of items that are not released and that are included in adjacent cycles. With the common items from successive assessments used as the basis for linking, scores from each new assessment cycle are then placed on the existing achievement scale from previous assessment cycles, so allowing differences from one assessment cycle to the next to be measured. The fact that some items are released and replaced by others means that the assessment changes somewhat from cycle to cycle, which introduces linking error.[1]

In large-scale assessments such as TIMSS and PIRLS, linking error is the error associated with placing achievement data from the most recent assessment cycle on a preexisting trend scale. As such, this error has two major sources:

1. Changes in the pool of items used to measure achievement as previously used items are released and replaced by new items; and
2. Shifts in the measurement properties of the common items from one assessment cycle to the next.

An effective linking method ensures that linking error is kept to a minimum while providing an estimate of the magnitude of the error and its impact on estimates of change in student achievement from cycle to cycle. In TIMSS and PIRLS, new replacement items are developed according to the same assessment frameworks as the released items, thereby ensuring that they have a similar focus and subject-matter coverage. In particular, the new items address mathematics and science content and cognitive domains that are the same as those for TIMSS and reading purposes and comprehension processes that are the same as those for PIRLS. Furthermore, in order to provide sufficient number of common items to maintain a stable link across assessment cycles, the TIMSS and PIRLS assessment designs currently specify that each assessment

1 Note that linking error applies only when comparing achievement results across cycles. Linking error is not an issue when making comparisons within the same assessment cycle.

shares 60% of its items with the next assessment cycle. For the TIMSS eighth-grade assessments, this specification resulted in 126 items common to the 2007 and 2011 mathematics assessments, and 125 items common to the science assessments.

Due to the increased visibility of and reliance on trend data, examining the precision of estimates of trends in student achievement and the error associated with these estimates has become an important area of research. This study describes the TIMSS and PIRLS approach to measuring student achievement trends, and presents a method for estimating linking error based on this approach. The method is then applied to estimate linking error between the PIRLS 2001 and 2006 assessments.

IRT APPROACHES TO MEASURING TRENDS

Large-scale assessments of student achievement, such as the National Assessment of Educational Progress (NAEP) in the U.S. and the TIMSS, PIRLS, and Program in International Student Assessment (PISA) international assessments, rely on item response theory (IRT) methods to construct achievement scales for reporting student achievement and measuring trends from assessment cycle to assessment cycle. IRT methods are valuable in this context because they provide a way to estimate achievement in a student population based on the measurement properties of the individual items comprising the assessment. The item properties, or item parameters, are not known in advance, but must first be estimated from the assessment data through a process known as item calibration.

On completion of the item calibration, the item parameters are used to produce estimates of student achievement, which in large-scale assessments are typically in the form of "plausible values." These are estimates of student performance on the entire assessment, conditional on the responses the students gave to the assessment items they were administered and on the students' background characteristics (Foy, Galia, & Li, 2007).

Typically, one of two methods is used to calibrate item parameters within the IRT framework. In the separate calibration method, assessment data from adjacent cycles are calibrated separately (Kolen & Brennan, 2004). That is, the assessment data for the previous cycle are calibrated first, after which the assessment data for the current cycle are calibrated. A scale linking method is then used to place parameter estimates from the two calibrations on the same scale.

In contrast to the separate calibration method, the concurrent calibration method uses all data from both the previous cycle and the current cycle to estimate item and person parameters at the same time (Kolen & Brennan, 2004). Given that all parameters are estimated at the same time and therefore items common to both cycles receive the same estimates, item parameters from both cycles are on the same scale when the concurrent calibration method is applied. An advantage of the concurrent calibration approach is that it makes maximum use of all available data to estimate the item parameters. Also, by recalibrating the item parameters for common items at each assessment cycle, it permits these parameters to evolve gradually across successive cycles as circumstances change.

After completion of the item calibration, student achievement is estimated using the newly calibrated item parameters. In large-scale assessments such as TIMSS and PIRLS, this procedure involves conducting a principal components analysis (PCA) using background variables and then developing a regression equation using both the principal component variables and the item responses to estimate plausible values for each student—a process known as "conditioning." Student achievement is estimated by using each of the sets of plausible values; variation across the sets of plausible values reflects the measurement error (Foy et al., 2007).

Once student achievement has been estimated, the scale-linking transformation places the current-cycle data onto the previously existing trend scale. In large-scale trend studies, the scale-linking transformation is typically determined by matching characteristics of the achievement distribution of the data from the previous cycle (obtained using the new item calibration) to characteristics of the achievement distribution of the same data on the existing trend scale (obtained using the previous item calibration). After the linear transformation that best matches these two distributions has been determined, the transformation is applied to the current cycle data, so allowing these data to be placed on the trend scale (Donoghue & Mazzeo, 1992; Muraki, Hombo, & Lee, 2000).

PISA is a large-scale international trend study that has investigated scale linking error in regard to trend estimation. The PISA approach to trend estimation incorporates the separate calibration method using Rasch scaling, followed by plausible value estimation and scale linking performed by matching characteristics of the achievement distributions. Linking error is then calculated based on Rasch item parameter estimates for the subset of common or link items. Specifically, linking error in PISA is calculated as the standard error of the mean difference in the Rasch difficulty parameters of the common items calibrated from two adjacent cycles.

To calculate linking error between the 2000 and 2003 reading and science administrations, for example, PISA first estimated the Rasch difficulty parameters for the 2003 assessments. The common item parameters were then centered by setting the mean to 0. The Rasch difficulty parameters for the common items in the 2000 administration were also centered by setting the mean to 0. The difference between the two item parameters was calculated and averaged across all items, thus representing the difference in relative difficulty of the common items across cycles. Finally, linking error was computed as the standard error of the mean difference using the traditional formula for calculating the standard error of a mean (Monseur & Berezner, 2007; OECD, 2005). This approach assumes that the common items are a random sample of all possible common items.

It is evident that linking error in PISA is a function of both item drift (differences in item difficulty from one administration to the next) and the number of common items across surveys. The linking error formula was modified for PISA 2006 to account for the clustering of items within passages/blocks for the reading assessment as well as for the inclusion of partial credit items (OECD, 2009).

Trend estimates for the PISA reading and science literacy scales linking PISA 2000, PISA 2003, and PISA 2006 were based on 22 to 28 common items and had linking error estimates ranging from 3.11 to 5.30 points on the PISA (500, 100) achievement scales. Trends in mathematics literacy from 2003 to 2006 were based on 48 common items and had a somewhat smaller link error of 1.38 points.

MEASURING TRENDS IN TIMSS AND PIRLS

Similar to other large-scale studies, TIMSS and PIRLS use IRT-based scaling methodology to estimate item parameters, generate plausible values for each student (with these values based on the conditioning model), and then link the current assessment scale to the trend scale. To make maximum use of the data from successive assessment cycles, TIMSS and PIRLS use the concurrent calibration method, as opposed to the separate calibration method, to estimate item parameters.

In this approach, all assessment items from both the current assessment and the previous assessment are included in the calibration for all countries participating in both assessment cycles. The PIRLS 2001 assessment, for example, consisted of eight passages, four literary and four informational, with a total of 98 items (133 score points). In the next cycle of PIRLS (2006), half of the PIRLS 2001 assessment (four passages and 49 items, worth 66 points) was reassessed for the purpose of trend measurement. The PIRLS 2006 assessment accordingly consisted of 10 passages,[3] five literary and five informational, with a total of 125 items with 165 score points. Four of these passages (and 49 items) were trend passages from 2001, and the other six passages and item sets (three literary and three informational) were newly developed to reflect the current environment and context of reading literacy (Martin, Mullis, & Kennedy, 2007). In summary, 49 items were unique to 2001, 49 items were common to 2001 and 2006, and 76 items were unique to 2006 (Foy et al., 2007).

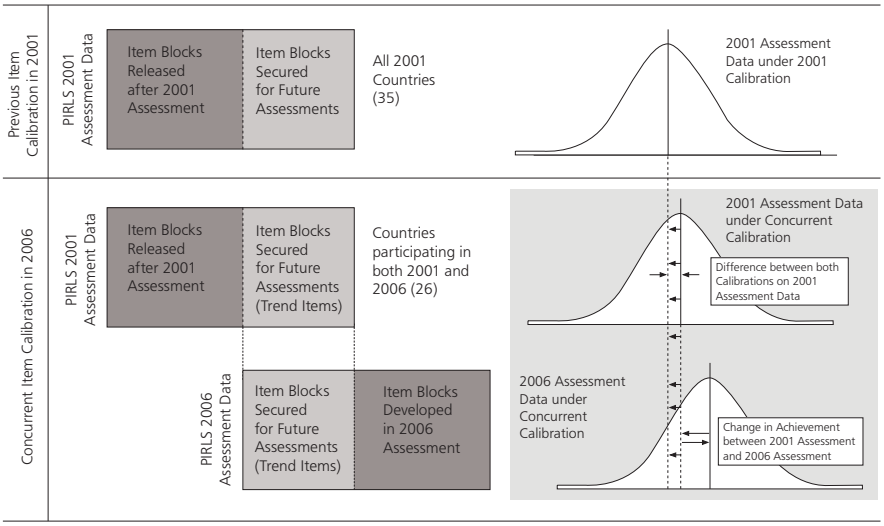
Figure 1 shows the concurrent calibration model used for PIRLS 2006. The right-hand side of the top panel shows that scaling the PIRLS 2001 achievement data resulted in a distribution of student results. That is, after item parameters were estimated in the PIRLS 2001 calibration, these item parameters were used to "score" the student responses and yielded the achievement results published in the *PIRLS 2001 International Report* (Mullis, Martin, Gonzalez, & Kennedy, 2003).

As explained in the *PIRLS 2001 Technical Report* (Martin, Mullis, & Kennedy, 2003), the process of estimating achievement used information about each student's responses to the administered items and the student's background characteristics to impute student scores, or plausible values, on the assessment as a whole. To quantify error in the imputation process, five plausible values were generated for each student, and all analyses were conducted five times. The average of the five analyses was taken as the

3 In PIRLS 2006, the assessment was extended to include 10 passages instead of eight passages as in PIRLS 2001. PIRLS 2011 also includes 10 passages, of which six passages containing 75 items are common to both PIRLS 2006 and PIRLS 2011.

best estimate of the statistic in question, and the variance among them reflected the imputation error. The PIRLS achievement scale metric was established, based on the 2001 data, as having a mean of 500 and a standard deviation of 100.

Figure 1: Concurrent calibration model used for PIRLS 2006



Source: IEA Trends in International Mathematics and Science Study (TIMSS) 2007.

Figure 1 also indicates (top left) that, after publication of the PIRLS 2001 achievement results, about half of the PIRLS 2001 items were released to the public and the other half were kept secure to be used again in PIRLS 2006. The lower panel of Figure 1 shows the concurrent calibration used in PIRLS 2006. The 2006 calibration included only those countries that participated in both 2001 and 2006. For these 26 countries, data from the entire PIRLS 2001 assessment were rescaled together with the new data from the PIRLS 2006 assessment.

To resolve the issue of metric indeterminacy inherent in IRT procedures, the scale was set such that the mean and standard deviation of the combined distribution were approximately 0 and 1, respectively. The item parameters from the 2006 concurrent calibration were used to score the 2006 data and produce the achievement results published in the *PIRLS 2006 International Report* (Mullis, Martin, Kennedy, & Foy, 2007).

As can be seen from the top panel of Figure 1, the PIRLS 2001 items had a set of item parameters from the original item calibration conducted in 2001. With the concurrent calibration approach, item parameters for the trend items are not fixed, but are re-estimated with each cycle. Thus, the PIRLS 2006 concurrent calibration, using all the data from 2001 and 2006 for countries that participated in both cycles, resulted in a second set of item parameters for the 2001 items. The second set of item parameters reflects changes in the pool of items used to measure achievement and shifts in the item parameter estimates. As part of this process, the two sets of parameters for the trend items were compared item by item, and the shifts typically were minor.

As shown in the lower panel of Figure 1, there was also a small change between the distribution of the PIRLS 2001 data under the original item calibration and the distribution of the PIRLS 2001 data under the concurrent calibration. The difference was due to changes in item parameters from the original item calibration in 2001 (based on 2001 data) and the concurrent calibration item parameters in 2006 (based on both 2001 and 2006 data).

PIRLS 2006 Scale Linking

The method used to place the PIRLS 2006 data on the trend scale is based on a procedure used in the National Assessment of Educational Progress (NAEP), described by Donoghue and Mazzeo (1992) and Muraki et al. (2000) and further investigated by Lee, Song, and Kim (2004). First, a linear transformation was performed to match the distribution of the 2001 data under the concurrent (2006) item calibration to the distribution of the 2001 data under the original (2001) item calibration. This transformation ensured that the mean and standard deviation of the distribution of 2001 data under the concurrent calibration aligned to the mean and standard deviation of the distribution of 2001 data from the original item calibration (i.e., the data were placed on the same scale). Next, the same linear transformation was applied to the PIRLS 2006 data. This placed the PIRLS 2006 data on the PIRLS 2001 metric, allowing the achievement trend between PIRLS 2001 and PIRLS 2006 to be estimated.

In summary, four steps were taken in PIRLS 2006 to estimate item parameters and ability distributions and to place these estimates on the trend scale:

1. Based on trend countries (i.e., countries that participated in both assessments), item parameters were estimated via concurrent calibration for all items on both the PIRLS 2001 and PIRLS 2006 assessments.
2. Achievement distributions were estimated for the PIRLS 2001 trend countries using the item parameters from the concurrent calibration.
3. The linear transformation was determined that best matched the PIRLS 2001 achievement distributions estimated under the concurrent calibration to the PIRLS 2001 achievement distributions published in 2001.
4. The linear transformation determined in (3) was applied to the PIRLS 2006 achievement distributions to place the current estimates on the trend scale.

Estimating Linking Error in PIRLS 2006

On completion of the concurrent calibration in 2006, each PIRLS 2001 item had two item parameter estimates: one estimate based on the 2001 calibration and one estimate based on the 2006 concurrent calibration. These made it possible to produce two achievement estimates from the 2001 data, one using the 2001 item parameters and the other using the 2006 parameters. Differences between these two estimates reflect linking error in PIRLS.

Linking error was computed as a function of the standard errors of the differences in achievement estimates resulting from the two sets of item parameters. First, plausible values for students in each trend country were generated using the 2001 student responses with the 2001 item parameters (i.e., the results published in the *PIRLS 2001 International Report*). All 2001 item parameters—as opposed to only common item parameters—were used to obtain achievement estimates so as to be consistent with the 2001 scaling and also to base achievement estimates on all available data. Plausible values were then generated for students in each trend country using the same data (i.e., the 2001 student responses), but substituting the reestimated (2006) item parameters for the original 2001 item parameters. Because the student responses were identical under both conditions, the difference between the two achievement estimates reflects the effect of the changes made to item parameters between 2001 and 2006, and its standard error may be considered an estimate of the linking error (Johnson, 2005).

The variability due to linking was estimated by jackknifing[4] the differences between the two achievement estimates across the five plausible values. Thus, student-level differences on the first plausible value were determined between estimates obtained on the 2001 data using the 2001 item parameters and estimates obtained on the 2001 data using the 2006 item parameters. These differences were then calculated for the second, third, fourth, and fifth plausible values. The average difference across each of the five plausible values was then calculated as the statistic of interest, and its standard error was estimated using the jackknife procedure.

The linking error was estimated separately for each country so that country-level results in PIRLS 2006 could be adjusted accordingly. Twenty-six countries and two Canadian provinces participated in both PIRLS 2001 and PIRLS 2006, and linking error was estimated for each participant. This approach differs from the approach utilized by PISA, which calculates only one overall linking error across all participating countries.

Linking Error Results for PIRLS 2006

Table 1 presents the PIRLS 2006 linking error results. The first two columns in the table present the average achievement and standard error for the PIRLS 2001 data based on the 2001 item parameters, which are the results originally reported in PIRLS

4 See Wolter (2007) for a description of the jackknife procedure.

2001 (Mullis et al., 2003). The next two columns display the average achievement scores and the respective standard errors based on the same data using the item parameters from the 2006 concurrent calibration. The last two columns present the differences between average achievement based on the original 2001 calibration and average achievement based on the 2006 concurrent calibration. The standard errors of the differences are the linking-error estimates for the link between PIRLS 2001 and PIRLS 2006. The differences between achievement estimates are very small, mostly less than half a score point. The linking errors are also very small, ranging from 0.6 to 2.0, with an international average of 1.1.

Table 1: Average achievement and linking-error estimates using PIRLS 2001 data

Country	2001 data estimated from PIRLS 2001 calibration		2001 data estimated from PIRLS 2006		Difference in average achievement	
	Average achievement	*SE*	Average achievement	*SE*	Achievement difference	*SE* of difference (linking error)
Bulgaria	550	3.8	551	3.8	0.3	1.0
England	553	3.4	552	3.3	-0.6	1.4
France	525	2.4	526	2.4	0.4	0.7
Germany	539	1.9	539	1.8	0.0	0.8
Hong Kong SAR	528	3.1	528	3.2	0.3	0.8
Hungary	543	2.2	544	2.1	0.3	1.1
Iceland	512	1.2	512	1.2	-0.7	1.1
Iran, Islamic Rep. of	414	4.2	414	4.4	0.6	1.5
Israel	509	2.8	509	2.8	0.2	1.2
Italy	541	2.4	540	2.4	-0.3	0.8
Latvia	545	2.3	544	2.2	-0.5	2.0
Lithuania	543	2.6	544	2.5	0.2	1.3
Macedonia, Rep. of	442	4.6	442	4.8	0.9	1.1
Moldova, Rep. of	492	4.0	492	4.2	0.0	1.0
Morocco	350	9.6	346	10.0	-3.3	1.5
Netherlands	554	2.5	554	2.7	0.1	1.2
New Zealand	529	3.6	529	3.8	0.2	1.4
Norway	499	2.9	500	2.8	0.7	1.1
Romania	512	4.6	512	4.6	-0.1	0.8
Russian Federation	528	4.4	528	4.2	0.0	1.0
Scotland	528	3.6	528	3.5	0.2	1.0
Singapore	528	5.2	528	5.2	0.1	0.6
Slovak Republic	518	2.8	518	2.8	0.2	1.2
Slovenia	502	2.0	502	1.9	0.3	1.3
Sweden	561	2.2	561	2.3	0.1	1.2
United States	542	3.8	542	3.8	0.3	0.9
International average	517	3.4	517	3.4	0.0	1.1
Ontario, Canada	548	3.3	548	3.3	-0.1	1.3
Québec, Canada	537	3.0	538	2.8	0.5	1.1

Table 2 presents average reading achievement and the respective standard errors in PIRLS 2001 and PIRLS 2006 (Mullis et al., 2007). This table also displays the average difference for each participant between the two cycles and its standard error, computed without reference to linking error. Reliance on these traditional standard error estimates in PIRLS 2006 led to 14 countries showing statistically significant changes in reading achievement between 2001 and 2006.

Table 2: Trends in reading achievement

Country	PIRLS 2001 average scale score		PIRLS 2006 average scale score		Difference between PIRLS 2001 and 2006 scores		
	Average achievement	*SE*	Average achievement	*SE*	Difference	*SE* without linking error	*SE* including linking error
Russian Federation [2a]	528	4.4	565	3.4	37	5.6*	5.6*
Hong Kong SAR	528	3.1	564	2.4	36	3.9*	4.0*
Singapore	528	5.2	558	2.9	30	5.9*	5.9*
Slovenia	502	2.0	522	2.1	20	2.9*	3.2*
Slovak Republic	518	2.8	531	2.8	13	4.0*	4.1*
Italy	541	2.4	551	2.9	11	3.8*	3.8*
Germany	539	1.9	548	2.2	9	2.9*	3.0*
Moldova, Rep. of	492	4.0	500	3.0	8	5.0	5.1
Hungary	543	2.2	551	3.0	8	3.7*	3.9*
Iran, Islamic Rep. of	414	4.2	421	3.1	7	5.2	5.4
Canada, Ontario [2a]	548	3.3	554	2.8	6	4.4	4.5
Israel [2b]	509	2.8	512	3.3	4	4.4	4.5
New Zealand	529	3.6	532	2.0	3	4.1	4.3
Macedonia, Rep. of	442	4.6	442	4.1	1	6.2	6.3
Scotland†	528	3.6	527	2.8	-1	4.6	4.7
Norway ‡	499	2.9	498	2.6	-1	3.9	4.0
Iceland	512	1.2	511	1.3	-2	1.8	2.1
United States†[2a]	542	3.8	540	3.5	-2	5.2	5.2
Bulgaria [2a]	550	3.8	547	4.4	-3	5.8	5.9
France	525	2.4	522	2.1	-4	3.1	3.3
Latvia	545	2.3	541	2.3	-4	3.3	3.8
Canada, Québec	537	3.0	533	2.8	-4	4.1	4.2
Lithuania	543	2.6	537	1.6	-6	3.1*	3.3
Netherlands†	554	2.5	547	1.5	-7	2.9*	3.2*
Sweden	561	2.2	549	2.3	-12	3.2*	3.4*
England	553	3.4	539	2.6	-13	4.3*	4.5*
Romania	512	4.6	489	5.0	-22	6.8*	6.8*
Morocco	350	9.6	323	5.9	-27	11.3*	11.4*

Notes:

* $p < 0.05$.

† Met guidelines for sample participation rates only after replacement schools were included.

‡ Nearly satisfied guidelines for sample participation rates after replacement schools were included.

2a National defined population covers less than 95% of national desired population.

2b National defined population covers less than 80% of national desired population.

Trend note: The primary education systems of the Russian Federation and Slovenia underwent structural changes. Data for Canada, Ontario include public schools only.

The standard error of the difference without including linking error is computed as $SE = \sqrt{SE_1^2 + SE_2^2}$, where SE_1 is the standard error from PIRLS 2001 and SE_2 is the standard error from PIRLS 2006. To include the linking error in the standard error of the difference, the estimate of the linking error for each participant was combined with the existing standard error of the difference as $SE = \sqrt{SE_1^2 + SE_2^2 + SE_L^2}$, where SE_1 and SE_2 are the standard errors from PIRLS 2001 and PIRLS 2006, respectively, and SE_L is the standard error of the link.

Results reflecting the revised standard error estimates are presented in the last column of Table 2. Comparing the published standard errors of the difference between the 2001 and 2006 scale scores to the standard errors that include the linking error demonstrates that the linking error had very little impact on the statistical significance of the PIRLS 2006 trend estimates. In most countries, the standard error that included the linking error increased by less than 0.1 of a point. Due to these small changes, the statistical significance of the trend estimates remained the same for all but one country—Lithuania. When the linking error was set aside, the difference between PIRLS 2001 and PIRLS 2006 average scale scores for Lithuania was -6.35 with a standard error of 3.1, which was a statistically significant decrease in reading achievement between 2001 and 2006. However, when the linking error was included, the standard error became 3.3, making the change in reading achievement not statistically significantly different from zero.

Despite this one change in significance, most participants' trend estimates were not greatly affected by the inclusion of the linking error. For example, Latvia had the highest linking error estimate (2.0). However, adding the linking error to the traditional standard error estimate ($\sqrt{2.3^2 + 2.3^2} = 3.25$) only increased the error by 0.6 points ($\sqrt{2.3^2 + 2.3^2 + 2.0^2} = 3.81$) and did not affect the statistical significance of the trend estimate (i.e., the difference remained nonsignificant).

CONCLUSION AND IMPLICATIONS

As is the case with other large-scale assessments, TIMSS and PIRLS measure trends in student achievement by administering assessments to national student samples every four or five years. About 60% of each assessment is reassessed from previous cycles, and the rest is newly developed. Linking the results of successive assessments to a common achievement scale is a statistical estimation process that necessarily involves some degree of estimation error.

This paper described a procedure for estimating the error in the TIMSS and PIRLS linking process. It also described an application of that procedure to the linking of the PIRLS 2006 data to the PIRLS achievement scale originally established by PIRLS 2001. Because new assessment items are introduced with each assessment cycle as replacements for released items, each cycle requires an item calibration to determine the values of the item parameters necessary to estimate the distribution of student achievement. The fact that item parameters are not constant from cycle to cycle introduces some uncertainty, or linking error, into the trend measurement process.

To minimize linking error, TIMSS and PIRLS use a concurrent calibration process involving items and data from the previous as well as the current assessment and including items common to both assessment cycles. As a result of this process, item parameters are not only determined for all new items but are also reestimated for items common to both the current and previous assessments. The reestimation allows the parameter values of the common items to gradually evolve from assessment cycle to cycle, while providing the best possible fit to the assessment data.

Because the concurrent calibration approach to item parameter estimation uses all available data from both the current and previous assessment cycles, this approach provides the best estimate of the item parameter values. As an additional advantage, it also provides the data to estimate the linking error that results from changes in item parameters from cycle to cycle. The concurrent calibration process results in two sets of item parameter estimates for the items in the "previous" assessment—one set from the calibration conducted for the previous cycle and a second, more recent set from the new concurrent calibration. The differences between the two sets of item parameters can be used to estimate linking error.

For the PIRLS example in this paper, the PIRLS 2001 items had item parameter estimates based on the original 2001 item calibration and a second set based on the 2006 concurrent calibration. Applying the two sets of item parameters to the same data (the PIRLS 2001 data in this paper) provided a measure of the effect on achievement estimation of using one set of parameters in place of the other. The difference between the achievement estimates is a measure of the linking error due to the change in item parameters.

In contrast to the PISA approach to linking error, which is based on the variance of the linking-item difficulty parameters, the TIMSS and PIRLS approach considers linking error in terms of differences in student achievement distributions due to item parameter changes. As such, the latter approach more closely addresses the consequences for student achievement of evolving assessment item pools from one assessment cycle to the next. An advantage of the focus on changes in student achievement distributions is that linking error can be estimated and reported country by country rather than as a single global estimate, as in PISA.

This study indicates that trend estimation in PIRLS 2006 was not greatly affected by including linking error in the computation of standard errors. Most likely, this outcome is a result of the relatively large number of trend items in the PIRLS 2006 assessment design, which reassessed approximately 50% of the assessment from PIRLS 2001 (Martin et al., 2007). The PIRLS linking error, based on 49 common items, averaged 1.1 across the countries, which is relatively close to PISA's estimate of 1.4 for mathematics literacy based on 48 items. These linking error estimates are considerably less than the PISA estimates of 3.1 to 5.3 for reading and science literacy, based on no more than 28 common items.

The findings of this study, and the comparison with PISA results, provide support for the idea that the key to reliable trend measurement lies in having a sufficiently large

number of common items in adjacent assessments. Including a component for linking error in tests of achievement differences between two PIRLS assessment cycles with 49 items in common made very little difference to the overall outcome. Nonetheless, to ensure that linking error is further reduced in PIRLS, the number of items common to PIRLS 2006 and PIRLS 2011 has been increased to 75. As indicated earlier, TIMSS also has large numbers of items common to the 2007 and 2011 assessments: 126 and 125 for eighth-grade mathematics and science, respectively, and 103 and 100 for fourth-grade mathematics and science, respectively.

References

Donoghue, J. R., & Mazzeo, J. (1992, April). *Comparing IRT-based equating procedures for trend measurement in a complex test design*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco.

Foy, P., Galia, J., & Li, I. (2007). Scaling the PIRLS 2006 reading assessment data. In M. O. Martin, I. V. S. Mullis, & A. M. Kennedy (Eds.), *PIRLS 2006 technical report* (pp. 149–172). Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Johnson, E. (2005). *Trend linking error in PIRLS and TIMSS*. Unpublished manuscript.

Kolen, M., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices* (2nd ed.). New York, NY: Springer.

Lee, W-C., Song, M-Y., & Kim, J-P. (2004, April). *An investigation of procedures for obtaining a common IRT scale*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego.

Martin, M. O., Mullis, I. V. S., & Kennedy, A. M. (Eds.). (2003). *PIRLS 2001 technical report*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Martin, M. O., Mullis, I. V. S., & Kennedy, A. M. (Eds.). (2007). *PIRLS 2006 technical report*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Monseur, C., & Berezner, A. (2007). The computation of equating errors in international surveys in education. *Journal of Applied Measurement*, *8*(3), 323–335.

Mullis, I. V. S., Martin, M. O., Gonzalez, E. J., & Kennedy, A. M. (2003). *PIRLS 2001 international report: IEA's study of reading literacy achievement in primary schools*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Mullis, I. V. S., Martin, M. O., Kennedy, A. M., & Foy, P. (2007). *PIRLS 2006 international report*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Muraki, E., Hombo, C. M., & Lee, Y-W. (2000). Equating and linking of performance assessments. *Applied Psychological Measurement*, *24*, 325–337.

Organisation for Economic Co-operation and Development (OECD). (2005). *PISA 2003 technical report*. Paris, France: OECD Publishing.

Organisation for Economic Co-operation and Development (OECD). (2009). *PISA 2006 technical report*. Paris, France: OECD Publishing.

Wolter, K. M. (2007). *Introduction to variance estimation* (2nd ed.). New York, NY: Springer.

Exploring the measurement profiles of socioeconomic background indicators and their differences in reading achievement: A two-level latent class analysis

Kajsa Yang Hansen and Ingrid Munck
Department of Education and Special Education, University of Gothenburg, Sweden

Applying a two-level mixture modeling technique, the study explored the psychometric profiles of socioeconomic status (SES) and examined reading achievement differences according to the latent SES profiles. The two-level latent class analysis (TLCA) takes into consideration the measurement error in the response patterns of the SES indicators, and the variation of the latent class indicators across different schools. It also controls for individual characteristics to assure precision in latent class estimation and differences in reading achievement between classes. The SES background variables and reading achievement variable were taken from the Swedish Grade 4 data of the Progress in Reading Literacy Study (PIRLS) 2006.

The analysis identified three latent classes of individuals, namely the economically and culturally affluent group, the culturally disadvantaged group, and the culturally well-off group. Reading achievement differed significantly across the three SES classes. About 16% of the differences in reading achievement could be attributed to the SES differences of individual students in different latent SES classes. The lowest achieving class was the immigrant-concentrated, culturally disadvantaged group. Only for this group, speaking the test language at home had a significant impact on reading achievement. The school-level continuous factor captured the variation in the SES composition of student intake across schools, and it explained almost half of the between-school differences in reading achievement. The findings in this study may imply that a mixed societal and school environment can compensate for students' disadvantaged family background. Educational investment in this group of students may reduce educational inequality and school segregation.

IERI Monograph Series: Issues and Methodologies in Large-Scale Assessments Volume 5

BACKGROUND

Numerous studies have established the association between students' socioeconomic status (SES) and academic achievement, making SES one of the powerful predictors of student performance (see, for example, Coleman et al., 1966). Typically, family SES is represented by parental education, occupation, and income. However, there is no consensus on how SES is composited and measured by its indicators. A variety of alternative measurements of SES has also been developed, which has led to an inconsistency of predictive value of SES on, for example, academic achievement and cognitive development (see, for example, Sirin, 2003, 2005; White, 1982).

As a background variable, SES fulfills one of the following functions:

1. A control variable to statistically adjust for background differences and confounding effects with other factors;
2. A stratification variable to increase the precision of comparing treatment effects or interaction effects of different treatments within different SES groups;
3. An independent variable in causal models to examine its effects on educational outcomes;
4. A descriptive aggregated variable at classroom or school level in contextual, teaching effect, or school-effect studies (White, 1982; see also Willms, 1992).

Depending on the function it may carry, SES is defined according to differentiated psychometric properties. In most of the effect studies, SES is measured as a one-dimensional composition. This line of research adopts the Weberian tradition, regarding SES as a combination of property, power, and prestige, the components of which are well reflected by family material assets, income, and parental education and occupation on an integrated continuum (see, for example, Bradley & Corwyn, 2002).

White (1982) reviewed over 200 studies and found an overall average SES-achievement correlation of 0.35. He also observed that the relationship between SES and achievement varied greatly across studies, contingent upon the measures of SES and the unit of analysis. When SES is defined as a composite of father's income, education, and occupation, and the individual is used as the unit of observation, SES is only weakly correlated with academic achievement (r = 0.22). When the SES-achievement relationship is determined at aggregate levels, such as the school level, the correlation is higher (r = 0.73).

This aggregated-level SES-achievement correlation is an ecological correlation, which is based on school average of SES and achievement. And it tends to be much higher compared to the individual correlation, due to the fact that measurement errors of the school average in SES and reading achievement are reduced by the aggregation across units. Thus, "there needs to be no correspondence between the individual correlation and the ecological correlation" (Robinson, 1950, p. 354). Sirin (2005) also noted in his meta-analysis of SES effect that much of the variation in the magnitude

of the SES-achievement relation is due to "methodological characteristics, such as the type of SES measures, and student characteristics, such as student's grade, minority status, and school location ... [all of which] moderated the magnitude of the relationship between SES and academic achievement" (p. 438).

However, a general trend, observed in the recent literature on measuring SES and its effects, is that of moving from a conception of SES as a single composition toward a multidimensional normally distributed continuous latent construct that imposes its effects differently at different levels of observation, for example, students and schools (Sirin, 2003, 2005). Bourdieu's cultural reproduction theory and forms of capital offer a conceptual framework with respect to defining the dimensionality and structure of SES and explaining educational inequality (see, for example, Lareau & Weininger, 2003).

According to Bourdieu (1997), social structure and function are impossible to explain unless all forms of capital are recognized. Capital in the monetary sense (i.e., economic capital) along with the intangible forms of capital, namely cultural capital and social capital, form the three fundamental manifestations of resources, and the possession of one form of capital can influence the chance to possess other forms of capital (Bourdieu, 1997). When explaining his cultural reproduction theory, Bourdieu argued that modern education systems transform the social hierarchies into academic hierarchies (Bourdieu, 1977; Bourdieu & Passerson, 1977). The transmission of status operates largely through cultural capital, by which the relative social advantages of individuals or groups are maintained and promoted. Lack of knowledge of institutionalized cultural norms, such as attitudes, preferences, formal knowledge, behaviors, goods, and credentials, negatively affects the academic outcomes of individuals and therefore hinders their upward mobility (Bourdieu & Boltanski, 1979; see also Lamont & Lareau, 1988).

Applying Bourdieu's capital concepts and a two-level structural equation modeling technique (see, for example, Muthén, 1989, 1991), Yang (2003) measured family and school SES simultaneously with a set of home possession items from the IEA survey studies. An economic capital factor and a cultural capital factor were identified at the individual level. However, only one general capital factor was found at the school level. The latent structure of SES differs across countries. Such differences are reflections of country-specific social, cultural, and economic factors, as well as differences in availability of SES data in IEA studies. It also was found that the cultural capital factor in general had a significant and positive impact on students' academic achievement, while the relationship between the economic capital factor and academic achievement was negative or non-significant. The school-level general capital factor, representing the SES composite of the school student bodies, alone explained a substantial amount of school differences in academic achievement (Yang, 2003; see also, Yang & Gustafsson, 2004). Such a multidimensional, multilevel measurement of SES can thus offer a more detailed understanding of the mechanisms through which SES exerts its impact on academic achievement.

Another major use of SES is, however, to classify individuals according to the level of family socioeconomic circumstances. In this context, SES is a categorical index that very often is based on the average of the observed SES indicators. Several issues are raised in respect of such an index. First, measurement errors and missing responses in the relatively few construct indicators may cause the construct index to have a low reliability. Second, the assumption that construct indicators contribute equally to the index they measure may not hold, since a construct can relate more or less strongly to different indicators. Third, little is known about the validity of the index.

One way to deal with these methodological problems is to adopt a multivariate latent variable modeling approach, whereby measurement models are used to represent the construct indicators (see, for example, Yang Hansen, Rosén, & Gustafsson, 2006). This approach allows testing of the hypothesized model against data, and it optimally weights the contribution of each of the indicators to its underlying latent variable. Measurement models also allow estimation of individual latent variable scores (i.e., factor scores) from partially missing data. Factor score estimates may be treated as observed variables and used both in regular analyses and categorization of individuals. However, because a factor score does not have a natural cut point for SES groups, the division of SES groups seems to be more arbitrary in the factor score approach.

The two-level latent class analysis (TLCA) approach offers new possibilities for exploring the psychometric profiles of SES (Vermunt, 2003). It is assumed that SES has a multinomial distribution and this assumption is conceptualized by forming discrete latent categories or typologies that are based on the prior and posterior probability distributions under conditional maximum likelihood estimation (Henry & Muthén, 2010; Muthén, 2007).

One of the advantages of TLCA is that it takes into consideration the measurement error (i.e., misclassification of individuals) in the response patterns of the SES indicators, and achieves more precise classification of individuals through estimated latent classes. The latent class membership can then be merged with the original data file and used as an estimated SES index. Another advantage of TLCA is that it allows researchers to adjust the biases caused by cluster sampling designs by modeling the hierarchical data structure and then simultaneously examining the SES effects on academic achievement at individual and collective levels.

The aim of the current study was to apply two-level latent class analysis in order to identify the unobserved categories of individuals according to a set of SES measures in the IEA Progress in Reading Literacy Study (PIRLS) 2006 data (Mullis, Martin, Kennedy, & Foy, 2007). A further aim was to examine the extent to which different latent SES classes and collective SES can account for differences in reading achievement.

METHOD

Sample and Variables

The current analysis drew on the Swedish data from the 2006 iteration of IEA PIRLS. The Swedish sample consisted of 4,393 Grade 4 students and 147 schools. Variables indicating family SES and school student-intake characteristics were selected from the study's student questionnaire (StQ) and home questionnaire (HQ), in order to identify latent SES classes of individuals while taking into account the clustering of individuals in different schools. A measure of student reading achievement (ASRREA01) was included so that the mean achievement differences among different latent classes could be examined. A set of student-level background variables was also included in order to describe the characteristics of the estimated latent classes.

Table 1 shows the descriptive statistics of SES indicators involved in identifying the latent classes, together with the covariate "use of the test language at home" (ASBGLNG1) and the distal reading outcome variable ASRREA01 at the individual level. The number of observations of certain categories in some of the SES indicators is fairly low, which may cause a sparse distribution of individuals in certain cells in the cross-tabulation among different variables. Because this, in turn, can affect maximum likelihood estimation, making it difficult for the statistical model to converge to the global maximum, categories with few observations were reclassified into a larger category of the variable.

As shown in Table 1, five SES indicators were taken from the HQ: highest level of parental education (ASBHHEDUP), number of books at home (ASBHBOOK), number of children's books at home (ASBHCHBK), family affluence (ASBHWELL), and highest level of parental occupation (ASBHHOCP). A sum score of the five items used in the StQ to denote possessions potentially found in students' homes (personal computer, desk, own books, newspapers, and own room) was also derived and then categorized into low, medium, and high levels. It was used to represent the level of educational aids at home (EDUAIDS). These six SES indicators were recoded into ordinal categorical variables, with the low or negative category being coded as 1 and the high or positive category being coded as 3.

Among these SES indicators, two categories of variables can be distinguished. One category, in line with Bourdieu (1984, 1997), is the objectified state of cultural capital, which signifies the cultural preferences of people in everyday life. Proxies for the concept of cultural capital are number of books at home and the educational level of the parents. Another category relates to family wealth (i.e., economic capital). Information about family affluence, parental occupation, and educational aids in the home for children functions as an indicator of the economic aspect of SES. Table 1 also presents frequencies of individuals in each category of the variables as well as the number of missing observations in each variable.

Table 1: Descriptive information on all variables included in the mixture modeling of SES

Variables	Label	Source	Scale	Value labels (percentage of individuals in each category)	Missing
ASBHHEDUP	Highest level of parental education	HQ	Ordinal	1 = finished upper-secondary school (29.2%) 2 = finished post-secondary but not university (36.6%) 3 = finished university or higher (34.2%)	696
ASBHBOOK	Number of books at home	HQ	Ordinal	1 = fewer than 25 (10.4) 2 = 26–100 (26.7%) 3 = more than 100 (62.9%)	303
ASBHCHBK	Number of children's books at home	HQ	Ordinal	1 = fewer than 25 (15.1%) 2 = 26–100 (57.2%) 3 = more than 100 (27.7%)	296
ASBHWELL	Family affluence status: being well-off family	HQ	Ordinal	1 = not well-off (7%) 2 = average (41%) 3 = well-off (52%)	
ASBHHOCP	Highest level of parental occupation	HQ	Ordinal	1 = skilled worker, general laborer, or never worked outside home for pay (8.6%) 2 = small business-owner or clerical (33.7%) 3 = professional (57.7%)	451
EDUAIDS	Level of educational aids in home: an index of summed scores of the five common home-possessions items (personal computer, desk, own room, newspaper, books)	StQ	Ordinal	1 = low 0–3 (8%) 2 = average 4 (27.5%) 3 = high 5 (64.5%)	65
ASBGLNG1	Use of the test language at home	StQ	Binary	1 = yes (94.8%); 0 = no (5.2%)	162
ASRREA01	Plausible value: overall reading achievement	StQ	Continuous	Mean = 548.7; *SD* = 63.5	0

Notes:

The percentages presented in the parentheses are valid percentages and the dataset was weighted by house weight (HOUWGT).

Sample size = 4,393.

The variable ASBGLAN1 (i.e., extent to which the language of the PIRLS test was being used at home) is an indicator of students' ethnic backgrounds. It was used as a covariate at the individual level to further describe the estimated latent SES classes in the latent class analysis (LCA) modeling.

An important aspect of LCA is that it allows researchers to investigate the characteristics of individuals within each latent class by relating the latent classes to auxiliary variables. Auxiliary variables can be, for example, covariates, concurrent outcomes, and distal outcomes that are not involved in the estimation of latent classes (see, for example, Clark & Muthén, 2009). Table 2 presents these auxiliary variables, all of which are, in their original scale, individual-level variables taken from the StQ or the HQ. These were used in the current study to further describe the latent SES classes.

Given the latent SES class membership, the class-specific means (i.e., conditional means) of the variables in Table 2 can be evaluated and compared by conducting a Wald test of mean equality using Mplus software (Muthén & Muthén, 2010a). The results of the comparison can then be used to validate the interpretation of the latent SES classes and further depict the class characteristics (Clark & Muthén, 2009; Muthén & Muthén, 2010a).

Table 2: Auxiliary variables of student characteristics used in the latent class analysis

Variables		Source	Highest category (no. of categories)	Missing
ASBHMJF	Main job of father	HQ	Professional (11c)	727
ASBHMJM	Main job of mother	HQ	Professional (11c)	682
ASBGTA5	Own room	StQ	Yes (dummy)	84
ASBHLEDF	Highest education of father	HQ	Beyond ISCED Level 5A (7c)	832
ASBHLEDM	Highest education of mother	HQ	Beyond ISCED Level 5A (7c)	861
ASBGBOOK	Number of books at home	StQ	Over 100 books (5c)	146
ASBGTA4	Daily newspaper	StQ	Yes (dummy)	97
ASDHHER	Index home educational resources	StQ	High (3c)	344
ASBGBRNM	Mother born in the country	HQ	Yes (dummy)	200
ASBGBRNF	Father born in the country	HQ	Yes (dummy)	159

Note: Sample size = 4,393.

Analytical Method

Latent class analysis (LCA, see, for example, Hagenaars & McCutcheon, 2002) was used in the first step to identify small numbers of unobserved homogenous groups of individual students inferred from the properties of the set of SES measures. A latent class is characterized by a pattern of conditional probabilities that indicates the degree of association of the observed dependent variables with each of the latent classes (i.e., SES indicators).

The conditional probabilities of SES indicators can be understood in the same way as the factor loadings in factor analysis. The conditional probability signifies the chance of a certain response pattern being chosen in the set of observed indicators, given the individual's membership within a latent class. In other words, conditional probabilities indicate the likelihood of an individual within a latent class giving a particular response on an observed measure. Like the factor loadings, the conditional probabilities offer measurement profiles that describe the latent classes. By looking at the pattern of responses for all SES measures, one obtains an overview of the nature of each latent class, thus assisting in interpreting the relationship between these latent classes and other outcome variables such as reading achievement.

As is the case with conventional covariance structure modeling, the observations are assumed to be independently sampled in LCA. However, the stratified multistage cluster sampling design used in PIRLS violates this assumption. For example, students in the same school will, when compared to students from different schools, achieve similarly and have a similar demographic background. For this reason, the multilevel technique was needed in the current study to account for such dependencies (Asparouhov & Muthén, 2008; Vermunt, 2003, von Davier, 2010).

The two-level latent class model applied in the current study not only allowed an examination of the multilevel structure of the data but also made it possible to detect subgroups of individuals according to the observed SES properties. In other words, the two-level latent class analysis took into account the nested structure of the data, thereby allowing the aggregated SES indicators to differ across second-level units (i.e., indicator-specific random effects; Henry & Muthén, 2010). This type of analysis can also detect whether, and if so how, the second-level units influence the lower-level latent classes. Here, the process involves comparing the latent class profiles of SES that were estimated by taking into account the indicator-specific random effects at school level with those that were estimated without (for a detailed example, see Henry & Muthén, 2010).

The parameters in the two-level latent class analysis were estimated by the maximum likelihood estimator with robust standard errors, with the combination of missing data analysis. The current analysis was carried out using Version 6 of the Mplus software program (Muthén & Muthén, 2010a).

The Modeling Process

In order to find the best-fitting latent class model, the modeling process followed an exploratory strategy (see, for example, Henry & Muthén, 2010b). The first step involved performing a one-level fixed-effect LCA using six SES measures: ASBHHEDUP, ASBHBOOK, ASBHCHBK, ASBHWELL, ASBHHOCP, and EDUAIDS. To determine the optimal number of latent classes, the LCA model was estimated stepwise, that is, a one-class LCA model was fitted first and an additional class was added sequentially until the model fitted the data well. These models were estimated by different sets of starting values to prevent any local maximum in the iteration processes and were compared with respect to the Bayesian information criterion (BIC). BIC is a popular

indicator for determining the number of latent classes among the information criterion measures (Schwartz, 1978; see also Kass & Wasserman, 1995). The model with the lowest BIC was preferred.

The next step involved evaluating the correct number of classes chosen for each model. The Vuong-Lo-Mendell-Rubin likelihood ratio test (LRT) (Lo, Mendell, & Robin, 2001) was used to compare the loglikelihood differences of the model with *k* classes with the one with *k*-1 classes. A significant improvement in loglikelihood difference implies that the *k*-class model fits the data better (Muthén & Muthén, 2010b; Nylund, Asparouhov, & Muthén, 2007). Moreover, the substantive meaningfulness of the latent classes is yet another factor to consider when determining the number of latent classes. If the latent class size is small, a substantive rationale based on previous research and theories has to be offered to support the inclusion of the class. This need arises because the extremely small group may be a statistical artifact that reflects only the measurement character in the sample (Samuelsen & Dayton, 2010).

In the next step, the single-level LCA model was extended to a two-level latent class model to capture the randomness in the latent class indictors at school level (i.e., variation in the intercepts of SES indicators across different schools). A continuous latent variable was specified by these aggregated means of SES indicators. A chi-square test of conditional mean equality was then carried out for a set of potential latent SES class predictors and reading achievement.

Finally, the covariate "use of test language at home" (ASBGLNG1) was introduced to the two-level latent class measurement model in order to control for differences in students' ethnic background. The reading achievement ASRREA01 was regressed on the covariate ASBGLNG1 and the latent class membership at individual level so that the latent class differences in reading achievement level could be examined. At school level, the aggregated reading achievement was regressed on the indicator-specific random effect factor at the school level so that the collective SES effect on reading achievement could be examined.

FINDINGS

Results from the Single-Level LCA Analysis

The LCA model was fitted at the individual level in order to classify students into possible latent groups. Table 3 shows the Bayesian information criterion (BIC) and entropy values associated with each of the LCA models. The BIC estimate decreased dramatically from the one-class LCA model to the four-class LCA model, and stabilized between the four-class and the five-class models. The BIC increased, however, for the six-class model, which indicated that the five-class model should be chosen as the preferred solution.

Table 3 also shows the *p*-values for the LRT test and the adjusted LRT test[1] that were obtained from the series of LCA models. The *p*-values for LRT and adjusted

1 That is, adjusting the conventional likelihood ratio test for *k* versus *k* +1 classes for violation of regularity conditions (see Lo et al., 2001).

Table 3: Evaluation of the single-level LCA measurement models of socioeconomic status

Fit criteria	Individual-level latent class models (complex model type with weight)					
	One-class model	*Two-class model*	*Three-class model*	*Four-class model*	*Five-class model*	*Six-class model*
No. of free parameters	12	25	38	*51*	64	77
Loglikelihood	- 22288	-20915	-20631	*-20488*	-20430	-20406
BIC	44677	42041	41581	*41405*	41399	41458
Comparison between		1- vs. 2-class	2- vs. 3-class	3- vs. 4-class	4- vs. 5-class	5- vs. 6-class
p-value for LRT		.0000	.0002	.0006	*.7409*	*.8150*
p-value for adjusted LRT		.0000	.0002	.0007	*.7414*	*.8150*

Note: The Vuong-Lo-Mendell-Rubin test (LRT) showed a significant difference in the loglikelihood estimation between the four-class and five-class model, thus leading to the choice of the four-class LCA model of the individual-level SES (see fit criteria in italic) even though the Bayesian information criterion for the four-class model was not the lowest among the estimated models.

LRT between the four- and five-class LCA models were non-significant, suggesting that four classes would be the sufficient number of latent groups of individuals with respect to the combined characteristics of the SES measures. Scrutiny of the BIC estimates of the four- and five-class LCA models revealed the difference as marginal. The size of the fifth class was rather small for the five-class solution, being about seven per cent of the total sample size (see Appendix 1). The inclusion of the extra class did not contribute more substantive meaning to the classification of the sample students. The four classes achieved from the LCA model were therefore taken as the optimal solution.

Item probabilities were estimated for each category of the SES measures, given the SES latent class membership. Table 4 shows the item probability of answering Response Category "3" for each measure of SES in the different latent classes. The SES indicators for Response Category 3 are "finished university education or higher," "more than 100 books at home," "more than 100 children's books at home," "being a well-off family," "highest parental occupation is professional," and "have five home-possession items" (see Table 1 on page 54 for detailed descriptions of the variables and labels).

For Latent Class 1, the conditional probabilities were very high for all six SES indicators, implying that the individuals belonging to this latent class very likely came from an economically and culturally affluent family. Latent Class 1 can therefore be termed the "economically and culturally affluent group." The opposite pattern was evident in Latent Class 2, where the item probability was low on all six SES indicators. The individuals in this group were very likely to have come from an economically and culturally disadvantaged family. This latent class can thus be termed the "economically and culturally disadvantaged group."

Table 4: The conditional probability of answering "Category 3" of the SES indicators estimated by the four-class single-level LCA measurement model

Variables	Label (Category 3)	Latent Class 1: Economically and culturally affluent group	Latent Class 2: Economically and culturally disadvantaged group	Latent Class 3: Economically well-off group	Latent Class 4: Culturally affluent group
ASBHHEDUP (C)	Parental educational level (finished university or higher)	0.65	0.06	0.34	0.03
SBHBOOK (C)	Books at home (more than 100 books)	1.00	0.02	0.34	0.59
ASBHCHBK (C)	Children's books at home (more than 100 books)	0.51	0.00	0.09	0.21
ASBHWELL (E)	Family well-off	0.65	0.40	0.55	0.39
ASBHHOCP (E)	Parental highest occupation (professional)	0.93	0.16	0.93	0.09
EDUAIDS (E)	Index of the level of educational aids at home (5, high)	0.74	0.39	0.68	0.62

Notes:

C and E in parentheses denote cultural capital indicator and economic indicator respectively.

Although the pattern of posterior probabilities of the economic capital indicator for Latent Class 3 was rather similar to the pattern for Latent Class 1, the former's cultural capital was much lower than the latter's. Latent Class 3 was thus named the "economically affluent group." Finally, Latent Class 4 can be seen as the "culturally affluent group" because the number of books, children's books, and educational aids in the home accounted for the distinction between Latent Classes 2 and 4. (For the graphical profiles of each latent class, see Appendix 2.)

Each student in the sample was assigned to one of the four latent classes according to the most probable likelihood of him or her belonging to that class. Class 1 contained the most students in the sample; in all, 1,700 students (38.9% of the total sample). The second largest group was Class 4, with 1,314 individuals (30% of the sample). The number of individuals in Classes 2 and 3 was smaller, being 609 (13.9%) and 755 (17.2%), respectively.

In summary, the single-level LCA achieved a four-class solution for the classification of students according to the response patterns of SES indicators. However, the stratified cluster sampling design in PIRLS 2006 implies that the data have a hierarchical structure (Martin, Mullis, & Kennedy, 2007). But ignoring the cluster effect in the data may lead to misclassification of individuals. A two-level LCA model is thus needed to correct such misclassification.

In Sweden, the increasing degree of residential segregation and the fact that more and more parents are determining which school their child should attend has led to an increasing incidence of cross-school differences in school SES composition (Björklund, Clark, Edin, Fredriksson, & Krueger, 2005; Yang Hansen, Cliffordson, & Gustafsson, 2010). Consequently, the collective SES, represented by the aggregated means of SES indicators (i.e., an indicator-specific random effect), becomes more homogeneous within each school and varies largely across different schools.

Results from the Two-Level Mixture Modeling

Determining the Structure of the Two-Level Latent Class Model

To adjust for the cluster effect of the hierarchical data structure and to capture the between-school difference in the aggregated means of the SES indicators, a set of two-level latent class models was estimated and evaluated. Various researchers, including Asparouhov and Muthén (2008) and Vermunt (2003), recommend identifying a common factor to represent the SES indicator-specific random effects between schools (i.e., the random means and associated covariance). The assumption here is that the random means are highly correlated and have different loadings on the common factor.

Table 5 presents the evaluation of the set of two-level latent class models, with the four-class single-level LCA model (Model 1) providing the reference. Model 2 is a two-level latent class model with four latent classes at the individual level and a continuous factor *fu* at the school level (Model 2). This model has a better BIC estimation compared to that of the single-level LCA model.

Table 5: Model evaluation for two-level LCA measurement models and two-level LCA models with a covariate and an outcome variable

	Model 1: Single-level four-class LCA	Model 2: Two-level LCA with cw4 & *fu*	Model 3: Two-level LCA with cw3 & *fu*	Model 4: Two-level mixture model with cw3, *fu*, x & y
No. of free parameters	51	57	44	57
Loglikelihood	- 20488	- 20063	- 20171	- 42587
BIC	41405	40604	40712	85650
Entropy	0.67	0.57	0.58	0.60
Lo-Mendell-Rubin test (LRT)	282.658	214.684	460.435	485.32
p-value for LRT	0.01	0.10	0.00	0.01
Test for H_0	3 vs. 4 classes	3 vs. 4 classes	2 vs. 3 classes	2 vs. 3 classes

Notes:

Model 1: Single-level LCA model with four latent SES classes;

Model 2: Two-level LCA with cw4 & *fu* = a two-level latent class model with four latent classes at individual level and a continuous factor *fu* at school level;

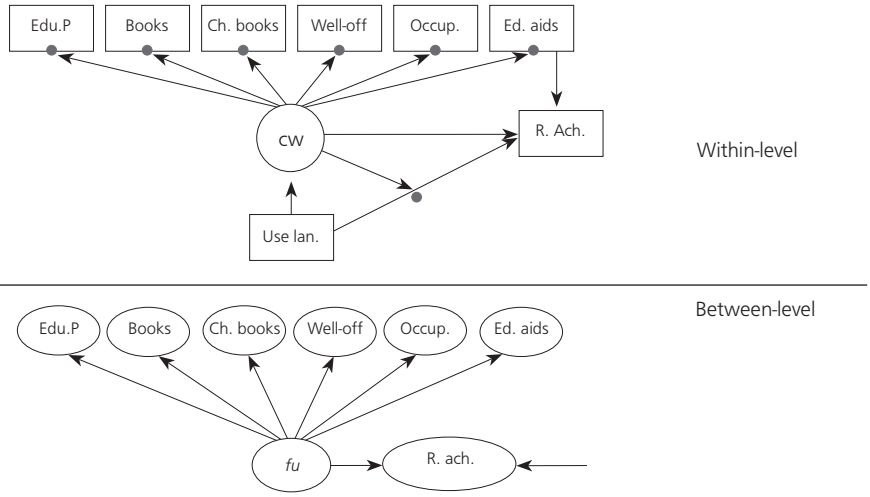
Model 3: Two-level LCA with cw3 & *fu* = a two-level latent class model with three latent classes at individual level and a continuous factor *fu* at school level;

Model 4: Two-level LCA with cw3, *fu*, x & y = a two-level mixture model with three latent classes at individual level and a continuous factor *fu* at school level.

An individual-level covariate x (use of the test language at home) and an outcome variable y (reading achievement) were also included in Model 4.

However, the non-significant *p*-value for LRT suggests that a three-class solution is satisfactory for classifying individuals, but only after the between-school differences in the collective SES of student intakes have been taken into account. A more parsimonious model was thus estimated, with a three-class latent SES variable at the individual level and the same between-school structure as in the previous model (Model 3). In the next step, Model 3 was extended to include an individual-level covariate "use of test language at home" (ASBGLNG1) and reading outcome (ASRREA01, Model 4). Figure 1 depicts the structure of the final two-level mixture model (see Appendix 2 for the Mplus model input).

Figure 1: Path diagram of the two-level mixture model with covariate and reading outcome (Model 4)



Notes:

Edu. P = Parental highest educational level (ASBHHEDUP); Books = number of books at home (ASBHBOOK); Ch. books = number of children's books at home (ASBHCHBK); Well-off = family affluence status, being well-off (ASBHWELL); Occup. = parental highest occupation (ASBHHOCP); Ed. aids = educational aids at home (EDUAIDS); R. Ach. = reading achievement; Use lan. = use of test language at home.

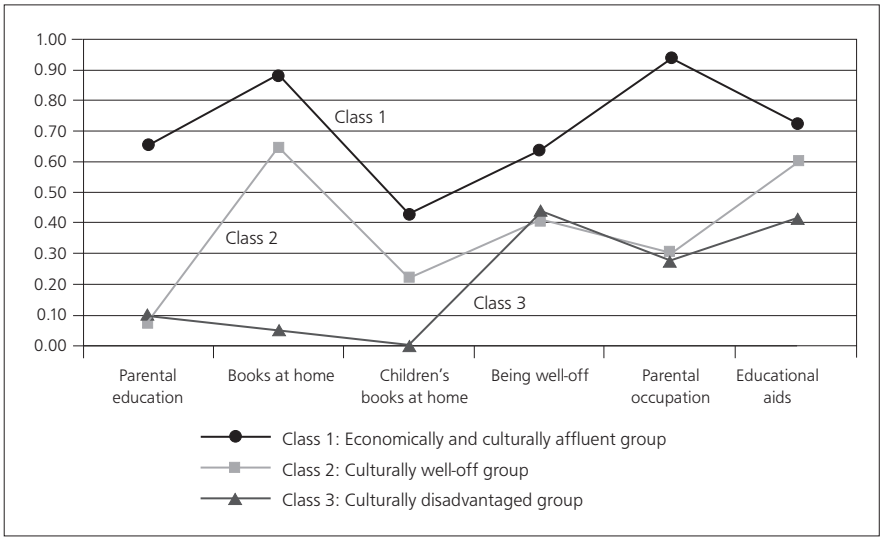
The filled dots under the SES indicators represent the randomness of aggregated mean across schools. *Fu* is a continuous latent variable reflecting the collective SES.

The continuous latent variable *fu* may be understood as the SES composite of all school intakes that captured the SES indicator-specific mean differences across schools and was assumed to affect average school reading achievement. In the within part of the model, the filled dots attached below the individual-level SES indicators represent the random means of the six indicators at the second-level unit, that is, the school. They were regarded as unobserved variables and are shown in circles in the school-level model. The within-level part of the model was relaxed, thus allowing for differences in ASBGLNG1 and reading achievement across latent classes, as well as the between-class differences in the impact of ASBGLNG1 on reading achievement ASRREA01. The latter was shown by the path from the latent class variable cw to the path between ASBGLAN1 to ASRREA01.

Latent Class Profiles Estimated from the Two-Level Latent Class Model

Three latent SES classes were chosen when comparing the fit of the different sequentially estimated two-level latent class models. The nature of these latent classes was determined by the posterior probabilities estimated for each of the SES indicators. Figure 2 depicts the latent class profiles based on item probabilities of students belonging to one of the latent SES classes when choosing the highest category in their response to the SES indicators.

Figure 2: Graphical description of the latent classes according to the profiles estimated in the two-level LCA model



Notes:

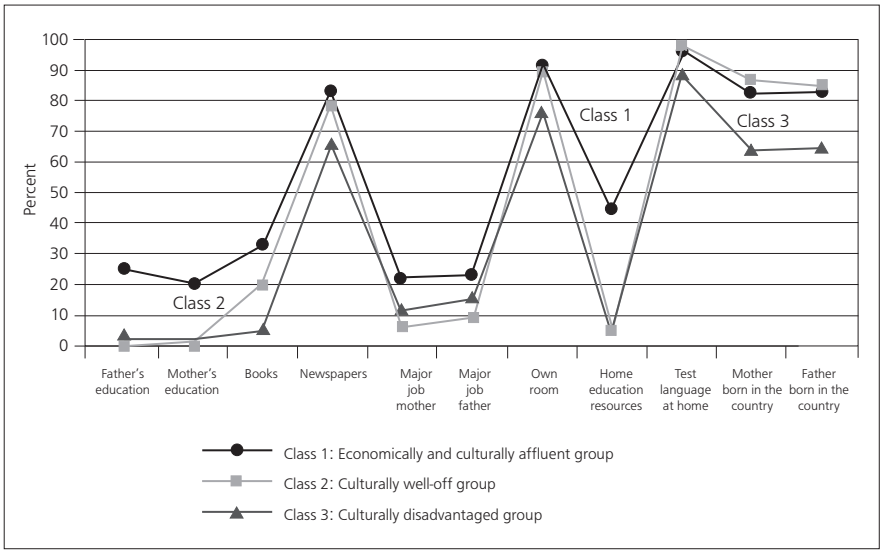
The y-axis represents the estimated mean probability of the third category (the highest category) in each given SES latent class indicator presented on the x-axis. The total number of individuals in each of the latent classes is as follows: Class 1: economically and culturally affluent group = 2,150 (49%); Class 2: culturally well-off group = 1,247 (28%); and Class 3: culturally disadvantaged group = 980 (22%).

It is evident in Figure 2 that the posterior probabilities of Class 1 were high in all SES indicators, meaning that the probability of being included in Class 1 was very high for those students whose parents graduated from university or gained a post-graduate qualification, had more than 100 books and children's books at home, were well-off and held a professional job, and could provide their children with basic educational resources. This latent class can thus be called the "economically and culturally affluent group." About half of the individuals in the sample belonged to this group (2,150 students).

Class 2 and Class 3 showed a fairly similar pattern with regard to the economic capital indicators and parental education level. However, Class 2 had relatively high probabilities for the books and educational aids at home variables. Class 3 had the lowest probabilities for almost all the SES indicators, especially with respect to the book and educational aids variables. Students in this class were therefore identified as the "culturally disadvantaged group" and students in Class 2 as the "culturally well-off group." There were 980 individuals in Class 2 (22% of the total student sample) and 1,247 (28% of the student sample) in Class 3.

So that the characteristics of the latent SES classes could be better understood, the proportions of individuals who chose the highest category in their response to a set of auxiliary variables (see Table 2) were plotted for each latent SES class. The outcome is shown in Figure 3.

Figure 3: Graphical description of students' characteristics in each of the latent SES classes by a set of background variables pertaining to their parents



Notes:

The y-axis represents the percentage of individuals who responded with the highest category in each given background variable presented on the x-axis. (For variable information, see Table 2.) The three latent classes referred to in this figure were estimated by the two-level LCA measurement model.

As shown in Figure 3, students in the economically and culturally affluent group (Class 1) were likely to be the ones who chose the highest category of the auxiliary variables. The proportion of such students is the highest among the three SES latent classes. The proportion of students in the culturally disadvantaged group (Class 3) is relatively low. However, of the students in Class 3, the percentages of individuals whose parents were born in the country and who spoke the test language at home were much lower than the corresponding percentages for the other two classes. This pattern suggests that the culturally disadvantaged group contained a large number of students with an immigrant background.

The pattern most distinguishing the culturally well-off group from the culturally disadvantaged group was the larger number of books that the former group had at home and the greater likelihood that these students had of being from a non-immigrant background. The low proportion of students in Class 2 (culturally well-off) with one or both parents having finished university and/or having a professional job may have reflected a large number of parents who began university at a relatively late age and who read a lot, both for themselves and to their children. Also, these parents may have still been in full-time study and/or engaged in only temporary or part-time employment. Being culturally well-off in Class 2 was relative to the SES pattern in Class 3. When compared to Class 1, however, the cultural capital, in terms of parental educational level and home educational resources, was much lower in Class 2.

The Wald chi-square test of conditional mean equality of the auxiliary variables across the latent SES classes confirmed the latent class profiles of the auxiliary variables in Figure 3. (For detailed Wald chi-square test results, see Appendix 4.) Reading achievement differed significantly among the three latent classes, with the economically and culturally affluent group achieving the highest average score on the reading test (565), the culturally disadvantaged group the lowest (528), and the culturally well-off group scoring at an intermediate level (544). Differences in conditional means between Class 2 and Class 3 in variables such as "mother's educational level," "father's educational level," and "mother's major job" were not significant, implying that these background variables were not strong latent class predictors, that is, able to distinguish Class 2 from Class 3. The same was true with respect to Class 1 and Class 2 for the variables denoting "use of the test language at home," "mother being born in the country," and "father being born in the country."

Typically, the greater the differences between these conditional means were across the different SES classes, the more likely it was that the predictors related to the latent SES classes (Muthén & Muthén, 2010). The SES indicators "parental highest educational level" and "parental highest occupation" in the current two-level latent class models were derived variables from mother's educational level and father's educational level, and from mother's and father's major job. Non-significant mean differences may increase the level of measurement error (i.e., misclassification in LCA) and therefore affect the quality of categorization of individuals—a possibility that may explain the low entropy level in the two-level latent class analysis, namely, 0.58 (see Table 5).

It should be noted that an auxiliary variable was not involved in the determination of the latent classes, but rather was used as a further description of the latent classes. The plot and Wald chi-square test, however, showed that reading achievement and ethnic background were two potential predictors of latent SES classes. Including these predictors may improve the precision of classification of individuals, thus improving the reliability in the estimates of SES effects on reading achievement at both individual and school levels. Therefore, a final two-level mixture model was developed by adding "use of test language at home" (ASBGLNG1) and reading achievement (ASRREA01) in the two-level latent class model (see Figure 1 and Appendix 3 for the model structure and Mplus input of the final model).

Results from the Two-level Mixture Model with a Covariate and Reading Outcome

In the final two-level mixture model, the latent class variable cw and reading achievement were regressed on the covariate ASBGLANG1 at the individual level in order to examine latent class differences in reading achievement, with ethnic background differences being controlled for. In order to take into account the between-school differences in SES and its effect on reading achievement, the reading achievement variable was regressed on the collective SES factor *fu* at the school level.

After the effects of the SES composite of the school intakes and the ethnic background of students had been accounted for, it was obvious that some of the individuals had moved from one latent SES class to another. This outcome implies that taking into account information on students' achievement levels and ethnic backgrounds further corrected misclassification of students. Compared to the distribution of individuals in the latent classes estimated by the previous two-level latent class model (Model 3), the proportion of individuals in the economically and culturally affluent group reduced from 49% to 43%, leaving 1,809 students in this group. The number of students in the culturally disadvantaged group dropped to 804, leaving only about 19% of the total student sample in this group. The culturally well-off group gained eight per cent of the individuals in the total sample, swelling the number of students in that group to 1,618. The entropy increased to a level of 0.60, indicating a more accurate classification. The move to another SES class for some individuals kept the nature of the SES latent classes intact, however.

Mean reading achievement differed significantly across the latent SES classes. The economically and culturally affluent group (Class 1) achieved the highest mean achievement score, 552, while the culturally disadvantaged group (Class 3) had the lowest, 499. For the culturally well-off group (Class 2), the average achievement in the reading test was at an intermediate level, 512. It is worth noting that the higher number of books and educational aids at home in Class 2 may function as compensation for the otherwise disadvantaged home background, therefore increasing the ability of students in this group to achieve a higher level of reading ability—13 points higher than the group of students with a similar level of SES but many fewer books (Class 3). It should also be emphasized that the three latent SES classes estimated in

this final model became more distinguishable in terms of the between-class differences in reading achievement, especially the differences between Class 1 and the other two latent classes, a finding that again provides evidence of better classification.

Table 6 presents the between-class differences for the individual-level covariate variable "use of test language at home" (ASBGLNG1) and for the effects of ASBGLAN1 on reading achievement. The language used at home variable indicates students' ethnic background and so has a different impact on the latent class variable cw, that is, a different odds ratio for belonging to a different latent SES group. When the culturally well-off group (Class 2) was set as the reference group, the chance of a student from a non-immigrant background belonging to the culturally disadvantaged group was slightly more than one tenth of the likelihood of that same student being in Class 2. The chance of a student from a nonimmigrant background belonging to the economically and culturally affluent group was about three fifths of the chance of him or her being in Class 2. After the between-group differences in ethnic background had been controlled for, the only significant impact of language used at home on reading achievement was found for the culturally disadvantaged group, with a correlation coefficient of 0.13.

Table 6: Latent class differences in the individual-level covariate ASBGLNG1 and the effect of the covariate on reading achievement in the two-level mixture model

	Relation	Odds ratio
Class 1: Economically and culturally affluent	cw#1 ON ASBGLNG1	0.64
Class 2: Culturally well-off	Reference group	1.0
Class 3: Culturally disadvantaged	cw#3 ON ASBGLNG1	0.12
	Relation	Regression coefficient
Class 1: Economically and culturally affluent	ASRREA01 ON ASBGLNG1	0.07 (ns)
Class 2: Culturally well-off	ASRREA01 ON ASBGLNG1	0.07 (ns)
Class 3: Culturally disadvantaged	ASRREA01 ON ASBGLNG1	0.13

Note: ns = not statistically significant. cw#1 and cw#3 are thresholds for Latent Class 1 and Latent Class 3, respectively. Latent Class 2 is the reference group.

At the individual level, about 16% of the reading achievement differences between individuals could be attributed to the different SES profiles across the latent classes, according to the effect size measure eta-square. The collective SES factor *fu* at the school level that captured the differences in the SES composite of school intakes, and which mirrored the degree of segregation in Swedish compulsory schools, had a considerable impact on school achievement. The estimated regression coefficient was 0.70, which implied that almost half of the between-school reading achievement differences could be explained by the differences in collective SES across different schools. This result reflects, to a large extent, the selection mechanism and segregation in Swedish compulsory schools (Skolverket, 1996, 2012; Yang Hansen, Rosén, & Gustafsson, 2011).

DISCUSSION AND CONCLUSIONS

The current study explored the latent profiles of individuals according to their socioeconomic background and examined reading achievement differences among these latent profiles by applying two-level latent class modeling techniques. Unlike the previous view of SES as simply being an observed index, the current measurement approach allows the formation of distinct and homogenous categories or typologies of individuals by conceptualizing SES as unobserved classes conditional upon its indicators. One advantage of two-level latent class modeling is that it can achieve a SES index variable by estimating latent groups through a set of SES indicators. It can also take into account the measurement error in SES indicators, the hierarchical data structure, and the variation of the SES indicators across collective-level units (schools in the case of this study). Accordingly, it can eliminate the misclassification of individuals in different observed categories of SES indicators and thus assure the precision of the estimated class membership of individuals. The categorical latent class variable of SES can be saved and used as an ordinary variable in further analysis, such as that pertaining to the impact of SES differences on academic achievement, attitude, and motivation. It can also be used to test the interaction between SES and other personal traits, such as gender.

Another advantage is that two-level mixture modeling can simultaneously examine SES effects on reading achievement at individual and collective levels. The way in which students were classified into model-based latent classes makes it possible to further explore and improve understanding of the effects of educational inequality and school segregation.

Three latent classes of individuals with different SES profiles were found in the present study, namely, the economically and culturally affluent class, the culturally well-off class, and the culturally disadvantaged class. Reading achievement differed significantly among the three latent SES classes, with the economically and culturally affluent group achieving the highest and the culturally disadvantaged group the lowest scores. It was evident that these latent groups could potentially be ordered in terms of the item probabilities decreasing from Class 1 to Class 2 to Class 3.

Student's ethnicity, indicated by whether or not students spoke the test language at home, was significantly related to reading achievement only for the culturally disadvantaged immigrant-concentrated class (Class 3). This finding implies that immigrant background does not affect the reading achievement of students in Classes 1 and 2 because the language ability of these students is boosted by their peers and surrounding environments, such as family, neighborhood, and school. For students in the disadvantaged group, however, such support may not have been available to the same extent as it was for students in Class 1 and Class 2. This finding raises concerns about the increasing segregation along socioeconomic lines and immigrant background in Sweden's schools and society (Skolverket, 2009, 2012). The concern rests on the premise that students in the disadvantaged group are not being exposed to an integrated learning environment where deleterious aspects of their home background may be compensated for by other cultural and educational resources.

At the individual level, about 16% of the reading achievement differences could be attributed to membership of the different latent SES classes, while the relationship between collective SES and reading achievement at school level reached 0.70. These estimates not only agree closely with the findings of meta-analyses of SES effects on academic achievement conducted by such researchers as Sirin (2003, 2005) and White (1982, 2005) but also confirm the results from previous SES effect studies (conducted by the likes of Yang Hansen et al., 2011).

Previous studies measuring SES and its effects on reading achievement, using data from IEA's Reading Literacy Study (Elley, 1994), identified two dimensions in SES at the level of the individual—a cultural capital dimension and an economic capital dimension at (Gustafsson, 1998; Yang & Gustafsson, 2004). In Sweden, the studies just cited found no significant relationship between economic capital and reading achievement. However, the cultural capital factor had a strong positive impact on reading achievement.

The three latent SES classes identified in the current study indeed reflect the two underlying dimensions and their relationship with reading achievement. The composition of the latent SES groups highly mirrors the level of cultural capital and its importance to reading achievement. Muthén (2001) observed that data that have a good fit with a *k*-class model often have a good fit with a *k*-1 dimensional factor analysis model, in terms of the LCA's ordering of the latent classes. An observation of the same relationship was also found in a latent profile analysis conducted by Bartholomew (1987). The empirical evidence from the current study provides yet another proof of this relationship.

One may argue that factor scores achieved by multilevel factor analysis can be used to divide individuals into groups of different SES levels. Moreover, compared to the two-level mixture model used in the current study, multilevel factor analysis has relatively less computation load and is much easier to converge. However, because a factor score does not have a natural cut point for SES groups, the division of SES groups seems to be more arbitrary. As Muthén concluded, latent class analysis is a better means than factor analysis of finding clusters of individuals (see, for example, Muthén, 2001).

The current analysis clearly shows that ignoring the cluster effect of data structure and multiple-level variation will bias the classification of individuals, whereas bringing in covariates and distal outcome variables to the two-level mixture model can further improve the quality of latent class estimation. However, including too many SES indicators in the model may cause a heavy computation and latent class patterns that are difficult to interpret. Selecting optimal SES indicators is therefore important, and this can be done with the help of a Wald chi-square test of mean equality. Results from a Wald test of a set of auxiliary variables in this study showed that ethnic background and reading achievement were two important sources of variation among the latent SES classes that were controlled for in the later stage of the two-level mixture model. The outcome was an improvement in the quality of the classification of individuals.

Unfortunately, the indicators of ethnic background in PIRLS 2006 were limited. "Use of test language at home," "father born in the country," and "mother born in the country" were the only available alternatives. As shown in the Wald test, however, these variables did not distinguish between the students in SES Class 1 and those in SES Class 2. This finding could be one of the reasons why the entropy level in the current study, although being improved, stayed at a relatively low level.

In order to gain a better understanding of this problem, and to gain confirmation about the feasibility and accuracy of the method of classifying students into unobserved SES groups, multiple countries with either similar (e.g., other Nordic countries) and/or rather different circumstances (e.g., England, Germany, and the United States) and societal and school characteristics need to be brought into the analysis.

Finally, a sensitivity study may be needed to test the stability of the results. In the current study, SES indicators were recoded into ordinal variables with three categories. It might be argued that the cut point of the recoding can affect the latent class estimation. Testing of the impact of recoding has been carried out during our analysis, using dummy variables of the SES indicators. The results were consistent with the findings presented in this paper. Further testing in which the SES indicators were treated as continuous variables in a latent profile analysis, in order to avoid the self-inflicted problem of having only a few categories in each variable, would also be useful.

Acknowledgements: This study was supported by the Center for Comparative Analyses of Educational Achievement (COMPEAT) at the Department of Education and Special Education (IPS), University of Gothenburg, financed by the National Bank of Tercentenary Foundation, Sweden. We would like to express our gratitude to Professor Jan-Eric Gustafsson and his research team FUR (Prerequisites, Education, and Results) at IPS for their valuable comments and advice on our analysis and report.

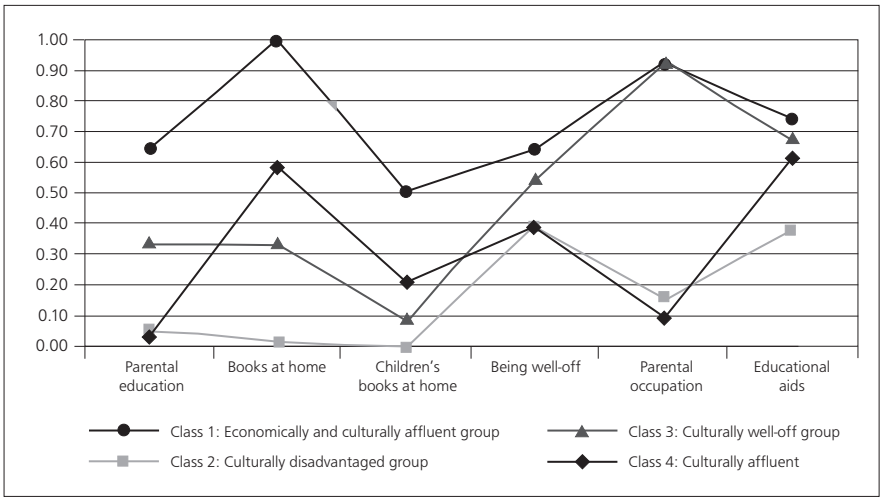
Appendix 1: Average latent class probabilities for most likely latent class membership (row) by latent class (column)

Four-class solution					
Subgroup of individuals belonging to:	Class 1	Class 2	Class 3	Class 4	Class size
Class 1	.859	.011	.086	.044	755
Class 2	.106	.827	.061	.006	1,700
Class 3	.049	.073	.807	.072	1,314
Class 4	.045	.012	.125	.818	609

Five-class solution						
Subgroup of individuals belonging to:	Class 1	Class 2	Class 3	Class 4	Class 5	Class size
Class 1	.773	.127	.027	.065	.007	1,522
Class 2	.010	.856	.097	.016	.021	813
Class 3	.000	.061	.842	.033	.064	877
Class 4	.045	.042	.098	.812	.003	828
Class 5	.017	.031	.162	.025	.766	339

The table shows a matrix of the average probabilities of the most likely latent class membership in each latent class, as well as the latent class size based on estimated posterior probabilities for both the four- and five-class solutions. For the subset of students with the most likely class membership of 1, for example, the posterior probabilities for each class are averaged, and presented in row 1 of Table 4. The average posterior probabilities should be high on-diagonal of the matrix, and the probabilities off-diagonal should be as low as possible, indicating precision of the latent class membership estimation. For the four-class solution, the average posterior probabilities on-diagonal are over .80, and the sizes of each latent class are rather large, being over 10% of the sample size. For the five-class solution, the average on-diagonal posterior probabilities are lower than .80 for Class 1 and Class 5. The size of Class 5 is relatively small, being about seven per cent of the total sample (4,393 individuals).

Appendix 2: Graphical depiction of the latent classes according to the profiles estimated in the single-level LCA model



Notes:

This is a graphic depiction of the conditional probabilities presented in Table 4. The y-axis represents the conditional probability level, and the x-axis represents the SES indicators. The first three variables on the x-axis are cultural capital indicators; the last three are economic capital indicators. The total number of individuals in each of the latent classes is as follows: economically and culturally affluent group = 1,700 (38.9%); culturally disadvantaged group = 609 (13.9%); economically well-off group = 755 (17.2%); culturally affluent group = 1,314 (30%).

Appendix 3: Mplus input of the final two-level mixture model with covariates and outcome variable

```
USEVARIABLES ARE u1-u6 y x;
        CATEGORICAL ARE u1-u6;
        MISSING ARE u1-u6 x(99);
        CLUSTER=IDSCHOOL;
        WEIGHT=HOUWGT;
            CLASSES = cw(3);
            WITHIN = x;
   ANALYSIS:   TYPE = TWOLEVEL MIXTURE;
            PROCESSORS = 4(STARTS);
            STARTS = 100 10;
            STITERATIONS = 20;
   MODEL:
            %WITHIN%
            %OVERALL%
           cw ON x (cw_x);
           cw#1 ON x;
           cw#2 ON x;
           y ON x;
           %cw#2%
           y
           y ON x;
           %cw#3%
           y ON x;
           y;
            %BETWEEN%
            %OVERALL%
            fu BY u3@1;
            fu BY u1;
            fu BY u2;
            fu BY u4;
            fu BY u5;
            fu BY u6;
            [fu@0];
            fu*0.043 (8);
            y ON fu;
    %cw#1%
    [u1$1 u1$2 u2$1 u2$2 u3$1 u3$2 u4$1 u4$2 u5$1 u5$2 u6$1 u6$2];
    %cw#2%
    [u1$1 u1$2 u2$1 u2$2 u3$1 u3$2 u4$1 u4$2 u5$1 u5$2 u6$1 u6$2 ];
    %cw#3%
    [u1$1 u1$2 u2$1 u2$2 u3$1 u3$2 u4$1 u4$2 u5$1 u5$2 u6$1 u6$2];

   OUTPUT: STANDARDIZED;
         TECH1 TECH11;
   SAVEDATA:
            FILE IS Model 4_cw3_fu_xy.dat;
                     SAVE = CPROBABILITIES;
```

Note:

Note: Variables u1 to u6 are the SES indicators ASBHWELL, ASBHHEDUP, ASBHBOOK, ASBHCHBK, ASBHHOCP, and EDUAIDS. "x" is the covariate at student level—"use of test language at home" (ASBGLNG1). The covariate at school-level "w" is "school intake characteristics" (INTAKECH). Outcome variable "y" is the standardized reading achievement ASRREA01

Appendix 4: Comparison of mean equality across latent SES classes by Wald chi-square test

Classes	Conditional mean									
	ASBHLEDF	ASBHLEDM	ASBGBOOK	ASBGTA4	ASBHMJF	ASBHMJM	ASBGTA5	ASBGLNG1	ASBGBRNF	ASBGBRNM
Class 1	4.9	5.3	3.8	0.83	8.1	8.8	0.91	0.96	0.83	0.83
Class 2	3.4	3.7	3.4	0.77	5.8	5.5	0.88	0.96	0.83	0.84
Class 3	3.4	3.6	2.8	0.70	6.2	5.8	0.80	0.90	0.66	0.66
Classes	*p*-values for Wald test of mean equality									
Overall	*	*	*	*	*	*	*	*	*	*
1 vs. 2	*	*	*	*	*	*	*	ns	ns	ns
1 vs. 3	*	*	*	*	*	*	*	*	*	*
2 vs. 3	ns	ns	*	*	*	ns	*	*	*	*

Classes	Conditional mean
	ASREA01
Class 1	565
Class 2	544
Class 3	528
Classes	*p*-values for Wald test of mean equality
	ASREA01
Overall	*
1 vs. 2	*
1 vs. 3	*
2 vs. 3	*

Note:

For more information on the auxiliary variables, see Table 2. Most of these variables have more than five ordinal categories or are dummy variables. They are taken as continuous variables in the Wald test of mean equality.

ns = not significant; * *p*-value < 0.05.

References

Asparouhov, T. M. B., & Muthén, B. O. (2008). Multilevel mixture models. In G. R. Hancock & K. M. Samuelsen (Eds.), *Advances in latent variable mixture models* (pp. 27–51). Charlotte, NC: Information Age Publishing.

Bartholomew, D. J. (1987). *Latent variable models and factor analysis*. New York, NY: Oxford University Press.

Björklund, A., Clark, M., Edin, P.-E., Fredriksson, P., & Krueger, A. B. (2005). *The market comes to education in Sweden: An evaluation of Sweden's surprising school reforms*. New York, NY: Russell Sage Foundation.

Bourdieu, P. (1984). *Distinction: A social critique of the judgment of taste*. Cambridge, MA: Harvard University Press.

Bourdieu, P. (1997). The forms of capital. In A. H. Halsey, H. Lauder, P. Brown, & S. A. Wells (Eds.), *Education: Culture, economy, and society* (pp. 46–58). Oxford, UK: Oxford University Press.

Bourdieu, P. (1977). Cultural reproduction and social reproduction. In J. Karabel & A. H. Haley (Eds.), *Power and ideology in education* (pp. 487–511). New York, NY: Oxford.

Bourdieu, P., & Boltanski, L. (1979). Changes in social structure and changes in the demand for education. *Information Sur Les Sciences Sociales*, *XII*(October), 61–113.

Bourdieu, P., & Passeron, J.-C. (1977). *Reproduction in education, society and culture*. Beverly Hills, CA: Sage.

Bradley, R. H., & Corwyn, R. F. (2002). Socioeconomic status and child development. *Annual Review of Psychology*, *53*(1), 371–399.

Coleman, J. S., Campbell, E. Q., Hobson, C. F., McPartland, A. M., Mood, A. M., Weinfield, F. D., & York, R. L. (1966). *Equality of educational opportunity*. Washington, DC: U. S. Government Printing Office.

Clark, S. L., & Muthén, B. (2009). *Relating latent class analysis results to variables not included in the analysis*. Unpublished manuscript, University of California, Los Angeles.

Elley, W. B. (1994). *The IEA study of reading literacy: Achievement and instruction in thirty-two school systems*. Oxford, UK: Pergamon Press.

Gustafsson, J.-E. (1998). Social background and teaching factors: Determinants of reading achievement at class and individual levels. *Journal of Nordic Educational Research*, *18*, 241–250.

Hagenaars, J., & McCutcheon, A. (Eds.). (2002). *Applied latent class analysis models*. New York, NY: Cambridge University Press.

Henry, K. L., & Muthén, B. (2010). Multilevel latent class analysis: An application of adolescent smoking typologies with individual and contextual predictors. *Structural Equation Modeling: A Multidisciplinary Journal*, *17*(2), 193–215.

Kass, R. E., & Wasserman, L. A. (1995). A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion. *Journal of the American Statistical Association*, *90*, 928–934.

Lamont, M., & Lareau, A. (1988). Cultural capital: Allusions, gaps, and glissandos in recent theoretical developments. *Sociological Theory*, *6*(2), 153–168.

Lareau, A., & Weininger, E. B. (2003). Cultural capital in educational research: A critical assessment. *Theory and Society*, *32*(5/6, Special issue on the sociology of symbolic power: A special issue in memory of Pierre Bourdieu), 567–606.

Lo, Y., Mendell, N. R., & Rubin, D. B. (2001). Testing the number of components in a normal mixture. *Biometrika*, *88*(3), 767–778.

Martin, M. O., Mullis, I. V. S., & Kennedy, A. M. (Eds.). (2007). *PIRLS 2006 technical report*. Chestnut Hill, MA: Boston College.

Mullis, I. V. S., Martin, M. O., Kennedy, A. M., & Foy, P. (2007). *PIRLS 2006 international report: IEA's Progress in International Reading Literacy Study in primary schools in 40 countries*. Chestnut Hill, MA: Boston College.

Muthén, B. O. (1989). Latent variable modeling in heterogeneous populations. *Psychometrika*, *54*, 557–585.

Muthén, B. O. (1991). Multilevel factor analysis of class and student achievement components. *Journal of Educational Measurement*, *28*, 338–354.

Muthén, B. O. (2001). Latent variable mixture modeling. In G. A. Marcoulides & R. E. Schumacker (Eds.), *New developments and techniques in structural equation modeling* (pp. 1–33). Mahwah, NJ: Lawrence Erlbaum Associates.

Muthén, B. O. (2007). Latent variable hybrids: Overviews of old and new models. In G. R. Hancock & K. M. Samuelsen (Eds.), *Advances in latent variable mixture models* (pp. 1–24). Charlotte, NC: Information Age Publishing.

Muthén, L. K., & Muthén, B. O. (2010a). *Mplus user's guide* (6th ed.). Los Angeles, CA: Muthén & Muthén.

Muthén, L. K., & Muthén, B. O. (2010b). Wald test of mean equality for potential latent class predictors in mixture modeling. *Mplus technical appendices* [website: http://www.statmodel.com/techappen.shtml].

Nylund, K. E., Asparouhov, T., & Muthén, B. O. (2007). Deciding on the number of classes in latent class analysis and growth mixture modeling: A Monte Carlo simulation study. *Structural Equation Modeling*, *14*(4), 535–569.

Robinson, W. S. (1950). Ecological correlations and behavior of individuals. *American Sociological Review*, *15*(3), 351–357.

Samuelsen, K. M., & Dayton, M. C. (2010). Latent class analysis. In G. R. Hancock & R. O. Mueller (Eds.), *The reviewer's guide to quantitative methods in the social sciences*. New York, NY: Routledge.

Schwartz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*, 461–464.

Sirin, S. R. (2003). *The relationship between socioeconomic status and school outcomes: Meta analytic review of research, 1990–2000*. College Hill, MA: Boston College.

Sirin, S. R. (2005). Socioeconomic status and academic achievement: A meta-analytic review of research 1990–2000. *Review of Educational Research*, *75*(3), 417–453.

Skolverket (National Agency for Education). (1996). *Att välja skola—effekter av valmöjligheter i grundskolan* [Choosing a school: Effects of free choice possibilities in elementary schools] (No. 109). Stockholm, Sweden: Author.

Skolverket (National Agency for Education). (2009). *Vad påverkar resultaten i svensk grundskola? Kunskapsöversikt om betydelsen av olika faktorer* [What influenced the results in the Swedish compulsory school?]. Stockholm, Sweden: Author.

Skolverket (National Agency for Education). (2012). *Likvärdig utbildning i svensk grundskola? En kvantitativ analys av likvärdighet över tid.* [Equal education in the Swedish compulsory school? A quantitative analysis of equality over time]. Stockholm, Sweden: Author.

Vermunt, J. K. (2003). Multilevel latent class models. *Sociological Methodology*, *33*, 213–239.

von Davier, M. (2010). Hierarchical mixtures of diagnostic models. *Psychological Test and Assessment Modeling*, *52*(1), 8–28. Available online at http://www. psychologie-aktuell. com/ fileadmin/download/ptam/1-2010/02_vonDavier.pdf

White, K. R. (1982). The relation between socioeconomic status and academic achievement. *Psychological Bulletin*, *91*, 461.

Willms, J. D. (1992). *Monitoring school performance: A guide for educators*. London, UK: The Falmer Press.

Yang, Y. (2003). *Measuring socioeconomic status and its effects at individual and collective levels: A cross-country comparison* (Gothenburg Studies in Educational Sciences, 193). Gothenburg, Sweden: Acta Universitatis Gothoburgensis.

Yang, Y., & Gustafsson, J.-E. (2004). Measuring socioeconomic status at individual and collective levels. *Educational Research and Evaluation*, *10*(3), 259–288.

Yang Hansen, K., Cliffordson, C., & Gustafsson, J.-E. (2010). *Changes in the variances of school grades and SES effect between school and individuals: A multilevel analysis*. Paper presented at the European Conference on Educational Research, August, 25–27, 2010, Helsinki, Finland.

Yang Hansen, K., Rosén, M., & Gustafsson, J.-E. (2006). Measures of self-reported reading resources, attitude and activities based on latent variable modeling. *International Journal of Research and Method in Education*, *29*(2), 221–237.

Yang Hansen, K. Rosén, M., & Gustafsson, J.-E. (2011). Changes in the multi-level effects of socio-economic status on reading achievement in Sweden 1991 and 2001. *Scandinavian Journal of Educational Research*, *55*(2), 197–211.

Leadership, learning-centered school conditions, and mathematics achievement: What can the United States learn from top performers on TIMSS?

Nianbo Dong and Xiu Chen Cravens
Vanderbilt University, Nashville, Tennessee, United States

Drawing on crossnational datasets, including contextual questionnaires and the mathematics achievement results of eighth graders, from the Trends in International Mathematics and Science Study of 2007 (TIMSS 2007), this study examines the viability of using the findings of international assessment reports to inform school leadership practice directed at enhancing learning conditions. The study first identifies core school conditions within the realm of influence that can be captured by the TIMSS conceptual framework, and then examines the association between these conditions and student achievement in mathematics. We focus on education systems whose eighth graders consistently gain higher average scores than their counterparts in the United States on mathematics assessments. These systems are Chinese Taipei, Korea, Singapore, Hong Kong SAR, and Japan. Analyses of the 2007 contextual survey results from the five education systems and the United States reveal interesting differences in school-level conditions for learning and in the associations between these conditions and mathematics achievement. Our initial crossnational analyses indicate links between achievement and learning conditions such as evaluation of the instruction curriculum and instructional implementation, and learning culture. However, the types and strengths of the associations appear to vary according to national context. Findings also indicate discrepancies between the perspectives of teachers and principals regarding school learning conditions.

IERI Monograph Series: Issues and Methodologies in Large-Scale Assessments Volume 5

INTRODUCTION

Around the world, governments are paying increasing attention to the results of large-scale crossnational assessments of student achievement, and are using the findings to inform educational reform initiatives. In the United States, studies of student performance on international assessments are playing an ever more important role in setting reform agendas at both state and national levels (Swanson & Barlage, 2006). In 2008, the National Governors Association (NGA), the Council of Chief State School Officers (CCSSO), and Achieve Inc. formed an advisory group to develop "international benchmarking" as a critical tool for creating a world-class education system for United States students. *Benchmarking for Success*, the report issued by the advisory group (NGA, 2008), cautioned that the United States "is falling behind other countries in the resource that matters most in the new global economy: human capital" (p. 5). Drawing together results from international assessments conducted during recent decades, the report called on both state and federal policymakers to provide stronger support for research and development in order to identify and learn from "top performers and rapid improvers," and thereby gain insights and ideas unlikely to be "garnered solely from looking within and across state lines" (p. 6).

Recent large-scale international assessments, such as the Trends in International Mathematics and Science Study (TIMSS) and the Programme for International Student Assessment (PISA), reveal United States students lagging behind their peers in other systems, especially those in East Asia, with respect to mathematics and science achievement (Ferraro & Van de Kerckhove, 2006; Mullis, Martin, & Foy, 2008; Provasnik, Gonzales, & Miller, 2009). For mathematics, in particular, the TIMSS 2007 results found that while the United States fourth-graders and eighth-graders scored, on average, above the international-scale and the US TIMSS 1995 averages, the performance of both cohorts was below that of their peers in several other systems. The fourth graders were outperformed by peers in eight of the 36 participating systems (Hong Kong SAR, Singapore, Chinese Taipei, Japan, Kazakhstan, Russian Federation, England, and Latvia) and the eighth graders by five of the 48 participating systems (Chinese Taipei, Korea, Singapore, Hong Kong SAR, and Japan; Provasnik et al., 2009).

The increasing recognition of the relevance that international assessments have for school improvement initiatives has intensified research interest in analyzing and gaining insight from the crossnational data now made publically available by the testing organizations (Rutkowski, Gonzalez, Joncas, & von Davier, 2010). In order to offer educational stakeholders a fuller appreciation of what the achievement results mean and how they can be used to improve student learning, international assessments such as TIMSS and PISA also incorporate contextual questionnaires designed within the context of theoretical frameworks that aim to capture attributes associated with student learning and achievement (Mullis et al., 2008).

The developing emphasis on studying the major characteristics of educational and social contexts with a view to improving student learning has led to crossnational

assessments gradually moving away over recent years from simple descriptions of contextual differences toward identification of factors that are malleable—changeable—and therefore meaningful to those involved in the process. In short, new research questions are being raised that differ from those asked and explored in much of the existing body of work (Baker, Goesling, & LeTendre, 2002; Chudgar & Luschei, 2009; Heyneman & Loxley, 1983).

Today, researchers want to know *what* makes a difference in student achievement in terms of input into schooling. Is it, for example, national income, school characteristics such as size and resources, or student socioeconomic background? They also want to explore how the schooling process affects learning outcomes. The purpose of the international benchmarking is seen as that of informing the development of domestic interventions—programs, practices, policies—that positively influence education outcomes. As such, the between- and within-country variations in achievement results and other factors arising out of the international assessments provide a rich opportunity for researchers engaged in comparative and international education studies to explore plausible associations between learning conditions and student performances (Baker, Lee, & Heyneman, 2003; Porter & Gamoran, 2002; Schmidt, Rotberg, & Siegel, 2003). Learning conditions define the contexts within which student learning takes place. More specifically, they are the factors that affect students' learning, such as national curriculum standards, resource allocation schemes for schools, classroom instructional approaches, teacher qualifications and professional development, student attitudes, and home support for learning.

The contextual questionnaires of the international assessments also afford opportunities for researchers to examine learning conditions from multiple levels and angles. At the national policy level, studies have examined curricular goals of education systems and how the systems were organized to attain those goals (Baker et al., 2003; Schmidt et al., 2003). Included in this level are, for example, curriculum standards, the rigor and coherency of textbook content, and characteristics of the teaching force (Akiba, LeTendre, & Scribner, 2007; Wang, 2004). At the school level, studies have looked into school characteristics in terms of student composition and resourcing, classroom activities, and pedagogical practices (Clarke et al., 2007; Wang & Lin, 2005). At the student level, research has explored the role that home support, parental involvement, and student attitudes toward learning play in academic achievement (Paik, 2004; Shen, 2005; Wang, 2004).

Conspicuously missing from this line of international assessment literature is the connection between school-level leadership and conditions of student learning. Extensive research on school leadership and educational outcomes identifies school principals as the keystone of successful educational reform; they are critical with respect to the successful interpretation and implementation of and support for improvement interventions (Elmore, 2000; Leithwood, Louis, Anderson, & Wahlstrom, 2004; Rice & Islas, 2001). However, because school leaders impact core learning conditions indirectly and holistically (Leithwood et al., 2004; Louis, Leithwood, Wahlstrom, & Anderson, 2010; Waters, Marzano, & McNulty, 2003), constructing a theoretical

framework to capture empirical evidence for the linkage between school leadership and student learning is challenging within any learning context, let alone cross-nationally.

As a reaction to the less than optimal performance of United States fourth and eighth graders in TIMSS 1999, the National Association of Secondary School Principals (NASSP) called for stronger instructor leadership on mathematics and science performance (Rice & Islas, 2001). The 2008 NGA report also underscored the importance of developing school leaders and holding principals accountable for ensuring instructional quality by learning from "international best practices" (p. 28). However, despite the strong rhetoric for action, existing international assessment research has yet to provide evidence that contextual questionnaires offer substantive insight into the school-level practices that are positively associated with student achievement.

In this paper, we examine high-performing education systems through the prism of TIMSS in search of valuable lessons for the United States. We focus on the learning conditions that are likely to be within the realm of influence of school leadership. Specifically, we focus on education systems that have consistently produced higher TIMSS average scores than those of the United States in mathematics—Chinese Taipei, Korea, Singapore, Hong Kong SAR, and Japan.

To present our study, we first provide a twofold background review of (a) international assessment research addressing learning conditions and educational outcomes, and (b) leadership research that identifies core components of learning conditions malleable at the school level by principals and their leadership teams. We then examine the TIMSS contextual questionnaires for the extent to which they capture the core learning conditions. We additionally explore, both theoretically and empirically, the possible associations between core components of learning-centered school conditions and student achievement outcomes, by using TIMSS datasets from 2007 that include contextual-questionnaire and mathematics-assessment results for the participating eighth graders. We ask the following research questions:

1. How do core components of learning-centered school conditions compare among the high-performing systems and with the United States as described by the TIMSS contextual questionnaires?
2. Are core components of learning-centered school conditions associated with student achievement in mathematics when contextual factors are held constant? If yes, how and to what extent?
3. Are the relationships between the key malleable variables and student achievement consistent across the six systems when within-system contextual factors such as curriculum standards, attitudes toward learning, school size, teacher qualifications, and student socioeconomic status are held constant?

BACKGROUND

International Assessments and Learning Conditions

Research on learning conditions can be categorized into two main dimensions: school resourcing (personnel selection, qualifications, and training) and the schooling process (curriculum content, teaching pedagogy, teacher collaborations, and classroom structure). In recent decades, crossnational studies utilizing international assessment results to connect these two dimensions of learning conditions with student achievement have been conducted mainly at the national and aggregate levels. The learning conditions that have attracted the most attention in these studies appear to comprise four areas: (a) curriculum standards, rigor, and coherence, (b) teacher qualifications, (c) pedagogical strategies, and (d) home support for learning.

From their analysis of data from the first TIMSS dataset (1995), Schmidt and his colleagues found strong relationships between curricular content and learning outcomes both across countries and across classrooms within countries, especially in the United States (Schmidt et al., 2003). They suggested that much of the poor performance of the United States students could be attributed to a poorly constructed curriculum. Specifically, from their comparison of the curriculum standards of the best-performing nations, Schmidt and his team identified three essential characteristics that the United States standards were lacking (Schmidt et al., 2003):

- *Focus:* covering a smaller number of topics in greater depth at every grade level, enabling teachers to spend more time on each topic so that all students learn it well before they advance to more difficult content;
- *Rigor:* demanding more advanced learning in subjects such as algebra and geometry; and
- *Coherence:* laying out an orderly progression of topics that follow the logic of the discipline, allowing thorough and deep coverage of content.

Although subsequent studies using TIMSS data yielded less convincing associations between curriculum foci and cross-national achievement variation (Baker et al., 2003; LeTendre, Baker, Wiseman, Boe, & Goesling, 2002), the Schmidt study of early TIMSS results served as a sounding alarm that drew national attention to the weakness of having a splintered curriculum and was widely cited as evidence for the necessity of developing coherent and consistent curriculum standards.

Improving teacher quality is another educational reform priority for the United States and other nations. Using TIMSS 2003 data from 46 systems, Akiba et al. (2007) tested the assumption that teacher quality, measured by certification rate, mathematics major, and teaching experience, is associated with student achievement. The authors also examined the association between access to qualified teachers and the socioeconomic status (SES)-based achievement gap. The authors found from their analyses that the achievement gap in the United States between high-SES and low-SES students was among the largest when compared with the relevant data from other nations that participated in TIMSS 2003. However, they also found that the gap

in access to qualified teachers was not significantly associated with the achievement gap between students with high and low SES. Their research left room for further investigation into other influential factors, such as instructional resources and teacher learning opportunities, that might help account for the student achievement variation unexplained by the teacher qualification measures alone.

Analyses of teacher questionnaire items regarding classroom activities and video archives provided data for deeper probes into instructional practices and pedagogical strategies. Givvin, Hiebert, Jacobs, Hollingsworth, and Gallimore (2005) conducted an ethnographic study that drew on the TIMSS 1999 video archives. They used three coding dimensions when analyzing the data—purpose of activity, interaction structure, and content activity. Their findings suggested that within the seven systems (Australia, the Czech Republic, Hong Kong SAR, Japan, the Netherlands, Switzerland, and the United States) that participated in the TIMSS 1999 Video Study, eighth-grade mathematics teachers within a country taught lessons in relatively similar ways. They also found that many of the features within the three dimensions examined were discernible in all seven systems (Givvin et al., 2005).

These findings supported the suggestions made by others (e.g., LeTendre, Baker, Akiba, Goesling, & Wiseman, 2001) that countries share patterns of teaching practice. Such convergence, according to Givvin and colleagues, provides opportunities for educators not only to share familiar notions regarding classroom practices but also to realize that seeing "the familiar in a new light might offer many opportunities for teachers to rethink taken-for-granted practices and to see them as choices rather than inevitabilities" (2005, p. 342). This view was expanded by the research of David Clarke and colleagues, which emphasized the complexity of international and comparative learning associated with instructional practices (Clarke et al., 2007). In differentiating the choices of instructional unit for analysis, they revealed significant structural variation in any one teacher's lesson sequence, suggesting that a single lesson pattern is unlikely to be an accurate or a useful representation of either an individual teacher's lessons or of any nationally representative sample of lessons. Furthermore, Clarke et al. (2007) raised important questions about the assumption that less-successful countries would necessarily do well to adapt instructional practices of countries consistently successful on international measures of mathematics performance, given that variations in student performance might be attributable to other differences in culture, societal affluence, or aspiration.

The search for other factors influencing student learning beyond and/or in connection with classroom instructions has led researchers to explore the role of home support, parental involvement, and student attitudes toward learning in academic achievement (Paik, 2004; Shen, 2005; Wang, 2004). For example, using TIMSS 1995 data, Wang (2004) compared the mathematics achievement of students from Hong Kong SAR with that of their peers from the United States. Wang also looked at a series of family background factors, such as mothers' expectations, presence of study aids, and extracurricular time spent on various activities. Finding that some of the factors differentially influenced the Hong Kong and the United States students, Wang

conjectured that the differences could be culture-dependent. Paik (2004) used a multiple-factor psychological model to analyze the home and school factors in TIMSS 1995 for Korean and United States students. The findings of the study suggested implications for family–school partnerships, after-school or weekend programs targeted at improving academic competencies, and family support, given that all appeared to be influential factors for learning.

Shen (2005) used multivariate discriminant analysis and data from TIMSS 1999 to make comparisons between the United States middle school system and five top-ranked Asian middle school systems, with respect to student achievement in mathematics and science. The analyses were based on variables related to school and classroom environment as well as students' out-of-school life, home background, and self-perceptions about mathematics and science ability. The results further illuminated the differences between American schools and their Asian counterparts, especially the somewhat more peripheral place of schooling in the lives of American adolescents versus the more central position of schooling in the lives of their East Asian peers.

Malleable Learning Conditions and School Leadership

Despite the complexity and difficulty of identifying factors and conditions that can be optimized for student learning, research suggests that strategic actions which integrate core components of school-wide improvement efforts are essential in terms of effective reform efforts for student learning (Desimone, 2006; Goldring, Porter, Murphy, Elliott, & Cravens, 2009; Rowan, Correnti, Miller, & Camburn, 2009). Furthermore, a relatively extensive research base supports the notion that school leaders play a pivotal role in interpreting, implementing, and sustaining such intervention measures by enhancing teaching and learning conditions (Hallinger & Heck, 2010; Leithwood et al., 2004; Waters et al., 2003). But why is leadership crucial? Louis and colleagues conjectured, on the basis of their six-year leadership study, that "leaders have the potential to unleash latent capacities in organizations" (Louis et al., 2010, p. 7). They pointed out that while most school variables, considered separately, have only small effects on student learning, it is possible to create synergy across the relevant variables operating among the key stakeholders in the process. Educators in leadership positions, as Louis and her colleagues asserted, are uniquely well positioned to ensure the creation and sustainability of this synergy.

Intervention-oriented research, therefore, focuses on understanding the nature of strong leadership and, more specifically, of identifying the pathways through which leadership affects learning conditions. Research in recent decades suggests that school leaders impact student learning by establishing school conditions which support and strengthen teaching and learning (Waters et al., 2003). Much of the evidence shows, however, that the connection is indirect and complex. Leadership studies suggest that the direct and indirect effects of school leadership on student learning are small but significant at about five to seven percent of the variation in student learning across schools, or about one quarter of the total across-school variation (12 to 20%) explained by all school-level factors after controlling for student characteristics (Creemers & Reezigt, 1996; Louis et al., 2010). Further empirical evidence suggests

that school principals, along with their leadership teams, influence student outcomes by enhancing curriculum structures and instruction practices as well as providing academic support for parents and students (Cohen & Hill, 2000; Goldring & Cravens, 2007; Leithwood & Jantzi, 1999; Smith, Desimone, & Ueno, 2005).

In preparation for reauthorization of the Elementary and Secondary Education Act, the U.S. Department of Education issued *A Blueprint for Reform*, which called on "states and districts to develop and implement systems of teacher and principal evaluation and support, and to identify effective and highly effective teachers and principals on the bases of student growth and other factors" (U.S. Department of Education, 2010, p. 4). While the term "effective" in conjunction with leadership or teaching is often treated with considerable skepticism, a comprehensive review of the research literature (Goldring et al., 2009; Porter, Goldring, Muphy, Elliott, & Cravens, 2006) reveals six core components of leadership that are highly effective with regard to student learning and achievement: holding high standards for student performance, a rigorous curriculum, quality instruction, a culture of learning and professional behavior, connections to external communities, and systemic performance accountability.

The learning-centered leadership framework has three strong features that provide the scaffold for empirical research designed to detect malleable school conditions for learning. First, the focus of this framework is on measureable leadership behaviors drawn from literature on effective schools and school districts. This framework fits within a more general leadership model (CCSSO, 1996; Glasman & Heck, 1992; Hallinger & Heck, 1996) of what qualifications school principals must have and how principals in the school system are expected to perform, but it does not try to address every aspect of the overall leadership process. The framework focuses on principal behaviors that are linked to teachers' opportunities to improve instructional practices. Not included in the framework are other aspects of leadership that are considered to be the precursors of leadership behaviors, such as knowledge and skills, personal characteristics, and beliefs (Murphy, Goldring, Elliott, & Porter, 2006). Second, the core components include standards, curriculum, instruction, culture, external environment, and performance accountability, which rest upon the same theoretical foundation as that of the international-assessment contextual frameworks (Mullis, Martin, Smith, Garden, Gregory, & Gonzalez, 2005). Third, the learning-centered core components assume that there are aspects of the context within which leadership and schooling take place that might moderate the impact of leadership effects. For example, everything else being equal, the evaluation of leadership quality might appropriately take into account systemic curriculum standards, experience of leadership, length of time in the same school, student body composition, staff composition, level of schooling, and the geographic setting of the school.

In summary, our review of literature indicates that, to date, while a substantial number of studies have been conducted using large-scale international assessment data (e.g., Rutkowski et al., 2010) and while some have explored the connection between school

contexts and student achievement, few have drawn on the available cross-national datasets to address the role of school leaders in improving learning conditions. Even fewer have applied theoretically-grounded frameworks and sophisticated analytical methodology.

METHOD

Data

TIMSS 2007 was the fourth administration (since 1995) of international benchmarking of Grade 4 and Grade 8 student achievement in mathematics and science undertaken by the International Association for the Evaluation of Educational Achievement (IEA). In 2007, 36 educational jurisdictions participated in the Grade 4 testing and 48 participated in the Grade 8 testing. Participating systems administered the TIMSS assessments to two system-wide probability samples of schools and their students, based on a standardized definition. Countries were required to draw samples of students who were nearing the end of their fourth year or eighth year of formal schooling. The sample included both public and private schools, randomly selected and weighted to be representative of the nation. Achievement results from TIMSS are reported on a scale that has a scale average of 500 and a standard deviation of 100 (Gonzales et al., 2008).

Although assessment results were available for both Grades 4 and 8 for this current study, we decided to focus on Grade 8 for several reasons. Middle school grades are considered critical to formative adolescent development and cognitive learning yet are also considered to be, within the context of school management, the grades most susceptible to adverse academic and social factors (Cobb & Smith, 2005). Furthermore, because elementary and secondary schools in most education systems have separate and often different administrative structures, focusing on learning conditions at the secondary school level (here, Grade 8) may add to the clarity of the findings.

Six education systems featured in our crossnational study: Chinese Taipei, Korea, Singapore, Hong Kong SAR, Japan, and the United States. Our choice of the first five systems was based on the fact that they had statistically higher overall average mathematics scores than the other systems, including the United States, participating in TIMSS in 2007 (see Table 1). The same five education systems also had consistently higher average mathematics scores than the United States in TIMSS 1999 and TIMSS 2003.

The dependent variable for the analyses was five plausible values of mathematics score for each student from TIMSS 2007. The independent variables were derived from the accompanying contextual questionnaires administered to the school principals, teachers, and students sampled in each system in 2007 and will be further discussed in this paper as measures for learning-centered and malleable school conditions.

Table 1: Weighted mean national mathematics achievement scores on TIMSS 2007

Education system	Mean	School *N*	Student *N*
Chinese Taipei	597	143	3,830
Korea	597	144	4,072
Singapore	592	155	4,351
Japan	571	141	4,151
Hong Kong SAR	569	106	3,040
United States	508	192	5,859

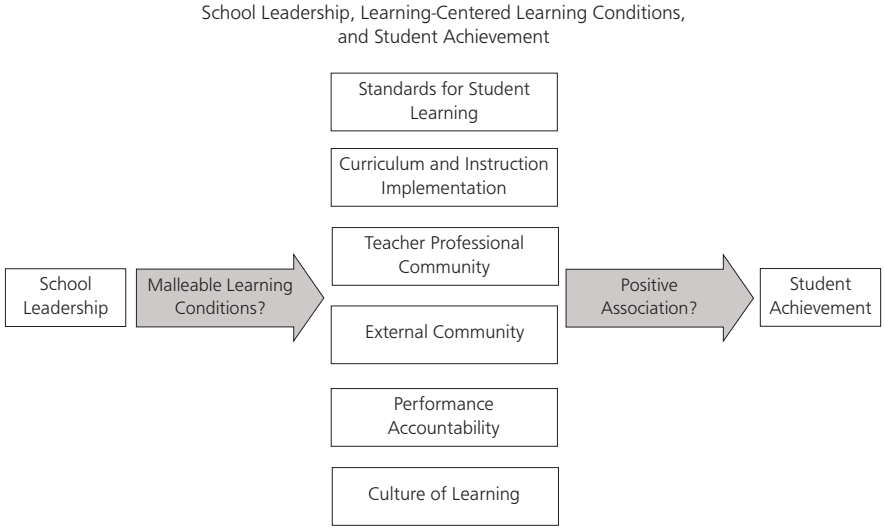
Source: TIMSS 2007 Mathematics Grade 8 database.

We handled missing data by first using listwise deletion for the observations with missing data in all key variables. We then applied the multiple imputation technique to impute the missing data. For the principal and teacher surveys, we imputed once. For student data, we imputed once based on each of five plausible values of the student mathematics score. We next merged principal, teacher, and student data, a process that gave us five imputed datasets for final analysis. The final dataset included a sample of 25,303 students nested in 881 schools of six education systems.

Our construction of measures for learning-centered school conditions using TIMSS 2007 contextual questionnaires was a three-step process. First, we studied the contextual framework document issued by IEA (Mullis et al., 2005) for intended construct domains to be covered by the school, teacher, and student questionnaires. Second, we combed through the questionnaires and identified items that met the face validity criteria for measuring dimensions of learning-centered school conditions. That is, we compared the items listed in the TIMSS conceptual framework as covering school learning conditions with the descriptions provided by the learning-centered leadership framework, and thereby formed the initial variables for the study. Third, based on preliminary exploratory factor analysis and the theory that we reviewed, we grouped selected variables into an analytic framework containing six measures of learning-centered school conditions: (a) standards for student learning, (b) curriculum and instruction implementation, (c) teacher professional community, (d) external community, (e) performance accountability, and (f) culture of learning (see Figure 1).

Our analytic framework essentially modified the core components of the Goldring et al. (2009) learning-centered leadership model, based on the literature review of international comparisons of student performance. We treated *culture of learning* as a stand-alone component in order to capture the importance of learning motivation in each national context. This consideration is especially relevant to the fact that all five systems with higher TIMSS average mathematics scores than those of the United States are East Asian and, as suggested in the literature, share a strong meritocratic cultural background that values education (Paik, 2004; Shen, 2005). We decided to merge and capture the core components of curriculum rigor and instructional quality with *curriculum and instruction implementation*, given that the available items in the TIMSS school and teacher questionnaires tended to address the two dimensions in a combined manner. Specific TIMSS questionnaire items and how they contributed

Figure 1: Conceptual framework



to each measure, whether independently or as parts of response scales, will be fully discussed in the results section, where we examine the extent to which TIMSS contextual questionnaires captured the learning-centered school condition framework. The items included and their coding can be found in Appendix Tables 1 and 2.

Noteworthy at this point is the fact that the TIMSS 2007 school and teacher questionnaires included a number of identical questions regarding school learning conditions. Thus, for the six main learning-centered conditions, we included these items to form four unique scales that allowed us to compare the perceptions of the principals and teachers in relation to

1. Teachers' expectations for student achievement;
2. Teachers' understanding of the curriculum goals and degrees of success in implementing the curriculum within the school;
3. Teachers' professional development in the school; and
4. Parents' and students' desire to do well in school.

We included control variables for background factors of schools, teachers, and students (Table 2). We carefully selected these variables after reference to the literature addressing the impact of available school resources, student SES, and home conditions (Baker et al., 2002; Heyneman & Loxley, 1983; Mullis et al., 2005). School characteristics refer to school-level background and composition factors that are typically considered as given conditions and therefore not malleable. In our study, these conditions included grade enrollment size, type of community (urban or rural), percentage of students from economically disadvantaged backgrounds, number of students in the class tested by TIMSS, and the percentage of students who took the

mathematics assessment in their native language. Teacher characteristics included whether the teacher majored in mathematics, whether he or she was a certified teacher, gender, and years of teaching. Lastly, we included the student characteristics to account for the predictive association between SES and achievement: gender, language, home resources (calculator, computer, desk, dictionary, Internet, number of books), father's education, and mother's education.

We also controlled, to a limited extent, for instructional structure, including ability grouping, use of mathematics textbooks, minutes per week for mathematics teaching, and the amount of homework for mathematics per week. These instructional practice variables, which could be considered within the realm of influence of school management and leadership, are, we consider, more closely connected with classroom-level teacher practice and therefore outside the scope of this paper with its focus on school-level leadership and conditions.

Analytic Strategies

Our analytic strategies encompassed three steps. During Step 1, we tested the construction of key school-level variables based on the learning-centered leadership framework in an iterative process that accounted for both theoretical justification (validity) and internal reliability. We started by grouping all the TIMSS questionnaire items that might cover the construct domain of learning-centered school conditions based on the original construct maps for the TIMSS contextual questionnaires (Mullis et al., 2008). We then calculated the Cronbach's alpha to test the internal consistency of these constructs by system. In some cases where the clustering pattern had the potential to mask the nuanced differences among various learning-centered conditions, we calculated the internal consistency of each subscale separately. For example, the TIMSS 2007 school and teacher questionnaires contain eight items that measure "school climate for learning" (National Center for Education Statistics [NCES], 2008). We created three subscales to measure teachers' curriculum understanding and implementation (two items), teachers' expectations for achievement (one item), and culture of learning (four items). Items that did not cluster but fitted in the theoretical domains were also identified and grouped accordingly. We continued this process until we had examined all items of the two questionnaires for schools and teachers.

During Step 2, we aggregated teacher-level data to the school level. The ideal situation would be for teacher level to serve as a separate level from student level and school level, such that we could investigate independent teacher-level effects in a three-level hierarchical linear model (HLM). However, the small numbers of teachers (classrooms) per school in our dataset limited our options to conduct three-level HLM analyses. For example, Hong Kong SAR and Korea had only one mathematics teacher per school in the TIMSS 2007 data, Chinese Taipei had one mathematics teacher in 97.7% of its schools, and Japan had one mathematics teacher in 83.7% of its schools. Singapore had only one mathematics teacher in 0.7% of its schools, and the United States had one mathematics teacher in 6.3% of its schools. To maintain the comparability of our analysis results across the six selected education systems, we decided to aggregate

Table 2: Weighted means and standard deviations of control variables

Control variables	Chinese Taipei		Hong Kong SAR		Japan		Korea		Singapore		United States	
School level												
Grade 8 enrollment	588.08	(337.46)	192.58	(30.78)	157.57	(71.45)	368.9	(151.97)	312.73	(74.74)	253.43	(175.06)
Type of community	4.58	(1.10)	5.11	(0.84)	4.60	(1.16)	5.29	(0.94)	6.00	(0.00)	3.43	(1.53)
Economically disadvantaged	1.57	(0.81)	2.89	(1.07)	1.55	(0.72)	2.35	(1.03)	1.81	(0.95)	2.88	(1.06)
Tested in native language	2.80	(1.20)	3.82	(0.57)	4.00	(0.00)	4.00	(0.00)	1.45	(0.88)	3.43	(0.98)
Teacher qualifications												
Math major	0.85	(0.36)	0.79	(0.38)	0.86	(0.30)	0.95	(0.20)	0.79	(0.30)	0.69	(0.38)
Female	0.56	(0.50)	0.43	(0.47)	0.42	(0.45)	0.63	(0.43)	0.63	(0.38)	0.70	(0.38)
Teaching certificate	0.98	(0.14)	0.98	(0.14)	0.99	(0.08)	0.99	(0.08)	0.98	(0.12)	0.96	(0.15)
Years in teaching	12.12	(8.23)	13.18	(9.38)	15.46	(8.50)	13.57	(8.83)	8.34	(7.96)	13.98	(8.67)
Ability grouping	0.15	(0.36)	0.42	(0.50)	0.30	(0.46)	0.76	(0.43)	0.34	(0.47)	0.74	(0.44)
Classroom context												
Use math textbooks	0.93	(0.25)	1.00	(0.05)	0.98	(0.11)	0.97	(0.15)	0.90	(0.25)	0.92	(0.23)
Number of students in TIMSS class	34.79	(6.35)	36.31	(8.66)	33.29	(7.67)	36.7	(5.20)	37.28	(4.17)	22.35	(6.63)
Minutes/week for math teaching	238.32	(56.13)	247.27	(63.48)	157.43	(21.91)	181.34	(13.60)	218.56	(29.00)	247.77	(67.42)
Amount of homework per week	2.22	(0.83)	2.57	(0.70)	1.81	(0.82)	1.76	(0.75)	2.49	(0.62)	2.75	(0.58)
Length of homework per week	2.47	(0.73)	2.38	(0.52)	1.98	(0.74)	2.14	(0.64)	2.43	(0.65)	2.10	(0.47)
Student characteristics												
Female	0.47	(0.14)	0.50	(0.24)	0.50	(0.12)	0.48	(0.34)	0.49	(0.23)	0.50	(0.12)
Speaks language of test at home	2.33	(0.39)	2.63	(0.29)	2.93	(0.07)	2.68	(0.15)	1.62	(0.41)	2.67	(0.33)
Possesses calculator	0.98	(0.03)	0.99	(0.02)	0.98	(0.03)	0.96	(0.04)	0.99	(0.02)	0.96	(0.05)
Possesses computer	0.94	(0.07)	0.98	(0.03)	0.88	(0.09)	0.99	(0.02)	0.93	(0.07)	0.94	(0.07)
Possesses study desk	0.89	(0.09)	0.77	(0.12)	0.95	(0.06)	0.96	(0.04)	0.88	(0.09)	0.84	(0.09)
Possesses dictionary	0.98	(0.03)	0.98	(0.03)	0.99	(0.03)	0.99	(0.02)	0.98	(0.03)	0.91	(0.07)
Possesses Internet connection	0.89	(0.11)	0.97	(0.05)	0.77	(0.14)	0.96	(0.04)	0.86	(0.12)	0.86	(0.12)
Mother's education	2.24	(0.73)	1.90	(0.58)	2.97	(0.43)	2.90	(0.55)	2.37	(0.70)	3.38	(0.81)
Father's education	2.48	(0.82)	2.15	(0.69)	3.32	(0.57)	3.46	(0.62)	2.68	(0.78)	3.32	(0.84)
Number of books at home	2.92	(0.55)	2.49	(0.58)	2.97	(0.41)	3.46	(0.43)	2.81	(0.49)	2.95	(0.62)

teacher-level data at the school level. However, we needed to keep in mind the possibility of aggregation bias (Snijders & Bosker, 1999).

In situations where more than one teacher per school was surveyed, we calculated the school means of teacher-level variables. For the constructs, which were the same in the principal and teacher questionnaires, we calculated their correlations. The small correlations (usually $r < 0.30$) indicated that the principal- and teacher-reported constructs were different enough to be analyzed. This approach allowed us not only to focus on the effects of the school-level variables in our analysis but also to handle the complicated data structure arising out of more than one teacher teaching students in the same class in some schools.

During Step 3, we constructed the two-level hierarchical linear model (HLM) with students nested within schools to account for clustering data structure (Raudenbush & Bryk, 2002). This allowed us to examine the associations among the explanatory variables for learning-centered school conditions and student achievement, while controlling for covariates at the school, classroom, and student levels. We conducted the analysis by running the unconditional model that did not include any covariates and the conditional (full) model that included all explanatory and control variables formed from the student, teacher, and school questionnaires. We then calculated the intra-class correlations (ICCs) for the unconditional and conditional models, and the percentages of variance explained by the explanatory and control variables. We used the SAS PROC MIXED procedure to conduct the weighted HLM analysis.

Because we had five imputed datasets that corresponded to five plausible dependent variables, which we created using multiple imputation (Foy & Olson, 2009), we ran the models five times with five plausible variables, respectively, then used the SAS PROC MIANALYZE procedure to summarize the results. The detailed two-level HLM was as follows:

Level 1 (student): $y_{ij} = \beta_{0j} + \beta_{xj} X_{ij} + e_{ij} \quad e_{ij} \sim N(0, \sigma^2)$

Level 2 (school): $\beta_{0j} = \gamma_0 + \gamma_w W_j + u_j \quad u_j \sim N(0, \tau^2)$

$\beta_{xj} = \gamma_x$

The reduced model was:

$$y_{ij} = \gamma_0 + \gamma_w W_j + \gamma_x X_{ij} + u_j + e_{ij}$$

where y_{ij} is Grade 8 overall mathematics achievement in the IRT (item response theory) scale for student *i* in school *j*, and where W_j is a vector of school-level covariates for school *j*. X_{ij} is a vector of the Level-1 covariate for student *i* in school *j*. γ_w are the coefficients of school-level covariate, W_j, and γ_x are the coefficients of the Level-1 covariate, X_{ij}.

Our primary interests lay in interpreting the coefficients of the school-level explanatory variables. For the purpose of simplification, we used the fixed effects of student-level variables across schools. We then compared the means of explanatory variables and their coefficients, which represented the average effects of the explanatory

variables on students' mathematics achievement across the selected systems. Given the cross-sectional nature of the TIMSS results and the contextual questionnaires, our methodological strategies addressed issues raised by Rutkowski et al. (2010) on causality claims, sampling, weights, proficiency estimates, imputed values, and generalization.

RESULTS

We first addressed this research question: How do the core components of learning-centered school conditions compare cross-culturally as described by the TIMSS contextual questionnaires? Our analyses of questionnaire content and construct factor patterns provided initial evidence that a set of variables reflecting the learning-centered leadership framework for school conditions might be identified with sufficient internal reliability using TIMSS teacher and school questionnaire items in the six selected systems. Some measures, however, appeared to have low internal reliabilities and some appeared to lack content validity due to limited coverage of the construct domain by the available questionnaire items.

Measures for learning-centered school conditions grouped into six sets:

1. Standards for learning;
2. Curriculum and instruction implementation;
3. Teacher professional community;
4. Parental involvement;
5. Performance accountability; and
6. Culture of learning.

Table 3 provides descriptive statistics for the key explanatory variables. There are several noteworthy findings. Under the main measure of c*urriculum and instruction implementation*, principals' direct efforts were measured by two items on the percentage of time spent on (a) "instructional leadership" and (b) teaching. The descriptive results show that the principal-reported time spent on activities related to instructional leadership varied across the six education systems. The United States average of 23.6% was lower than that of Korea (25.8%) and Chinese Taipei (25.0%), but was higher than that of Japan (22.9%) and Hong Kong SAR (19.9%).

The 2007 TIMSS questionnaire defines instructional leadership as "e.g., developing curriculum and pedagogy." The term may have been perceived very differently by the responding principals, however. In recent decades, the research-based focus on instructional leadership has led to the development of conceptual frameworks and instruments that aim to capture the complexity of this construct domain. For example, the widely used Principal Instructional Management Rating Scale (PIMRS; Hallinger, 1990, 2011; Hallinger & Murphy, 1985), which has 10 subscales and 50 items, proposes three dimensions to the role of an instructional leader—defining the school's mission, managing the instructional program, and promoting a positive school learning climate.

Table 3: Weighted means, standard deviations, and Cronbach's alphas for factors: explanatory variables from 2007 TIMSS contextual surveys

Explanatory variables	Chinese Taipei		Hong Kong SAR		Japan		Korea		Singapore		United States	
	Mean (*SD*)	Alpha	Mean (*SD*)	Alpha	Mean (*SD*)	Alpha	Mean (*SD*)	Alpha	Mean (*SD*)	Alpha	Mean (*SD*)	Alpha
Learning-centered leadership framework												
1. Standards for student learning												
Principal-reported expectation	4.11 (0.66)		3.80 (0.71)		3.53 (0.64)		3.81 (0.75)		3.84 (0.70)		3.97 (0.73)	
Teacher-reported expectation	3.91 (0.74)		3.58 (0.72)		3.45 (0.71)		3.69 (0.68)		3.62 (0.69)		3.95 (0.70)	
2. Curriculum and instruction												
Principal-reported curriculum rigor (2 items)	4.07 (0.47)	0.70	3.88 (0.46)	0.69	3.57 (0.57)	0.70	3.89 (0.53)	0.73	3.86 (0.49)	0.65	3.99 (0.63)	0.79
Teacher-reported curriculum rigor (2 items)	3.98 (0.55)	0.73	3.57 (0.56)	0.80	3.36 (0.54)	0.74	3.63 (0.49)	0.83	3.61 (0.48)	0.71	3.89 (0.59)	0.83
Time in instructional leadership (%)	25.00 (11.78)		19.91 (7.88)		22.86 (9.43)		25.83 (12.67)		21.35 (11.29)		23.56 (12.94)	
Time in teaching (%)	7.59 (10.1)		3.56 (7.84)		6.87 (8.24)		12.17 (12.67)		1.91 (2.92)		3.02 (6.62)	
3. Professional community												
Principal-reported PD (5 items)	2.99 (0.86)	0.87	3.55 (0.84)	0.86	2.84 (0.85)	0.77	2.77 (0.79)	0.86	4.25 (0.76)	0.86	4.09 (0.93)	0.89
Teacher-reported PD (6 items)	0.70 (0.29)	0.77	0.66 (0.31)	0.82	0.48 (0.27)	0.69	0.37 (0.28)	0.78	0.72 (0.22)	0.69	0.71 (0.27)	0.78
Teacher-reported collaboration (4 items)	1.67 (0.44)	0.64	1.81 (0.41)	0.60	1.95 (0.54)	0.60	1.73 (0.32)	0.54	1.83 (0.36)	0.63	1.77 (0.52)	0.73
4. External community												
Parental involvement (5 items)	0.77 (0.21)	0.48	0.78 (0.22)	0.46	0.59 (0.18)	0.16	0.62 (0.19)	0.26	0.83 (0.18)	0.32	0.92 (0.15)	0.37
5. Performance accountability												
Observation by principal	0.64 (0.48)		0.97 (0.17)		0.92 (0.27)		0.93 (0.26)		0.99 (0.08)		0.98 (0.14)	
Observation by external inspection	0.12 (0.32)		0.31 (0.47)		0.62 (0.49)		0.57 (0.50)		0.08 (0.28)		0.31 (0.46)	
Evaluated with student achievement	0.76 (0.43)		0.80 (0.40)		0.58 (0.50)		0.89 (0.32)		0.97 (0.16)		0.73 (0.45)	
Teacher peer review	0.33 (0.47)		0.70 (0.46)		0.50 (0.50)		0.83 (0.37)		0.53 (0.50)		0.27 (0.44)	
Incentive to recruit or retain teachers	0.06 (0.24)		0.05 (0.21)		0.18 (0.38)		0.10 (0.31)		0.14 (0.35)		0.06 (0.24)	
6. Culture of learning												
Principal's perception of parent and student desire to do well (4 items)	3.72 (0.61)	0.81	3.33 (0.59)	0.76	3.25 (0.56)	0.71	3.34 (0.63)	0.76	3.38 (0.57)	0.79	3.29 (0.72)	0.85
Teacher's perception of parent and student desire to do well (4 items)	3.11 (0.66)	0.84	2.98 (0.63)	0.81	3.02 (0.63)	0.71	2.88 (0.59)	0.78	2.93 (0.66)	0.85	2.93 (0.78)	0.90

On average, the Korean principals were those (from across the systems) who were most likely to be engaged in teaching (12.7%). The principals from Singapore appeared to be those least engaged in direct teaching (1.9%). The United States principals reported spending 3.2% of their time in direct teaching. Time spent in teaching could be affected by the size of school and the subject-matter training of the principal. Whether a principal carries a teaching workload may also depend on the career pathways of school personnel. For example, Singapore schools offer three separate tracks for career advancement, such that a teacher can aspire to be a master teacher, an administrator, or an instructional specialist with the Ministry of Education (Tucker, 2011).

The descriptive differences under *performance accountability* are interesting. When the principals were asked whether (yes or no) each of the four methods for evaluating teacher practice was being used in their schools, more than 95% of them in five of the systems (the exception was Chinese Taipei at 64%) said that they or senior staff used observations. However, external inspection practices appeared to be done very differently among the systems selected, with the range extending from rarely done in Singapore (8%) and Chinese Taipei (12%) through to being somewhat more regularly done in the United States (31%), and on to being considerably more common in Japan (62%) and Korea (57%). As many as 83% of the principals from Korea and 70% from Hong Kong SAR reported utilizing teacher peer review to evaluate mathematics practices, and more than 50% in Singapore and Japan reported the same. The United States principals reported the lowest occurrences (27%).

Table 3 also includes Cronbach's alphas for the variables constructed from the multiple questionnaire items. The alphas were calculated separately for each system. Overall, the constructed variables demonstrated high internal consistency, with the alphas ranging from 0.60 to 0.90. However, the internal consistency for the five items that formed the parental involvement scale appear to be in question across the six systems, given the consistently low Cronbach's alphas, which ranged from 0.16 in Japan to 0.48 in Chinese Taipei, and with the United States at 0.37. The five items (attend special event, raise funds, volunteer for school projects, ensure homework is completed, and serve on a school committee) may be covering very different construct dimensions of parental involvement and so do not tap into a common domain.

There were also discernible differences between the perceptions of the principals and the perceptions of the teachers across the six systems on the four sets of scales formed from the common items in the school and teacher questionnaires. With respect to teachers' expectations for student achievement in the school, principals reported higher average ratings (on a scale from 1 = very low to 5 = very high) than the teachers of the sampled classes in all five East Asian systems, while the average ratings were about the same in the United States. On teachers' understanding of curriculum goals and their degree of success in curriculum implementation, principals consistently reported higher ratings than the teachers for the sampled classes. The largest difference between principal and teacher perceptions was for the importance that learning held for parents and students. Here, teacher ratings (also on a scale

from 1 = very low to 5 = very high) were significantly lower than the principal ratings across all six systems.

Table 4 shows the principal–teacher correlations for the four scale means (expectation, curriculum, professional development, and culture of learning). Among the six systems, teachers' and principals' viewpoints on learning culture appeared to be relatively highly correlated (correlations ranging from 0.27 in Korea to 0.60 in the United States). However, correlations on the other three measures were much lower. For example, correlations ranged from a low of 0.14 in Chinese Taipei to 0.38 in Singapore on teachers' expectations for student achievement, and from a negative 0.09 in Singapore to 0.28 in the United States. Given the low correlations, we decided to include the principal and teacher scale measures separately in the HLM analysis.

Table 4: Correlations of principals' and teachers' perceptions

Variable	Chinese Taipei	Hong Kong SAR	Japan	Korea	Singapore	United States
Expectation	0.14	0.30	0.26	0.17	0.38	0.25
Curriculum	0.07	0.20	0.30	0.13	0.24	0.35
Professional development	0.00	0.07	0.06	-0.03	-0.09	0.28
Culture of learning	0.36	0.36	0.42	0.27	0.53	0.60

We then addressed the second and third research questions regarding the associations between core components of learning-centered school conditions and student achievement in mathematics. The intra-class correlations (ICCs) and percentages of variance explained by the explanatory and control variables are reported in Table 5. The ICCs for the unconditional model varied from 0.09 to 0.64, with a mean of 0.33 across the six education systems, indicating considerable variation in the proportion of between-school mathematics achievement across the six systems; a high average of 33% of variation was due to between-school variation. The percentages of between-school variance explained by the explanatory and control variables varied from 54.5% to 74.6%, with a mean of 64.8% across the six systems, which meant they were relatively stable. These results suggest that these variables could account for more than 50% of between-school variation.

The fixed results of the full HLM (Table 6) show that some learning-centered school conditions were associated with student achievement in mathematics when background factors were held constant. However, which element of a learning-centered school condition and the extent to which the element was significant appeared to vary by national context. We will explain each condition specifically.

Table 5: Variances, ICC, and percentage of variance explained by the explanatory and control variables

Model	Variance and ICC	Chinese Taipei	Hong Kong SAR	Japan	Korea	Singapore	United States
Unconditional model[a]	Between school variance	2428.10	5892.37	1405.48	794.71	4108.09	2078.35
	Within school variance	8895.11	3351.38	5805.35	7690.23	4546.53	3787.35
	Total variance	11323.21	9243.75	7210.83	8484.94	8654.62	5865.70
	ICC	0.21	0.64	0.19	0.09	0.47	0.35
Conditional model[b]	Between school variance	807.52	2683.75	451.38	202.22	1451.65	816.61
	Within school variance	7646.14	3235.91	4923.08	6335.18	4116.13	3491.82
	Total variance	8453.66	5919.67	5374.46	6537.40	5567.78	4308.43
	ICC	0.10	0.45	0.08	0.03	0.26	0.19
Percentage variance explained[c]	Between school	66.7	54.5	67.9	74.6	64.7	60.7
	Within school	14.0	3.4	15.2	17.6	9.5	7.8
	Total variance	25.3	36.0	25.5	23.0	35.7	26.5

Notes:

a Unconditional model is the two-level HLM without any covariates.

b Conditional model is the two-level HLM including all variables in Tables 1 and 2.

c Percentage of variance explained was calculated by 100*(1-variance in conditional model/variance in unconditional model).

High Standards for Student Learning

Among the six selected systems, one point of increase in teachers' expectations on student achievement was positively associated with 7.77 points of increase ($p < 0.10$) in Japan and 22.63 points of increase ($p < 0.01$) in Singapore on the mathematics assessment scores. As Rutter and Jacobson (1986) have suggested, the perceptions that teachers have of student ability might affect their engagement in teaching and school improvement. Betts and Grogger (2003) found that, on average, higher grading standards were associated with higher Grade 12 test scores. However, among the selected systems, the positive associations were only statistically significant with respect to the teacher-reported measure, and for these two systems only.

Rigorous Curriculum and Instruction

We found that teachers' understanding of curriculum goals and their success in curriculum implementation at the school level did not appear to be strongly associated with the TIMSS 2007 Grade 8 students' mathematics achievement results. This finding held for both principal-reported and teacher-reported measures. In fact, there was a negative and statistically significant association between principal-reported curriculum-instruction implementation and student achievement results in Chinese Taipei (-12.97, $p < 0.10$). It will be interesting to further disentangle this association in terms of school, teacher, and student characteristics within the Chinese Taipei system.

Table 6: Learning-centered school conditions and mathematics achievement: fixed effects from two-level HLM

Explanatory variables	Chinese Taipei		Hong Kong SAR		Japan		Korea		Singapore		United States	
Learning-centered leadership framework												
1. Standards for student learning												
Principal-reported expectation	4.16	(5.93)	0.51	(10.43)	3.33	(4.7)	-4.00	(3.27)	2.46	(6.99)	4.65	(4.87)
Teacher-reported expectation	-2.90	(5.30)	5.91	(9.80)	7.77^{+}	(4.42)	4.66	(3.44)	22.63**	(8.04)	-3.60	(4.66)
2. Curriculum and instruction												
Principal-reported curriculum rigor	-12.97^{+}	(5.75)	22.15	(15.38)	5.08	(5.23)	4.47	(4.32)	-8.22	(9.85)	-2.88	(5.62)
Teacher-reported curriculum rigor	0.35	(6.78)	-1.44	(12.31)	-1.26	(5.05)	-6.17	(4.65)	9.31	(9.9)	-1.98	(5.08)
Time in instructional leadership (%)	0.24	(0.27)	-0.49	(0.88)	-0.54*	(0.26)	-0.01	(0.16)	-0.36	(0.32)	-0.25	(0.20)
Time in teaching (%)	0.15	(0.31)	0.13	(0.79)	-0.16	(0.3)	-0.05	(0.16)	0.35	(1.21)	-0.35	(0.38)
3. Professional community												
Principal-reported PD	-1.18	(4.05)	-2.37	(7.34)	0.56	(2.87)	1.35	(2.64)	3.53	(4.94)	0.19	(2.84)
Teacher-reported PD	-6.88	(11.47)	-10.43	(19.34)	-17.17^{+}	(9.08)	0.04	(7.62)	-17.53	(17.4)	9.06	(10.28)
Teacher-reported collaboration	5.59	(7.77)	15.52	(15.45)	-11.63*	(4.98)	11.05	(6.88)	18.10^{+}	(10.19)	0.59	(5.07)
4. External community												
Parental involvement	-1.65	(15.13)	-32.12	(29.13)	-4.96	(13.83)	-3.34	(10.57)	28.78	(21.47)	-3.84	(15.89)
5. Performance accountability												
Observation by principal	9.16	(7.88)	29.85	(36.22)	-28.45**	(9.64)	7.16	(8.08)	52.63	(50.93)	20.16	(17.87)
Observation by external inspection	5.51	(9.70)	-4.03	(13.27)	-4.29	(5.39)	-13.52**	(4.22)	2.62	(12.52)	-4.58	(5.42)
Evaluated with student achievement	4.82	(8.24)	3.18	(16.88)	-2.63	(4.87)	2.51	(6.48)	7.36	(24.84)	-5.96	(5.51)
Teacher peer review	7.76	(7.03)	4.14	(13.28)	-1.18	(4.75)	10.10^{+}	(5.63)	-1.53	(7.52)	-3.95	(5.64)
Incentive to recruit or retain teachers	20.74	(14.1)	46.98	(32.58)	-6.24	(6.50)	0.83	(6.16)	8.22	(10.91)	-10.51	(10.5)
6. Culture of learning												
Principal's perception of parent and student desire to do well	4.09	(6.90)	10.75	(13.71)	9.93^{+}	(5.56)	-0.97	(4.05)	26.98**	(10.00)	4.91	(5.54)
Teacher's perception of parent and student desire to do well	6.45	(6.49)	27.67*	(12.73)	3.77	(5.27)	2.42	(4.60)	5.23	(8.38)	12.54*	(5.36)

Table 6: Learning-centered school conditions and mathematics achievement: fixed effects from two-level HLM (contd.)

Control variables	Chinese Taipei		Hong Kong SAR		Japan		Korea		Singapore		United States	
School level												
Grade 8 enrollment	0.02^{+}	(0.01)	0.18	(0.22)	-0.01	(0.04)	-0.02	(0.02)	0.13*	(0.06)	0.04*	(0.02)
Type of community	2.05	(3.29)	5.52	(8.07)	0.04	(2.16)	1.20	(2.49)	na		-4.44*	(1.82)
Economically disadvantaged (%)	-2.38	(4.20)	-16.03*	(6.28)	-7.08^{+}	(3.84)	-4.87*	(2.43)	-1.28	(4.01)	-11.85**	(3.37)
Tested in native language (%)	-0.46	(2.82)	2.41	(10.59)	na		na		-3.36	(4.40)	2.03	(2.76)
Ability grouping	4.13	(9.44)	-29.10*	(12.58)	13.30*	(5.42)	-2.56	(5.04)	-1.78	(7.45)	8.70	(5.65)
Teacher qualifications												
Math major	4.61	(9.16)	-13.13	(15.85)	-1.82	(7.89)	-14.39	(10.34)	21.42^{+}	(11.80)	4.78	(6.81)
Female	4.05	(7.04)	7.98	(13.41)	9.81^{+}	(5.17)	4.06	(5.33)	8.31	(9.80)	-8.22	(6.41)
Teaching certificate	29.62	(23.52)	-28.00	(42.08)	na		na		7.16	(30.98)	-10.92	(16.96)
Years in teaching	0.11	(0.39)	0.06	(0.68)	-0.09	(0.28)	0.00	(0.24)	-0.27	(0.46)	-0.08	(0.30)
Classroom context												
Use math text books	0.40	(12.75)	na		-27.37	(21.84)	-5.33	(14.44)	6.33	(15.62)	-15.43	(11.33)
No. of students in TIMSS class	0.75	(0.58)	1.78*	(0.73)	0.62^{+}	(0.34)	0.31	(0.50)	-0.73	(1.05)	-0.34	(0.39)
Minutes/week for math teaching	0.09	(0.06)	-0.15^{+}	(0.09)	0.01	(0.11)	-0.20	(0.15)	-0.26*	(0.13)	0.05	(0.04)
Amount of homework per week	5.01	(3.84)	19.16*	(8.98)	-2.59	(3.06)	-3.92	(2.58)	11.99*	(6.06)	2.64	(4.81)
Length of homework per week	17.22**	(4.49)	19.27^{+}	(11.31)	4.47	(3.15)	2.28	(3.00)	11.08^{+}	(5.75)	13.46*	(5.60)

Table 6: Learning-centered school conditions and mathematics achievement: fixed effects from two-level HLM (contd.)

Control variables	Chinese Taipei		Hong Kong SAR		Japan		Korea		Singapore		United States	
Student characteristics												
Female	-3.05	(3.28)	-16.51**	(3.13)	-6.50*	(2.69)	-3.86	(2.95)	6.09*	(2.42)	-8.01**	(1.76)
Speaks language of test at home	17.84**	(2.38)	6.79**	(2.05)	24.01**	(3.76)	9.43**	(2.39)	-0.08	(1.34)	3.33+	(1.68)
Possesses calculator	26.42*	(12.00)	47.63**	(14.57)	16.79+	(9.97)	-1.67	(7.74)	43.62**	(11.27)	11.78**	(4.54)
Possesses computer	39.78**	(8.82)	-16.02+	(9.71)	5.01	(4.79)	17.53	(14.98)	10.55	(6.08)	7.27+	(4.37)
Possesses study desk	14.82**	(5.11)	-10.29**	(2.68)	16.22**	(5.61)	12.5	(8.15)	13.45**	(3.68)	4.10+	(2.39)
Possesses dictionary	42.16**	(11.90)	9.94	(10.31)	64.14**	(11.46)	82.13**	(12.27)	33.53**	(10.00)	3.71	(3.26)
Possesses Internet connection	4.44	(5.81)	16.84*	(6.98)	16.99**	(3.64)	58.38**	(8.51)	28.31**	(4.53)	1.00	(3.72)
Mother's education	-0.05	(1.52)	-0.13	(1.45)	3.08	(1.72)	2.90**	(1.11)	-0.87	(0.88)	1.13+	(0.58)
Father's education	6.46**	(1.44)	1.04	(1.46)	10.47**	(1.31)	5.42**	(1.16)	4.26**	(0.80)	3.12**	(0.78)
Number of books at home	17.21**	(1.38)	2.05+	(1.09)	9.35**	(1.10)	21.54**	(1.19)	7.90**	(1.03)	10.48**	(0.78)
Sample size												
School	143		106		141		144		155		192	
Student	3,830		3,040		4,151		4,072		4,351		5,859	

Notes:

Entries are coefficients (standard errors in parentheses).

na represents that either data were not available or the variable was the same within the system; + $p < 0.10$, *$p < 0.05$, **$p < 0.01$.

The extent to which principals spent their time on curriculum development and instructional quality (called "instructional leadership" in the TIMSS questionnaires) or directly engaged in teaching was only weakly associated with student achievement across all six selected systems. This finding is consistent with prior research that found little direct impact of principal instructional involvement on student achievement (Heck, 1993; Leithwood et al., 2004).

Teacher Professional Community

Both the principal-reported and teacher-reported measures for teacher professional development yielded statistically nonsignificant associations with student achievement, with the exception of Japan, where a one unit of increase of frequency in teacher-reported professional development was negatively associated with 17.17 ($p < 0.10$) of mathematics scores. It is conceivable that professional development may be used as additional training for teachers of low-performing students. Teacher collaboration, defined by discussions, joint lesson planning, and observations of peer teaching, was positively associated with student learning in Singapore (18.10, $p < 0.10$). However, the association was negative (-11.63, $p < 0.05$) in Japan. Again, it is possible that the same activities were being utilized for different purposes in the different settings.

Connections with External Community

The parental involvement measure did not appear to be associated with student achievement, perhaps because of the low internal reliability of the measure. In other words, the types of parental activities included in the questionnaires may be beneficial for a positive learning environment in various ways, but a composite score (average of the five dichotomous items) may not be directly associated with student achievement. Qualitative research may be necessary to fully account for the richness of community involvement as one of the key leadership effects on school improvement, as described by Chen (2008) and Pan and Yu (1999) in their qualitative studies of Taiwanese schools.

Performance Accountability

The performance evaluation measures for teacher practices yielded largely insignificant coefficients on student achievement across the six systems, except for Japan and Korea. In Japan, being observed by the principal or senior staff for mathematics teaching evaluation was negatively associated with student achievement by 28.45 ($p < 0.01$) points of decrease. In Korea, being observed by external inspectors was also negatively associated with achievement (-13.52, $p < 0.01$). However, teacher peer review as a form of evaluation had a positive association with student achievement (10.10, $p < 0.10$) in Korea. It may be that external observation, whether by the principal, senior staff, or inspectors, reinforces standards and enhances instructional quality, especially for teachers with low-performing students or who need to engage in relevant professional development. On the other hand, peer review might occur when a teacher is considered exemplary. It would be interesting to examine if and the extent to which evaluation patterns are associated with teacher qualifications in our next steps of data analysis.

Culture of Learning

The principal-reported and teacher-reported measures that capture students' (and their parents') desire to do well produced strong and large coefficients on student achievement in all six systems, but especially in the United States (12.54, $p < 0.05$), Hong Kong SAR (27.67, $p < 0.05$), and Singapore (26.98, $p < 0.01$). While the finding here is consistent with the theoretical assumptions about cultural values relating to education and student achievement (Paik, 2004; Shen, 2005; Wang, 2004), the four items that formed the measure are broadly defined: parental support for student achievement, parental involvement in school activities, students' regard for school property, and students' desire to do well in school. These dimensions may be confounded with other explanatory factors. Culture of learning as a construct domain will need to be further extricated in order to identify elements that are malleable by principals, teachers, and other key stakeholders.

Control variables at the school, classroom, and student levels yielded results largely consistent with previous literature and theoretical assumptions. For example, the percentage of economically disadvantaged students in the school, being female, not tested in the native language, and lack of Internet access at home were negatively associated with student achievement. Moreover, frequency of mathematics assignments and length of time for homework were positively associated with mathematics scores when all other factors were held constant. The association between ability grouping and mathematics achievement varied among the selected systems; it was a positive one in Japan (13.30, $p < 0.05$), but negative in Hong Kong SAR (-29.10, $p < 0.05$). It is plausible that ability grouping as a practice is used for different purposes, ranging from addressing the diverse needs of low-performing students (and thus being a negative association with student achievement) to meeting the needs of advanced students (and thus a positive association).

DISCUSSION

The TIMSS contextual questionnaires have been recognized for providing important background information on the learning conditions of students in participating education systems. In this study, we sought a fuller appreciation of how TIMSS data might be used to inform the field of school administration and leadership.

Our findings indicate that the TIMSS contextual questionnaires provide an interesting crossnational snapshot of learning conditions in the participating nations. We found this snapshot multidimensional and informative. Specifically, by using the modified learning-centered leadership framework to identify items from the questionnaires that reflected the core components of malleable learning conditions, we were able to tap into all six dimensions of the main construct of learning-centered leadership, albeit not fully on all fronts. We then explored the cross-national evidence for the associations between core learning conditions within the realm of influence of school leadership and student achievement by connecting mathematics achievement outcomes with

contextual questionnaire results. We also controlled for factors that might confound the relationships between school learning conditions and educational outcomes at multiple contextual levels.

The results showed that a number of learning conditions pertaining to accountability measures, classroom instructional practices, and attitudes toward learning were significantly and strongly associated with student learning in many of the selected systems. Our analyses of the selected systems also affirmed the notion that school leadership cannot simply be measured by the amount of direct instruction-related efforts of the principals (Hallinger & Heck, 2010; Leithwood & Jantzi, 1999; Louis et al., 2010). In essence, we operationalized, to an extent, the theoretical assumptions regarding leadership, learning-centered school conditions, and student achievement results within the TIMSS data framework. However, our probe also underscored the challenge of using TIMSS questionnaires to study the role of school leadership in student learning.

First, the TIMSS questionnaires did not appear to cover the full domain of the learning-centered leadership framework. The items available from the school, teacher, and student questionnaires were designed with minimizing the survey burden to respondents in mind. For example, only two items are available to reflect *standards for student learning* at the school level, and the items for connections with *external community* reflect parental involvement only.

Second, the items tend to be written in generic terms and lack the necessary specificity to describe what leaders should do to impact school conditions. For example, instructional leadership is defined as "developing curriculum and pedagogy," which is a very obvious simplification of the complex actions involved in providing this type of leadership (Hallinger & Murphy, 1987). Such vagueness may present a threat to the construct validity of the measures.

Time spent on classroom teaching by the principal is another example that deserves further consideration. Teaching may well afford principals the opportunity to be directly involved in classroom interaction with students. However, if schools have head teachers who take the lead in lesson planning and pedagogy development, principals may then be able to have a larger impact because of having the time to exercise instructional leadership at the school level OECD, 2010; Tucker, 2011).

In fact, the definitions and applications of learning-centered leadership will most certainly vary from country to country. Given that TIMSS 2007 questionnaires offer limited content validity on important learning conditions, in-depth and qualitative probes into how the same term is applied to these diverse definitions and practices may be necessary. For example, more in-depth understanding of what learning culture as a learning condition entails is needed beyond the items that we could identify from TIMSS in this study.

Third, the cross-sectional nature of the questionnaires and mathematics achievement provides limited insight into the relationship between the actions of school personnel in ensuring optimal learning conditions and student learning outcomes. When crossnational analysis shows opposing signs or varying degrees of associations, little explanation can be offered without more grounded investigation into the contexts of the schools and classrooms. Furthermore, even though the five East Asian systems selected for this study had average scores in mathematics that were statistically significantly higher than the corresponding scores in the United States, small point differences do not necessarily say something about the quality differences in education. If the influence of learning-centered leadership is to be analyzed in general—or in comparison with the United States—then the low-achieving systems should be chosen for reasons of validity, in the future. Overall, the limitations of this study underscore the need not only to broaden but to deepen our understanding of variation in educational and social contexts across countries so that we can fully appreciate the utility of international benchmarking for student achievement.

The 2008 NGA and CCSSO reports called for international benchmarking and proposed revising "state policies for recruiting, preparing, developing, and supporting teachers and school leaders to reflect the human capital practices of top performing nations and states around the world" (NGA, 2008, p. 27). The challenge, however, is to identify effective practices that take multilevel contextual factors into consideration. Using the TIMSS crossnational datasets from 2007, our study set out to examine the viability of using international assessment reports to inform school leadership practices. While our analysis did not include all participating nations, the results for the selected systems reveal interesting differences in school-level conditions for learning and how such conditions are associated with mathematics achievement. We hope to deepen our probe into available empirical evidences and identify convergent and divergent themes as compared with theoretical assumptions in the field.

Future research could further investigate the nature of linkage between school conditions malleable by leadership by:

(a) Including more countries in the analysis in order to identify any systematic differences between high- and low-achieving nations in the learning-centered conditions;

(b) Selecting a few countries of interest and conducting cohort comparisons among the four sets of TIMSS results (1995, 1999, 2003, and 2007) to determine if there are changes regarding important core learning-centered conditions and their associations with student learning over time;

(c) Analyzing local policies relevant to the design and implementation of improving learning-centered conditions; and

(d) Conducting analyses on whether learning-centered conditions are different with regard to the disaggregation of student subpopulations and school types, such as performance quartiles, racial and ethnic groups, public schools serving students with different socioeconomic concentrations, and urban versus rural schools.

To the extent that such insight identifies leadership components that are within the realm of control of school principals, information about the underlying processes gained from this study could be useful in informing cross-national research on school leadership regarding the development or the modification of existing professional training and evaluation. For the United States, such an approach might lead to the high levels of student achievement that other countries currently experience.

Appendix Table 1: Measures for learning-centered school conditions from 2007 TIMSS teacher questionnaire (TQ) and principal questionnaire (PQ)

Measures	Questionnaire items from TIMSS
	Standards for student learning
Teachers' academic expectation (TQ and PQ)	How would you characterize teachers' expectations for student achievement?
	Curriculum and instruction implementation
Curriculum rigor (TQ and PQ, 2 Items)	How would you characterize: • Teachers' understanding of the school's curricular goals within your school? • Teachers' degree of success in implementing the school's curriculum within your school?
Percentage of time in instructional leadership (PQ) Percentage of time in teaching (PQ)	By the end of this school year, approximately what percentage of time in your role as principal will you have spent on: • Instructional leadership (e.g., developing curriculum and pedagogy)? • Teaching?
	Teacher professional community
Professional development (PQ, 5 Items)	During this school year, how often have your eighth-grade teachers been involved in professional development opportunities for mathematics and science: • For mathematics and science targeted at supporting the implementation of the national or regional curriculum? • Targeted at designing or supporting the school's own improvement goals? • Targeted at improving content knowledge? • Targeted at improving teaching skills? • Targeted at using information and communication technology for educational purposes?
Professional development (TQ, 6 Items)	In the past two years, have you participated in professional development in: • Mathematics content? • Mathematics pedagogy/instruction? • Mathematics curriculum? • Integrating information technology into mathematics? • Improving students' critical thinking or problem-solving skills? • Mathematics assessment?
Teacher collaboration (TQ, 4 Items)	How often do you: • Have discussions about how to teach a particular concept with other teachers? • Have worked on preparing instructional materials with other teachers? • Visited another teacher's classroom to observe his/her teaching? • Have informal observations of your classroom by another teacher?

Appendix Table 1: Measures for learning-centered school conditions from 2007 TIMSS teacher questionnaire (TQ) and principal questionnaire (PQ) (contd.)

	External Community
Parental involvement (PQ, 4 Items)	Does your school expect parents to: • Attend special events (e.g., science fair, concert, sporting events)? • Raise funds for the school? • Volunteer for school projects, programs, and trips? • Ensure that their child completes his/her homework? • Serve on school committees (e.g., select school personnel, review school finances)?
	Performance accountability
Observation by principal	Are observations by the principal or senior staff used to evaluate the practice of eighth-grade mathematics teachers?
Observation by external inspection	Are observations by inspectors or other persons external to the school used to evaluate the practice of eighth-grade mathematics teachers?
Evaluated with student achievement	Is student achievement used to evaluate the practice of eighth-grade mathematics teachers?
Teacher peer review	Is teacher peer review used to evaluate the practice of eighth-grade mathematics teachers?
Incentive to recruit or retain teachers	Does your school currently use any incentives to recruit or retain eighth grade teachers in mathematics?
	Culture of learning
Parent and student desire to do well (PQ and TQ, 4 items)	How would you characterize Parental support for student achievement within your school?
	Parental involvement in school activities within your school?
	Students' regard for school property within your school?
	Students' desire to do well in school within your school?

Note: *Internal consistency, measured by Cronbach's alpha, for each scale specific to each system can be found in Table 2.

Appendix Table 2: Coding of measures

Variables	Coding
Learning-centered leadership framework	
1. Standards for student learning	
Principal-reported expectation	1= very low, 2 = low, 3 = medium, 4 = high, 5 = very high
Teacher-reported expectation	1= very low, 2 = low, 3 = medium, 4 = high, 5 = very high
2. Curriculum and instruction	
Principal-reported curriculum rigor	1= very low, 2 = low, 3 = medium, 4 = high, 5 = very high
Teacher-reported curriculum rigor	1= very low, 2 = low, 3 = medium, 4 = high, 5 = very high
Percentage time leadership	Percentage of time
Percentage time in teaching	Percentage of time
3. Professional community	
Principal-reported PD	1= none, 2 = 1–25%, 3 = 26–50%, 4 = 51–75%, 6 = 76–100 %
Teacher-reported PD	1 = yes, 0 = no
Teacher-reported collaboration	1= never or almost never, 2 = 2 or 3 times per month, 3 = 1–3 times per week, 4 = daily or almost daily
4. External community	
Parental involvement	1 = yes, 0 = no
5. Performance accountability	
Observation by principal	1 = yes, 0 = no
Observation with external inspection	1 = yes, 0 = no
Evaluated with student achievement	1 = yes, 0 = no
Teacher peer review	1 = yes, 0 = no
Incentive to recruit or retain teachers	1 = yes, 0 = no
6. Culture of learning	
Principal's perception of parent and student desire to do well	1= very low, 2 = low, 3 = medium, 4 = high, 5 = very high
Teacher's perception of parent and student desire to do well	1= very low, 2 = low, 3 = medium, 4 = high, 5 = very high

Appendix Table 2: Coding of measures (contd.)

Variables	Coding
Control variables	
School level	
Grade 8 enrollment	Number of students
Type of community	1 = 3,000 people or fewer, 2 = 3,001 to 15,000 people, 3 = 15,001 to 50,000 people, 5 = 100,001 to 500,000 people, 6 = more than 500,000 people
Percentage economically disadvantaged	1 = 0 to 10%, 2 = 11–25%; 3 = 26–50%, 4 = more than 50%
Percentage tested in native language	1 = less than 50%, 2 = 51–75%, 3 = 26–50%, 4 = more than 50%
Ability grouping	1 = yes, 0 = no
Teacher qualifications	
Math major	1 = yes, 0 = no
Female	1 = female, 0 = male
Teaching certificate	1 = yes, 0 = no
Years in teaching	Years
Classroom context	
Use math textbooks	1 = yes, 0 = no
Number of students in TIMSS class	Number of students
Minutes per week for math teaching	Minutes
Amount homework per week	0 = some homework, 1 = some lessons, 2 = about half the lessons, 3 = every or almost every lesson
Length of homework per week	1 = fewer than 15 minutes, 2 = 15–30 minutes, 3 = 31–60 minutes, 4 = 61–90 minutes, 5 = more than 91 minutes
Student characteristics	
Female	1 = female, 0 = male
Speak language of test at home	0 = never, 1 = sometimes, 2 = almost always, 3 = always
Possesses calculator	1 = yes, 0 = no
Possesses computer	1 = yes, 0 = no
Possesses study desk	1 = yes, 0 = no
Possesses dictionary	1 = yes, 0 = no
Possesses Internet connection	1 = yes, 0 = no
Mother's education	0 = ISCED Level 1 or 2, or did not go to school, 1 = ISCED 2; 2 = ISCED 3, 3 = ISCED 4, 4 = ISCED 5B, 5 = ISCED 5A, first degree, 6 = beyond ISCED 5A, first degree
Father's education	0 = ISCED Level 1 or 2, or did not go to school, 1 = ISCED 2; 2 = ISCED 3, 3 = ISCED 4, 4 = ISCED 5B, 5 = ISCED 5A, first degree, 6 = beyond ISCED 5A, first degree
Number of books at home	1 = 0–10, 2 = 11–25, 3 = 26–100, 4 = 101–200, 5 = over 100 books

References

Akiba, M., LeTendre, G. K., & Scribner, J. P. (2007). Teacher quality, opportunity gap, and national achievement in 46 countries. *Educational Researcher*, *36*(7), 369–387.

Baker, D. P., Goesling, B., & LeTendre, G. K. (2002). Socioeconomic status, school quality, and national economic development: A cross-national analysis of the "Heyneman-Loxley effect" on mathematics and science achievement. *Comparative Education Review*, *46*(3), 291–312.

Baker, D. P., Lee, J., & Heyneman, S. P. (2003). Should America be more like them? Cross-national high school achievement and U.S. policy. *Brookings Papers on Education Policy*, *6*, 309–338).

Betts, J. R., & Grogger, J. (2003). The impact of grading standards on student achievement, educational attainment, and entry-level earnings. *Economics of Education Review*, *22*, 343–352.

Chen, P. (2008). Strategic leadership and school reform in Chinese Taipei. *School Effectiveness and School Improvement*, *19*(3), 293–318.

Chudgar, A., & Luschei, T. F. (2009). National income, income inequality, and the importance of schools: A hierarchical cross-national comparison. *American Educational Research Journal*, *46*(3), 626–658.

Clarke, D., Mesiti, C., O'Keefe, C., Xu, L. H., Jablonka, E., Mok, I. A. C. ... Shimuzu, Y. (2007). Addressing the challenge of legitimate international comparisons of classroom practice. *International Journal of Educational Research*, *46*(5), 280–293.

Cobb, P., & Smith, T. (2008). District development as a means of improving mathematics teaching and learning at scale. In K. Krainer & T. Wood (Eds.), *International handbook of mathematics teacher education: Vol. 3. Participants in mathematics teacher education: Individuals, teams, communities and networks* (pp. 231–254). Rotterdam, the Netherlands: Sense Publishers.

Cohen, D., & Hill, H. (2000). Instructional policy and classroom performance: The mathematics reform in California. *Teachers College Record*, *102*, 294–343.

Council of Chief State School Officers (CCSSO). (1996). *Interstate school leaders licensure consortium standards for school leaders*. Washington, DC: Author.

Creemers, B. P. M., & Reezigt, G. J. (1996). School level conditions affecting the effectiveness of instructions. *School Effectiveness and School Improvement*, *7*, 197–228.

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

Elmore, R. F. (2000). *Building a new structure for school leadership*. Washington, DC: The Albert Shanker Institute.

Ferraro, D., & Van de Kerckhove, W. (2006). *Trends in International Mathematics and Science Study (TIMSS) 2003: Nonresponse bias analysis*. Washington, DC: National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education.

Foy, P., & Olson, J. F. (2009). *TIMSS 2007 user guide for the international database*. Boston, MA: Boston College.

Givvin, K. B., Hiebert, J., Jacobs, J. K., Hollingsworth, H., & Gallimore, R. (2005). Are there national patterns of teaching? Evidence from the TIMSS 1999 Video Study. *Comparative Education Review*, *49*(3), 311–343.

Glasman, N. S., & Heck, R. H. (1992). The changing leadership role of the principal: Implications for principal assessment. *Peabody Journal of Education*, *68*(1), 5–24.

Goldring, E., & Cravens, X. C. (2007). Teachers' academic focus on learning in charter and non-charter schools. In M. Berends, M. G. Springer, & H. J. Walberg (Eds.), *Charter school outcomes*. New York, NY: Lawrence Erlbaum Associates.

Goldring, E., Porter, A., Murphy, J., Elliott, S. N., & Cravens, X. (2009). Assessing learning-centered leadership: Connections to research, professional standards, and current practices. *Leadership and Policy in Schools*, *8*, 1–36.

Gonzales, P., Williams, T., Jocelyn, L., Roey, S., Kastberg, D., & Brenwald, S. (2008). *Highlights from TIMSS 2007: Mathematics and science achievement of U.S. fourth- and eighth-grade students in an international context*. Washington, DC: National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education.

Hallinger, P. (1990). *Principal Instructional Management Rating Scale*. Sarasota, FL: Leading Development Associates.

Hallinger, P. (2011). A review of three decades of doctoral studies using the Principal Instructional Management Rating Scale: A lens on methodological progress in educational leadership. *Educational Administration Quarterly*, *47*(2), 271–306.

Hallinger, P., & Heck, R. H. (1996). Reassessing the principal's role in school effectiveness: A review of empirical research, 1980–1995. *Educational Administration Quarterly*, *32*(1), 5–44.

Hallinger, P., & Heck, R. H. (2010). Collaborative leadership and school improvement: Understanding the impact on school capacity and student learning. *School Leadership & Management*, *30*(2), 95–110.

Hallinger, P., & Murphy, J. (1985). Assessing the instructional leadership behavior of principals. *Elementary School Journal*, *86*, 217–248.

Hallinger, P., & Murphy, J. (1987). Instructional leadership in the school context. In W. D. Greenfield (Ed.), *Instructional leadership: Concepts, issues, and controversies* (pp. 179–203). Boston, MA: Allyn & Bacon.

Heck, R. H. (1993). School context, principal leadership, and achievement. *The Urban Review*, *25*(2), 151–166.

Heyneman, S. P., & Loxley, W. A. (1983). The effect of primary school quality on academic achievement across 29 high- and low-income countries. *American Journal of Sociology*, *88*(6), 1162–1194.

Leithwood, K., & Jantzi, D. (1999). The relative effects of principal and teacher sources of leadership on student engagement with school. *Educational Administration Quarterly*, *35* (Supplemental), 679–706.

Leithwood, K., Louis, K. S., Anderson, S., & Wahlstrom, K. (2004). *How leadership influences student learning: A review of research for the Learning from Leadership Project*. New York City, NY: Wallace Foundation

LeTendre, G., Baker, D., Akiba, M., Goesling, B., & Wiseman, A. (2001). Teachers' work: Institutional isomorphism and cultural variation in the U.S., Germany, and Japan. *Educational Researcher*, *30*(6), 3–15.

LeTendre, G., Baker, D. P., Wiseman, A., Boe, E., & Goesling, B. (2002). *Classroom implementation of national curricula and cross-national patterns of achievement* (Vol. 19, Working Paper Series, Pennsylvania State University, Education Policy Studies). University Park, PA: Pennsylvania State University.

Louis, K. S., Leithwood, K., Wahlstrom, K. L., & Anderson, S. E. (2010). *Learning from leadership: Investigating the links to improved student learning*. New York, NY: Wallace Foundation.

Mullis, I. V. S., Martin, M. O., & Foy, P. (2008). *TIMSS 2007 international mathematics report: Findings from IEA's Trends in International Mathematics and Science Study at the fourth and eighth grade*s. Chestnut Hill, MA: Boston College.

Mullis, I. V. S., Martin, M. O., Smith, T. A., Garden, R. A., Gregory, K. D., & Gonzalez, E. J. (2005). *TIMSS assessment frameworks and specifications 2003*. College Hill, MA: Boston College.

Murphy, J., Goldring, E., Elliott, S. N., & Porter, A. (2006). *Learning-centered leadership: A conceptual foundation*. New York, NY: Wallace Foundation

National Center for Education Statistics (NCES). (2008). *TIMSS 2007 assessment frameworks*. Washington, DC: Institute of Education Sciences, U.S. Department of Education.

National Governors Association (NGA). (2008). *Benchmarking for success: Ensuring U.S. students receive a world-class education*. Washington, DC: Author.

Organisation for Economic Co-operation and Development (OECD). (2010). *Strong performers and successful reformers in education: Lessons from PISA for the United States*. Paris, France: Author. Retrieved from http://www.pisa.oecd.org/dataoecd/32/50/46623978.pdf

Paik, S. J. (2004). Korean and U.S. families, schools, and learning. *International Journal of Educational Research*, *41*(1), 71–90.

Pan, H. L., & Yu, C. (1999). Educational reforms with their impacts on school effectiveness and school improvement in Chinese Taipei, R.O.C. *School Effectiveness and School Reform*, *10*(1), 72–85.

Porter, A., & Gamoran, A. (2002). *Methodological advances in cross-national surveys of educational achievement*. Washington, DC: National Academy Press.

Porter, A., Goldring, E., Murphy, J., Elliott, S. N., & Cravens, X. (2006). *A conceptual framework for the assessment of principal and team school leadership*. New York, NY: Wallace Foundation.

Provasnik, S., Gonzales, P., & Miller, D. (2009). *U.S. performance across international assessments of student achievement: Special supplement to "The Condition of Education 2009"* (NCES 2009-083). Washington, DC: National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage Publications.

Rice, R. C., & Islas, M. R. (2001). TIMSS and the influence of instructional leadership on mathematics and science performance. *NASSP, 85*(623), 5–9.

Rowan, B., Correnti, R., Miller, R. J., & Camburn, E. M. (2009). *School improvement by design: Lessons from a study of comprehensive school programs*. Madison, WI: Consortium for Policy Research in Education.

Rutkowski, L., Gonzalez, E., Joncas, M., & von Davier, M. (2010). International large-scale assessment data: Issues in secondary analysis and reporting. *Educational Researcher*, *39*(2), 142–151.

Rutter, R., & Jacobson, J. (1986). *Facilitating teacher engagement*. Madison, WI: National Center on Effective Secondary Schools.

Schmidt, W. H., Rotberg, I. C., & Siegel, A. (2003). Too little too late: American high schools in an international context. *Brookings Papers on Education Policy*, *6*, 253–307.

Shen, C. (2005). How American middle schools differ from schools of five Asian countries: Based on cross-national data from TIMSS 1999. *Educational Research and Evaluation: An International Journal on Theory and Practice*, *11*(2), 179–199.

Smith, T., Desimone, L., & Ueno, K. (2005). "Highly qualified" to do what? The relationship between NCLB teacher quality mandates and the use of reform-oriented instruction in middle school mathematics. *Educational Evaluation and Policy Analysis*, *27*(1), 75–109.

Snijders, T. A. B., & Bosker, R. J. (1999). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. London, UK: Sage.

Swanson, C. B., & Barlage, J. (2006). *Influence: A study of the factors shaping education policy*. Bethesda, MD: Editorial Projects in Education.

Tucker, M. (2011). *Standing on the shoulders of giants: An American agenda for education reform*. Washington, DC: National Center on Education and the Economy.

U.S. Department of Education. (2010). *A blueprint for reform: The reauthorization of the Elementary and Secondary Education Act*. Washington, DC: Office of Planning, Evaluation, and Policy Development.

Wang, D. B. (2004). Family background factors and mathematics success: A comparison of Chinese and U.S. students. *International Journal of Educational Research*, *41*(1), 40–54.

Wang, J., & Lin, E. (2005). Comparative studies on U.S. and Chinese mathematics learning and the implications for standards-based mathematics teaching reform. *Educational Researcher*, *34*(5), 3–13.

Waters, T., Marzano, R. J., & McNulty, B. (2003). *Balanced leadership: What 30 years of research tells the U.S. about the effects of leadership on student achievement*. Aurora, CO: Mid-continent Research for Education and Learning.

IERI TECHNICAL NOTES

Rescaling sampling weights and selecting mini-samples from large-scale assessment databases

Eugenio J. Gonzalez
Educational Testing Service, Princeton, New Jersey, United States

This edition of IERI Technical Notes addresses two different, yet related, issues that researchers often encounter when working with large-scale assessment (LSA) databases. The first issue discussed is that of rescaling sample weights so that they add to a predetermined number that is set according to design or computational needs. This procedure is necessary when we want to work with entire databases, such as those assembled by international large-scale assessments, and when we want groups of cases that exist in different proportions in the population, such as countries, to contribute in equal proportions to summary statistics. A case in point is calculating item parameters with data from across all participating countries so that each country contributes equally in the calculation of these estimates. Rescaling the sample weights is also useful when we want the sum of the sample weights to add to a number equal to, for example, the sample size or effective sample size, in cases where the statistical software used has no straightforward way to adjust for the use of sampling weights.

The second issue is that of selecting subsets of cases from the database while taking into account their original selection probabilities and sampling weights. The procedure described in this technical note involves selecting one or multiple samples from the database, such that the resulting subset of cases can be treated as a simple random sample from the population. As a result of this selection, each record has equal weight, eliminating the need to use sampling weights in the calculations because the resulting data are the equivalent of selecting a simple random sample. This type of selection is useful when, as with rescaling sample weights (issue one above), we want to select samples of equal size from within different groups of the population. Using a file built in this way leads to each group of the population contributing equally to summary statistics. This type of selection is also useful when we want to select multiple samples

IERI Monograph Series: Issues and Methodologies in Large-Scale Assessments Volume 5

from a database and to use these to study the variability of estimates, a process that is akin to using bootstrapping techniques, but involves selecting without replacement.

SAMPLE SELECTION IN LSA

Sample selection in LSA usually involves two or more stages. During the first stage, schools are selected proportional to their size. During the second stage, students are chosen within the selected school. Depending on the sample design, classrooms might be selected within the school, and then including some or all of the students in the selected classroom. Students might also be selected directly across classes. Although these procedures vary somewhat, they all make it possible to calculate the probability of selection at each stage. Because the probabilities of selection at each stage are independent of one another, the overall probability of selection of any case can also be calculated as the product of the individual probabilities at the different stages. Further detail on how specific international LSA programs carry out their sample selection is beyond the scope of this paper, but can be found in the corresponding study technical reports (see, for example, Martin, Mullis, & Kennedy, 2007; Olson, Martin, & Mullis, 2008; Organisation for Economic Co-operation and Development [OECD], 2009). Kalton (1983), Kish (1968), and Ross (2005) provide additional useful information on sampling principles and methods.

SAMPLING WEIGHTS IN LSA

A sampling weight is the inverse of the probability of selection of a unit; it can be loosely interpreted as the number of units in the population that are represented by the selected unit. In the case of LSA, these units are usually students, although they can also be teachers, schools, parents, and so on. We therefore need to use sampling weights when analyzing the data so that each student's contribution to the statistical estimates is proportional to the number of students represented in the population. Using sampling weights helps to adjust the proportional contribution of the elements that make up the total estimate. In most cases, the sampling weights incorporate one or more adjustments for non-participation or post-stratification.

Using sampling weights to calculate a mean is relatively simple. We sum the product of each observed value by its weight, and divide by the sum of the weights. For example, if we have the values 10, 12, and 4, with weights 40, 60, and 80, respectively, the weighted average of these values would be as follows: (10*40+12*60+4*80) / (40+60+80) = 8.0.

The magnitude of the weights, simply calculated as the inverse of the probability of selection, is proportional to a population total. But the key property of the sampling weights is that they capture the relative proportion of each sampled unit with respect to the overall population. As a consequence, the sampling weights can be transformed by multiplying them by a constant, a step that preserves this proportionality. It also simply alters the overall sum of the weights, but not the statistic calculated. If, in the case of the example above, we divide the weights by 10, we get the same result: (10*4+12*6+4*8) / (4+6+8) = 8.0.

A simple example helps illustrate this point. Let's assume we select a simple random sample of 10% of students from a city with 1,800 students. Each of these students is enrolled in one of three educational tracks: Track V, with 1,000 students; Track A, with 500 students; and Track G, with 300 students. We would expect that our sample of 10% will have approximately 100 students from Track V, 50 students from Track A, and 30 from Track G. Because we have selected 10% of students, or 1 out of every 10, the probability of selection of each student is 1/10, and the corresponding weight, or the inverse of this probability, would be 10/1. Therefore, we have in our sample 100 Track V students, each "weighing" or representing 10; 50 Track A students, each weighing or representing 10; and 30 Track G students, each weighing or representing 10. If we add the number that each selected student represents, we obtain 1,800, which is the same as the overall number of students in the population.

In the example above, all selected students, regardless of the track they are in, have the same probability of selection and therefore the same sample weight. But what if we notice that we have only 30 students from Track G, but have 100, or over three times more, from Track V? We could then decide to select 10% of the students from the population, but this time select an equal number of students from within each track. We then proceed to select 60 out of 1,000 students from Track V, 60 out of 500 students from Track A, and 60 out of 300 students from Track G. As is evident in Table 1, the probabilities of selection within each track have changed, and so have the weights that would be assigned to each student. However, the proportional contribution of each group to the overall, when using the weights, has remained constant. The sampling weights thus allow us to adjust the relative contribution of the sampled elements when we need to sample at different rates within different groups.

Table 1: Weights with unequal sampling probabilities, by group

Description	*N*	Number of selections	Probability of selection (*N* of selections/*N*)	Sample weight (1/probability of selection)	Units represented (number of selections * sample weight)
Track V	1,000	60	0.06	16.67	1,000
Track A	500	60	0.12	8.33	500
Track G	300	60	0.20	5.00	300
Total	1,800	180			1,800

Table 2 provides another example. In it, we present summary statistics for five educational systems that participated in the Trends in International Mathematics and Science Study (TIMSS) 2007, conducted by the International Association for the Evaluation of Educational Achievement (IEA). We calculated these numbers from the publicly available international database. Even though each of the represented countries is very different in population size, a similar number of students (between 4,117 and 5,726) were selected within each country to participate in the study.

Table 2: Sample sizes in selected countries in TIMSS 2007

Country	Sample size	Average sample weight	Population estimate
Australia	4,791	53.73	257,407
Bahrain	4,199	2.51	10,543
Armenia	5,726	9.52	54,502
Bulgaria	4,117	21.28	87,603
Belgium (Flemish)	4,970	14.21	70,637

RESCALING THE WEIGHTS

As we indicated earlier, when we rescale weights, all we are doing is multiplying them by a constant. We do this so that the sum of the weights will be a number that is set according to design and computational needs. For example, some statistics use the sum of the weights as part of the calculations, such as with categorical data analysis statistics. In this case, we want the weights to add up to the sample size so that when we use this value in the calculations, it will be equivalent to the sample size. In other cases, we might want to control the proportional contribution of different groups that contribute to a statistic.

For example, let's say we want to compute an international average for the five countries in the data presented in Table 2 above. We could have each country contribute proportionally to however many students exist in the population, in which case Australia would contribute almost 25 times as much as Bahrain would. Alternatively, we could have each country contribute equally to a summary statistic. But doing this depends on how we define the average. If the average for the population is defined as the total of the scores for all students over the population, regardless of where they come from, then each country should contribute proportionally to however many students exist in the population, in which case Australia would contribute about 25 times as much as Bahrain would. However, if the average for the population of interest is defined as an arithmetic average of the mean scores for the countries, we would need to calculate an international average whereby each country contributes equally.

The equal contribution from each country can be achieved in two ways: by calculating the statistic for each country and then taking the average, or by rescaling the weights within each country to add to a constant and then using these rescaled weights in the calculation. In general, when we calculate so called "international" statistics, the

recommended approach is first to compute the individual country estimates, and then to take the simple average of these. This approach has two main advantages. First, by allowing us to easily calculate the standard error of the international average as the standard error of aggregated means from independent samples, it prevents us from making assumptions about how the replication procedure for calculating the standard error was conducted across the countries. The second, and perhaps most important, advantage is that we do not have the problem of unequal representation that arises when data are missing at different rates within countries for the same variable, or when groups within the population exist in different proportions.

Take, for example, the case where we want to compute the international percent of correct answers for several items, and we want each country to contribute equally. Because, in LSA, each item is taken by different numbers of students within each country, with each set of students being a representative sample of the student population within the country, we need to adjust the contribution of each group to the overall estimate. If we were to use the rescaled weights in this analysis, we would need to rescale them for each of the items to ensure that the pattern of missing data does not affect the desired equal contribution from each of the countries.

One common misconception found among analysts is their belief that rescaling the weights needs to be done only once. For example, the IEA TIMSS and Progress in International Reading Literacy Study (PIRLS) databases include a "senate weight" that adds up to 500 within a country, and the OECD Programme for International Student Assessment (PISA) databases contain a "weight factor" that can be used to transform the weights so that the sum within each country equals 1,000. Using these weights or factors will make the sum of the weights overall for each country equal to a constant. But the sum of the weights for subgroups within each country will vary. So, in the IEA databases, for example, the sum of the senate weights for boys within each country is proportional to the number of boys in the country in comparison to the number of girls and the amount of missing data for the variable for that country.

Rescaling the weights is as simple as multiplying the original sampling weights by a constant that will yield a desirable result based on a design or computational need. This occurs when we rescale weights in order to add to a constant across different groups, as with the senate weights in the IEA databases, or when we transform them to add to a specific number within each country (as is the case for the "house"[1] weights in the IEA databases).

The mechanics of rescaling the weights are as follows:

1. Select the constant to which the sum of the weights will be rescaled (in our example, "K").
2. Compute the sum of the sampling weights within each of the groups within which the rescaling will be done.
3. Multiply the sample weights by the result of dividing the constant selected in Step 1 above by the sum of the weights in Step 2.

1 House weights included in the IEA databases are transformed within each country to add to the overall sample of the country.

RESCALING THE WEIGHTS USING SPSS

As part of this technical note, we provide the SPSS[2] code that can be used to rescale the weights to a specified constant. The code is available online at www.ierinstitute.org/IERI_TechNote1.zip. It consists of two parts—a macro and a macro call—which we have placed in two separate files. The macro can be called from within any SPSS command syntax and executes a specified set of commands. Figure 1 shows the macro "RescaleWGTS" that we use in the example below.

To execute the macro, we need to specify a set of parameters in the call of the macro. The call of the macro is an SPSS command syntax that contains the necessary parameters for the macro to run. The macro can be executed from within any SPSS syntax window. Therefore, when using the macro RescaleWGTS, we need to specify these parameters:

- INFILE: The name of the file that has the original sampling weights. This file is not overwritten by the program.
- OUTFILE: The name of the file where the new weights will be saved. This file will preserve all the variables in the original file. It will also have a variable with the rescaled weights.
- DIR: The name of the directory where the INFILE and OUTFILE are located.
- CVAR: This lists the classification variables used to group the data. The sum of the weights will add to a constant within each unique combination of the variables defined by these classification variables.
- WGTS: The name of the variable in the original file that has the sampling weight. Here we can specify one or more weights in case we want to rescale several weights simultaneously, as would be the case when rescaling replicate weights.
- NEWWGTS: The name of the variable in the new file that will contain the rescaled sampling weights. There needs to be as many entries for this parameter as were made for the WGTS parameter. Not having these will have unexpected results.
- K: The constant that we want the weights to add up to.

Figure 2 shows an example of how the macro is called. In this example, we are doing the following:

1. Working with files located in "C:\IERI_TechNote1."
2. Reading the data from the file "RescaleFrom.sav."
3. Rescaling the weight variables called "TOTWGT."

These steps lead to:

4. The rescaled weights being saved to variable "R_TOTWGT."
5. The weights adding up to 1,000 for each IDCNTRY by ITSEX combination.
6. The resulting file being saved to "RescaleTo.sav."

2 The code is also available in SAS from www.ierinstitute.org/IERI_TechNote1.zip

We need, at this point, to offer several clarifications and recommendations:

1. Although not strictly necessary, saving the resulting file with a different name from that of the original file is good practice. Not doing this could have unexpected results.
2. The weight variable specified in the WGTS parameter must exist in the original file. One or more weights can be specified.
3. The variable name for the rescaled weights (NEWWGTS) should be different from the original weight variable. Not doing this will overwrite the original weights.
4. The weights should be rescaled to a constant that is "reasonable." While defining reasonable might be difficult, recognizing unreasonable is not. LSA databases tend to provide rescaled weights that add up to a number close to, yet not necessarily the same as, the effective sample size for a country.
5. Checking the results and verifying that the outcome is the desired one is very important. As part of quality control, the macro computes the sum of the weights within each of the groups and presents the results. This is shown in Figure 3, where we notice that the sum of the new variable, called "r_totwgt", equals 1,000 within each IDCNTRY by ITSEX combination, even in the case where ITSEX is coded as a user-defined missing value.

Figure 1: SPSS macro to rescale sampling weights

```
SET Length = None Width = 255
SET format f8.2.

* Rescales the weights within each <cvar> grouping and makes them add to <k>.
* The variable "newwgt" contains the rescaled weight.
* The resulting file is saved to <outfile> and has all the records from the original file.
* All variables in the original file are preserved.

define RescaleWgts
        (dir     = !charend('/') /
         infile  = !charend('/') /
         outfile = !charend('/') /
         cvar    = !charend('/') /
         wgts    = !charend('/') /
         newwgts = !charend('/') /
         k       = !charend('/') ).

set mprint = on.

* Count number of weights to rescale.
!let !nw = !null
!do !w !in(!wgts)
!let !nw = !concat(!nw,"w")
!doend
!let !nwgts = !length(!nw)

get file = !quote(!concat(!dir,"\",!infile,".sav")).
weight off.
sort cases by !cvar.

save outfile = !quote(!concat(!dir,"\",tmp0)).
```

Figure 1: SPSS macro to rescale sampling weights (contd.)

```
aggregate outfile = !quote(!concat(!dir,"\",tmp1))
  / break = !cvar
  / !do !w !in(!wgts) !concat("t",!w) !doend  = sum(!wgts).

match files
 / file = !quote(!concat(!dir,"\",tmp0))
 / table = !quote(!concat(!dir,"\",tmp1))
  / by !cvar.

!let !tmpnew = !newwgts.
!let !tmpold = !wgts.
!let !tmptot = !null.

!do !w !in(!wgts)
!let !tmptot =  !concat(!tmptot," t",!w).
!doend

!do !w = 1 !to !nwgts
compute !head(!tmpnew) = !head(!tmpold) * (!k / !head(!tmptot)).
!let !tmpnew = !tail(!tmpnew).
!let !tmpold = !tail(!tmpold).
!let !tmptot = !tail(!tmptot).
!doend.

execute.

mean tables = !newwgts !do !cv !in(!cvar) by !cv !doend
 / cells = sum min max count
 / missing = include.

save outfile = !quote(!concat(!dir,"\",!outfile,".sav"))
  / drop = !do !w !in(!wgts) !concat("t",!w) !doend.

new file.

erase file=!quote(!concat(!dir,"\",tmp0)).
erase file=!quote(!concat(!dir,"\",tmp1)).

!enddefine.
```

Figure 2: SPSS syntax to rescale the sampling weights

```
* Sample call of the macro to rescale the weights.

include file = "C:\IERI_TechNote1\RescaleWGTS.spm".

RescaleWGTS dir     = C:\IERI_TechNote1
        / infile  = RescaleFrom.sav
        / outfile = RescaleTo.sav
        / cvar    = idcntry itsex
        / wgts    = totwgt
        / newwgts = r_totwgt
        / k       = 1000.
```

Figure 3: SPSS output from rescaling the weights

COUNTRY ID	*SEX OF STUDENTS*	Sum	Minimum	Maximum	*N*
Australia	GIRL	1,000.00	0.01	1.54	2,443
	BOY	1,000.00	0.01	1.61	2,348
	Total	2,000.00	0.01	1.61	4,791
Bahrain	GIRL	1,000.00	0.19	0.71	2,025
	BOY	1,000.00	0.19	0.77	2,174
	Total	2,000.00	0.19	0.77	4,199
Armenia	GIRL	1,000.00	0.13	1.45	3,003
	BOY	1,000.00	0.15	1.62	2,696
	OMITTED	1,000.00	23.30	160.76	27
	Total	3,000.00	0.13	160.76	5,726
Bulgaria	GIRL	1,000.00	0.12	2.31	2,015
	BOY	1,000.00	0.11	2.15	2,102
	Total	2,000.00	0.11	2.31	4,117
Belgium (Flemish)	GIRL	1,000.00	0.22	1.03	2,620
	BOY	1,000.00	0.25	1.14	2,350
	Total	2,000.00	0.22	1.14	4,970
Total	GIRL	5,000.00	0.01	2.31	12,106
	BOY	5,000.00	0.01	2.15	11,670
	OMITTED	1,000.00	23.30	160.76	27
	Total	11,000.00	0.01	160.76	23,803

SELECTING MINI-SAMPLES FROM LSA

Despite recent increases in computing speed and capacity, those of us using LSA databases might still want to work with a smaller subset of the entire database, or select multiple subsets of the larger database. There are many reasons for taking these actions. For example, we might want to conduct preliminary or exploratory analysis with the data without the burden of having to use the entire dataset; or we might want to create multiple samples to validate results obtained using a particular statistical procedure, such as validating a factor analysis solution. In some instances, newer software systems are not very efficient at handling large databases, and for some procedures the facility of using sampling weights might yet not be available. In these cases, the use of mini-samples or subsets from the entire database can be useful.

The procedure described in this technical note amounts to selecting one or more samples from the database, where the resulting subset of cases is statistically equivalent to a simple random sample from the population. As a result of this selection, each record has equal weights, thus eliminating the need to use sampling weights in the calculations. This selection is useful when, as in the previous issue, we want to select samples of equal size from within different groups of the population.

Using a file built in this way effectively makes each group of the population contribute equally to summary statistics. This type of selection is also useful when we want to select multiple samples from a database and then use these to study the variability of estimates, a procedure akin to using bootstrapping techniques.

The results obtained from using these mini-samples are not expected to match exactly those results from the entire dataset, nor are they a substitute. Because LSA data are collected using complex sampling procedures, the process of selecting a subset of records from the database that is still representative of the population is not straightforward.

When selecting a subset, we need to take two matters into account: the probability of selecting the sampled units, and the desired composition of the resulting subset. We can take the first matter into account by sampling the units with probability proportional to their sampling weight. This results in a self-weighted sample in which all the units have equal weight. The sampling weight is the inverse of the probability of selection, and the sampling weights resulting from each stage are multiplicative, meaning the final sampling weight is the product of the weights at the different stages of selection.

This outcome can be easily shown. If we select the units from within the larger sample with probability proportional to their sampling weight, their probability of selection from the subset is then equal to:

$$Probability\ of\ Selection\ into\ subset = \left(\frac{Weight}{\Sigma\ Weights} * Number\ of\ Selections\right)$$

The sampling weight for this stage is the inverse of the probability of selection, or:

$$Weight\ for\ Selection\ into\ subset = \left(\frac{\Sigma\ Weights}{Weight * Number\ of\ Selections}\right)$$

Therefore, the resulting sampling weights, once the records have been selected for inclusion into the mini-sample, equal:

$$Final\ Weight\ for\ Selection\ into\ subset = Weight * \left(\frac{\Sigma\ Weights}{Weight * Number\ of\ Selections}\right)$$

After the redundant terms have been eliminated, the final weight for selection into the subset equals:

$$Final\ Weight\ for\ Selection\ into\ subset = Weight * \left(\frac{\Sigma\ Weights}{Number\ of\ Selections}\right)$$

As we can see, the final sampling weight for the units selected into the subset of cases simply equals the overall sum of the sampling weights divided by the number of selections.

A few points of clarification are necessary at this point:

1. The sum of the weights refers to the sum of the weights for the cases within the stratum or population group from where the selection is made.
2. The number of selections refers to the number of selections made from within the stratum from where the specific case is selected.
3. The weights of the units selected from within each stratum will be equal for each of the selected units.

Addressing the composition of the resulting subset requires a little more explanation. For example, assume we want to select a given number of cases from across the entire database, say a subset of 10,000 students. The selection results in a subset that will include different numbers of students from within each country, depending mostly on the overall size of the population for the country. Alternatively, we might be interested in selecting equal numbers of students from within each country or group (as defined by a set of grouping variables) while preserving the information provided by the sampling weights. Such is the case, for example, when we want to select a subset of students so that each country is represented by equal numbers of cases. Of course, we could address the equal representation of the countries in our analysis by transforming the weights to add up to a constant within each country and then using these transformed weights. But this approach will still leave us with all the cases in the data file and the computing burden unchanged, and it will not provide us with different samples to use in the validation of results.

Selecting representative subsets of students from within each stratum could be the technique to use when, for example, we want to validate data-processing procedures without using the entire database, or when we want to compare the results from two comparable samples of the population to assess the stability of a particular statistical estimate. LSA programs used this procedure during the 1990s in order to study the stability of item parameters at a time when computing capabilities made it prohibitive to use the entire database for calibration purposes.

The purpose of selecting a mini-sample is to have a more manageable set of cases that can be used in preliminary analyses, for didactic purposes, and as other examples. For example, the National Assessment of Educational Progress (NAEP) primer used mini-samples to describe and give examples of the use of NAEP data (Beaton & Gonzalez, 1995). Because individuals wanting to use the complete NAEP database have to have a license to do so, using mini-samples enables individuals without the license to run many graphs and tables. Mini-samples are also used to describe and provide examples of various techniques for estimating population parameters using the NAEP data (e.g., the average NAEP scores or percentages of students exceeding NAEP achievement levels in various demographic groups).

It is important to note that national and international results computed using these mini-samples are expected to be close—but not identical to—published results in the reports. The reason, of course, is that these mini-samples are a random subsample from the full database and therefore subject to sampling fluctuations. National or

international estimates should not be made with these data; nor should they be published as official estimates of the LSA database. Results from these mini-samples should be used for exploratory training purposes only.

The mechanics of selecting a mini-sample are as follows:

1. Specify the number of cases to select within each group (in our example, NSEL).
2. Calculate the sum of the weights within each group (in our example, TMPSWGT, for total measure of size), and find the number of times NSEL fits within TMPSWGT. This calculation will result in the selection interval of our systematic random sampling selection (in our example, INTERVAL).
3. Pick a uniform random number between 0 and 1 (in our example, RNDSTART); this will determine the case to select within each INTERVAL.
4. Compute the boundaries of the measure of size for each case in the data file (in our example, TMPBMOS is the lower boundary for begin-measure-of-size, and TMPCMOS is the upper boundary for cumulative-measure-of-size).
5. The selection then proceeds as follows:
   a. The first element to select is that which contains [RNDSTART * INTERVAL] between TMPBMOS and TMPCMOS.
   b. The next element to select is that which contains [RNDSTART * INTERVAL + (TMPSelNum-1)*INTERVAL].
   c. Depending on the number of selections, sample size, measure of size, and overall weight of the individual case, the same case could be selected more than once.
6. After completing the selection, we save only those records that were selected in Step 5 above. Should a record be selected more than once, we would want to write it as many times as it was selected. We would achieve this with the XSAVE command available in SPSS and would then add to this file a variable that indicates the selection sequence of the case (in our example, SELVEC).

SELECTING A MINI-SAMPLE USING SPSS

At this point, we want to provide the SPSS code for selecting a mini-sample while taking into account the sampling weights. The code consists of two parts that we have placed in two separate files: a macro and a call to the macro. The code is available online at (www.ierinstitute.org/IERI_TechNote1.zip).

The macro can be called from within any SPSS command syntax and executes a specified set of commands. Figure 4 on page 131 shows the macro "SamplePPS", which we use in our example. To execute the macro, we need to specify a set of parameters in the call of the macro. The call of the macro is an SPSS command syntax that contains the necessary parameters for the macro to run. The macro can be executed from within any SPSS syntax window.

When using the macro "SamplePPS" to select a mini-sample, we need to specify the following parameters:

- INFILE: The name of the file that has all the records. This file is not overwritten by the program.
- OUTFILE: The name of the file where the selected records will be saved. This file will preserve all the variables in the original file, but only those records that were selected into the mini-sample.
- DIR: The name of the directory where the INFILE and OUTFILE are located.
- CVAR: The listing of the classification variables that will be used to group the data. Equal numbers of selections will be made from within each unique combination of these grouping variables.
- WGT: The name of the variable in the original file that has the sampling weight.
- NEWWGT: The name of the variable that has the new sampling weight for the selected case. This is simply the sum of the sampling weights within the group divided by the number of selections. For most applications of the selection of cases, this weight will be discarded, but it is saved in the file as a measure of quality control.
- NSEL: The number of selections to make within each group.
- NSAMPLES: The number of mini-samples to select. Each of these samples is expected to be different from the others.
- SEED: The number used as the seed for the generation of random numbers. A different value for SEED will yield a different selection of cases into the mini-file. Only the same seed used on a file with the same number of records, sorted in the exact same order, will yield the exact same selection of cases. Any other combination will result in a randomly different selection of cases.
- IDVAR: Identification variables in the file that are used to sort the cases in the file prior to selection. If we want to ensure cases are selected from across different groups in the population, we need to specify the variables that identify these groups here. Although this parameter is optional, we highly recommend it. It works as an implicit stratification variable for the selection of cases to ensure cases are selected from a diversity of records.

Figure 5 shows an example of how the macro can be called, with some sample parameters. In this example, as evident in Figure 5, we are doing the following:

1. Working with files located in "C:\IERI_TechNote1."
2. Reading the data that are located in "SamplePPSFrom.sav."
3. Making 500 selections from within each IDCNTRY by ITSEX combination.

These steps lead to these outcomes:

4. The selection is made using the variable TOTWGT as the measure of size.
5. The cases are sorted, for the PPS selection, by the variables IDSCHOOL and IDSTUD.

6. The resulting file is saved to "SamplePPSTo.sav."
7. Ten samples in total are selected.

We again, at this point, need to offer several clarifications and recommendations, some of which are similar to the ones presented earlier:

- Although saving the resulting file with a different name from that of the original file is not strictly necessary, it does prevent us from overwriting the original file.
- The weight variable must exist in the original file.
- Depending on how many selections we choose to make as well as on sample size and magnitude of the sampling weights, it is possible for the same case to be selected more than once into the mini-sample file. In our example, the variable SELVEC in the resulting file contains the number of times a case was selected into the file. Cases selected more than once will be repeated in the output file as many times as necessary.
  - The variable SELSEQ in the output file has a sequential selection number for each selected case. This number will start at 1 and continue sequentially until reaching a maximum of the number of selections per subgroup multiplied by the number of possible unique groups that can be formed with the grouping variables.
  - The resulting file does not contain the weight variable from the original file. Because this weight is no longer useful, we drop it from the file. However, we then include in the file a new weight equal for all cases within each grouping, with the name specified by using the parameter NEWWGT.
  - An important part of the process involves checking the results and verifying that the outcome is the desired one. As part of quality control, the macro computes the sum of the weights and number of cases within each of the groups, and presents the results. This is shown in Figure 6, which presents summary statistics for the first sample selection. Here we can see that the number of cases equals 500 within each IDCNTRY by ITSEX combination. The values for "maximum" indicate the maximum number of times a single case was selected to be in the sample. In the figure shown, a girl in Bulgaria was selected twice, and individuals with omitted values in the ITSEX variable were selected multiple times because there were very few cases in this group.
  - Last, but not least, we cannot emphasize enough that national and international results computed using these mini-samples are expected to be close—but not identical to—published results in the reports or when the complete dataset is used. The reason, of course, is that these mini-samples are a random subsample from the full database and therefore subject to sampling fluctuations. National or international estimates should not be made with these data, and they should not be published as official estimates of the LSA database. Results from these mini-samples should be used for exploratory training purposes only.

Because the resulting mini-samples are equivalent to a simple random sample, there is no need to use complex methods for estimating sampling variance to obtain variances of the estimates.

Figure 4: SPSS macro to select mini-samples

```
SET Length = None Width = 255.
SET format f8.2.

* Selects <nsel> cases PPS from <infile> within each <cvar> grouping using <wgt> as the measure of size.
* It makes <nsamples> selections and saves each to a different file numbered sequentially.
* The PPS selection is done within each group using systematic SRS with the records sorted by <idvar>.
* The variable "selvec" contains the number of times a record is selected (could be > 1).
* The resulting <outfile> has <nsel> records.
* The variable "selseq" contains the selection sequence.
* All variables in the original file are preserved.
* The <seed> is used to initialize the random number generator.
* Different values for <seed> will result in different case selection.

define samplepps
        (dir     = !charend('/') /
         infile  = !charend('/') /
         outfile = !charend('/') /
         idvar   = !charend('/') /
         cvar    = !charend('/') /
         wgt     = !charend('/') /
         newwgt  = !charend('/') /
         nsel    = !charend('/') /
         nsamples= !charend('/') /
         seed    = !charend('/') ).

set seed = !seed.
set mprint = on.

get file = !quote(!concat(!dir,"\",!infile,".sav")).
weight off.
sort cases by !cvar !idvar.

save outfile = !quote(!concat(!dir,"\",tmp0)).

aggregate outfile = *
  / break = !cvar
  / tmpswgt = sum(!wgt).

!do !s = 1 !to !nsamples
compute !concat(rstart,!s) = uniform(1).
!doend

save outfile = !quote(!concat(!dir,"\",tmp1)).

match files
 / file  = !quote(!concat(!dir,"\",tmp0))
 / table = !quote(!concat(!dir,"\",tmp1))
  / by !cvar
   / first = tmpfirst
   / last  = tmplast.

save outfile = !quote(!concat(!dir,"\",tmp2)).

execute.
```

Figure 4: SPSS macro to select mini-samples (contd.)

```
save outfile = !quote(!concat(!dir,"\",tmp2)).

execute.

!do !s = 1 !to !nsamples

get file = !quote(!concat(!dir,"\",tmp2)).

compute interval = tmpswgt / !nsel.
compute selvec = 0.

do if (tmpfirst=1).
compute tmpbmos = 0.
compute tmpcmos = !wgt.
else.
compute tmpbmos = tmpcmos.
compute tmpcmos = tmpbmos + !wgt.
end if.
leave tmpcmos.
execute.

do if (tmpfirst = 1).
compute tmpselnum = !concat(rstart,!s) * interval.
end if.

do if (tmpselnum ge tmpbmos and tmpselnum lt tmpcmos).
compute selvec = 1+trunc((tmpcmos-tmpselnum)/interval).
compute tmpselnum = tmpselnum + selvec * interval.
end if.

leave tmpselnum.
execute.

select if SelVec > 0.
loop #i = 1 to SelVec.
xsave outfile = !quote(!concat(!dir,"\",!outfile,!s,".sav")).
end loop.
execute.

get file = !quote(!concat(!dir,"\",!outfile,!s,".sav")).

compute selseq = selseq + 1.
leave selseq.

mean tables = selvec !do !cv !in(!cvar) by !cv !doend
 / cells = sum min max count
 / missing = include.
compute !newwgt = tmpswgt / !nsel.

save outfile = !quote(!concat(!dir,"\",!outfile,!s,".sav"))
  / drop = tmpfirst tmplast tmpselnum tmpcmos tmpbmos tmpswgt !wgt.

new file.

!doend

erase file=!quote(!concat(!dir,"\",tmp0)).
erase file=!quote(!concat(!dir,"\",tmp1)).
erase file=!quote(!concat(!dir,"\",tmp2)).

!enddefine.
```

Figure 5: SPSS syntax to select mini-samples

```
include file = "C:\IERI_TechNote1\SamplePPS.spm".

samplepps dir     = C:\IERI_TechNote1
      / infile  = SamplePPSFrom
      / outfile = SamplePPSTo
      / idvar   = idschool idstud
      / cvar    = idcntry itsex
      / wgt     = totwgt
      / newwgt  = n_totwgt
      / nsel    = 500
      / nsamples= 10
      / seed    = 72864.
```

Figure 6: SPSS output from selecting mini-samples

COUNTRY ID	*SEX OF STUDENTS*	Sum	Minimum	Maximum	*N*
Australia	GIRL	500.00	1.00	1.00	500
	BOY	500.00	1.00	1.00	500
	Total	1,000.00	1.00	1.00	1,000
Bahrain	GIRL	500.00	1.00	1.00	500
	BOY	500.00	1.00	1.00	500
	Total	1,000.00	1.00	1.00	1,000
Armenia	GIRL	500.00	1.00	1.00	500
	BOY	500.00	1.00	1.00	500
	OMITTED	13,548.00	12.00	80.00	500
	Total	14,548.00	1.00	80.00	1,500
Bulgaria	GIRL	502.00	1.00	2.00	500
	BOY	500.00	1.00	1.00	500
	Total	1,002.00	1.00	2.00	1,000
Belgium (Flemish)	GIRL	500.00	1.00	1.00	500
	BOY	500.00	1.00	1.00	500
	Total	1,000.00	1.00	1.00	1,000
Total	GIRL	2,502.00	1.00	2.00	2,500
	BOY	2,500.00	1.00	1.00	2,500
	OMITTED	13,548.00	12.00	80.00	500
	Total	18,550.00	1.00	80.00	5,500

One additional point that we need to address when selecting mini-samples is that of the optimal sample size. There is no single answer to what this should be. In general, the samples selected should be large enough to reach desired effect sizes yet small enough to achieve the computational efficiencies sought. There should also be sufficient variability across the selected samples to obtain an optimal measure of variability of the estimates. In addition, we would recommend avoiding samples where the same records are selected multiple times in high frequencies.

References

Beaton, A., & Gonzalez, E. (1995). *The NAEP primer*. College Hill, MA: Center for the Study of Testing, Evaluation, and Educational Policy, Boston College.

Kalton, G. (1983). *Introduction to survey sampling* (SAGE university paper series, No. 35). Newbury Park, CA: SAGE Publications.

Kish, L. (1968). *Survey sampling*. New York, NY: Wiley.

Martin, M., Mullis, I., & Kennedy, A. (2007). *PIRLS 2006 technical report*. Chestnut Hill, MA: Boston College.

Olson, J., Martin, M., & Mullis, I. (2008). *TIMSS 2007 technical report*. Chestnut Hill, MA: Boston College.

Organisation for Economic Co-operation and Development (OECD). (2009). *PISA 2006 technical report*. Paris, France: Author.

Ross, K. (2005). *Sample design for educational survey research*. Paris, France: UNESCO.

INFORMATION FOR CONTRIBUTORS

Content

IERI Monograph Series: Issues and Methodologies in Large-Scale Assessments is a joint publication between the International Association for the Evaluation of Educational Achievement (IEA) and Educational Testing Service (ETS). The goal of the publication is to contribute to the science of large-scale assessments so that the best available information is provided to policy-makers and researchers from around the world. Papers accepted for this publication are those that focus on improving the science of large-scale assessments and that make use of data collected by programs such as IEA-TIMSS, IEA-PIRLS, IEA-Civics, IEA-SITES, U.S.-NAEP, OECD-PISA, OECD-PIAAC, IALS, ALL, etc.

If you have questions or concerns about whether your paper adheres to the purpose of the series, please contact us at IERInstitute@iea-dpc.de.

Style

The style guide for all IERI publications is the *Publication Manual of the American Psychological Association* (5th ed., 2001). Manuscripts should be typed on US letter or A4 format, upper and lower case, double spaced in its entirety, with one-inch margins on all sides. The type size should be 12 point. Subheads should be at reasonable intervals to break the monotony of lengthy text. Pages should be numbered consecutively at the bottom of the page, beginning with the page after the title page. Mathematical symbols and Greek letters should be clearly marked to indicate italics, boldface, superscript, and subscript.

Author Identification

The complete title of the article and the name of the author(s) should be typed only on the submission form to ensure anonymity in the review process. The pages of the paper should have no author names, but may carry a short title at the top. Information in the text or references that would identify the author should be deleted from the manuscript (e.g., text citations of "my previous work," especially when accompanied by a self-citation; a preponderance of the author's own work in the reference list). These may be reinserted in the final draft. The author (whether first-named or co-author) who will be handling the correspondence with the editor and working with the publications people should submit complete contact information, including a full mailing address, telephone number, and email addresses.

Review Process

Papers will be acknowledged by the managing editor upon receipt. After a preliminary internal editorial review by IERI staff, articles will be sent to two external reviewers who have expertise in the subject of the manuscript. The review process takes approximately three to six months. You should expect to hear from the editor within that time regarding the status of your manuscript. IERI uses a blind review system, which means the identity of the authors is not revealed to the reviewers. In order to be published as part of the monograph series, the work will undergo and receive favorable technical, substantive, and editorial review.

Originality of Manuscript and Copyright

Manuscripts are accepted for consideration with the understanding that they are original material and are not under consideration for publication elsewhere. If another version of the paper is being considered by another publication, or has been or will be published elsewhere (even as a working paper), authors should clearly indicate this at the time.

To protect the works of authors and the institute, we copyright all of our publications. Rights and permissions regarding the uses of IERI-copyrighted materials are handled by the IERI executive board. Authors who wish to use material, such as figures or tables, for which they do not own the copyright must obtain written permission from IERInstitute and submit it to IERI with their manuscripts.

Comments and Grievances

The Publications Committee welcomes comments and suggestions from authors. Please send these to the committee at IERInstitute@iea-dpc.de.

The right-of-reply policy encourages comments on articles recently published in an IERI publication. Such comments are subject to editorial review and decision. If the comment is accepted for publication, the editor will inform the author of the original article. If the author submits a reply to the comment, the reply is also subject to editorial review and decision.

If you think that your manuscript is not reviewed in a careful or timely manner and in accordance with standard practices, please call the matter to the attention of the institute's executive board.

Publication Schedule

The IERI Monograph is published annually, in October. Manuscripts can be submitted for review any time of the year. Most fitting to the review and editing process would be to submit a paper around 12 months before the monograph's publication date.

Manuscripts are reviewed and processed in the order they are received, and are published in the next available monograph if accepted for publication. If manuscripts move through the review and editing process considerably before publication of the print version, IERI offers an online version of it on the IERI website, in advance of its publication in the upcoming volume.

Each monograph consists of five to seven research papers. If, in a single year, there are more than seven accepted manuscripts, the editorial committee will determine whether the manuscript(s) will be published in the following monograph or in an additional monograph in the same year.

Submission of Articles

The monograph series *IERI Issues and Methodologies in Large-scale Assessment* welcomes the submission of original, research-based articles, in English. Manuscripts should be between 7,000 and 10,000 words in length, including abstract, notes, and references.

All contributors must submit:

- The text of the paper in editable format;
- All included tables and figures in editable format;
- A 100- to 150-word abstract briefly describing the content and central hypothesis of the paper;
- A completed article submission form, which can be found below or obtained from the IERI website: www.ierinstitute.org

IERI accepts only electronic submissions. Please send the manuscript to be considered for publication and any supplemental files (graphs and tables) to ierinstitute@iea-dpc.de.