Providing School-Level Reports from International Large-Scale Assessments: Methodological Considerations, Limitations, and Possible Solutions

Plamen Mirazchiyski
IEA Data Processing and Research Center

Providing School-Level Reports from International Large-Scale Assessments: Methodological Considerations, Limitations, and Possible Solutions

Plamen Mirazchiyski
IEA Data Processing and Research Center

March 2013

ISBN/EAN: 978-90-79549-19-1

Copies of this publication can be obtained from:

The Secretariat
International Association for
the Evaluation of Educational Achievement
Herengracht 487
1017 BT Amsterdam
The Netherlands

IEA Data Processing and Research Center
Mexikoring 37
22297 Hamburg
Germany

By email:
Department@iea.nl
mail@iea-dpc.de

Website:
www.iea.nl
www.iea-dpc.de

The International Association for the Evaluation of Educational Achievement, known as IEA, is an independent, international consortium of national research institutions and governmental agencies, with headquarters in Amsterdam. Its primary purpose is to conduct large-scale comparative studies of educational achievement with the aim of gaining more in-depth understanding of the effects of policies and practices within and across systems of education.

Production Editor
Marta Kostek IEA Data Processing and Research Center

Copyeditors: Paula Wagemaker Editorial Services, Christchurch, New Zealand, with David Robitaille
Design and production by Becky Bliss Design and Production, Wellington, New Zealand

Foreword

Countries that participate in international large-scale assessments (ILSA) are increasingly interested in providing schools with feedback on their performance. While ILSA are designed to provide detailed system and subnational-level information for monitoring purposes, greater public awareness about these assessments has stimulated an increased demand for information at the individual school level. In some countries, individual school feedback is considered to be an incentive for schools, principals, and teachers to participate in the studies.

This publication, however, considers some of the major challenges involved in providing valid school-level data. The author discusses in detail the design limits of ILSA, including sampling issues and such matters as the uncertainty of multiple matrix designs in terms of providing item-level reports. Other factors that need to be taken into account, the author argues, are methodological and statistical sources of error, as well as standards for calculations and reporting of errors.

Also considered are several options for changes in the assessment design of ILSA that would allow for the provision of feedback to small groups without compromising the studies' overall goals. In addition to describing some current practices for providing school-level data, the author suggests alternative strategies for developing reports. He also offers suggestions for variables that might be included and the level of detail that permits the provision of valid and reliable feedback to schools.

We would like to thank the National Center for Education Statistics (NCES) for supporting and funding this paper under contract No. ED-08-CO-0117 with the International Association for the Evaluation of Educational Achievement (IEA). Mention in this publication of trade names, commercial products, or organizations does not imply personal endorsement by the United States Government.

Dirk Hastedt

CO-DIRECTOR IEA DATA PROCESSING AND RESEARCH CENTER

CONTENTS

1 Introduction

The International Association for the Evaluation of Educational Achievement (IEA) has conducted international large-scale assessments (ILSA) for more than 50 years. The Pilot Twelve-Country Study, conducted between 1959 and 1962 (Foshay, Thorndike, Hotyat, Pidgeon, & Walker, 1962), and the First International Mathematics Study (FIMS), conducted from 1963 to 1967 (Husén, 1967), were among the association's first studies. Two of the better-known contemporary IEA studies are the Trends in International Mathematics and Science Study (TIMSS) and the Progress in International Reading Literacy Study (PIRLS). TIMSS measures the mathematics and science achievement of fourth and eighth grade students in four-year cycles; the first one took place in 1995, and the most recent one was completed in 2007 (Mullis et al., 2005). PIRLS assesses the reading literacy of Grade 4 students in five-year cycles. The first one was in 2001 and the second in 2006 (Mullis, Kennedy, Martin, & Sainsbury, 2006). The third cycle was conducted in 2011 (Martin, Mullis, Foy, & Stanco, 2012; Mullis, Martin, Foy, & Arora, 2012; Mullis, Martin, Foy, & Drucker, 2012).

The purpose of ILSA, since their onset, has been to compare the achievement of students from the participating countries in different subject domains. The participating schools within the countries invest time and effort to administer the assessments. Their only return on this investment is feedback on their students' performance. Some researchers (Australian Council for Educational Research, n. d.; Bos & Schwippert, 2003; Buckingham, 2008; PISA-Konsortium Deutschland & Leibniz-Institut für Pädagogik der Naturwissenschaften, n. d.; *Schoolfeedbackproject*, n. d.; Van Petegem & Vanhoof, 2004, 2005) outline or even recommend possible approaches for providing such feedback.

However, providing feedback to schools from ILSA may not be as straightforward as the process might seem. Sampling students

in ILSA is usually implemented with the intention of optimizing samples for reporting at the national—not the school—level. Under the general sampling design, student selection is carried out in two stages. During the first stage, schools are sampled using probability proportional to their size. In most of these large-scale studies, one or more (usually two) intact classrooms are selected within each school during the second stage. Because the purpose of ILSA is to provide aggregated results across many schools, this sampling approach is optimal because it allows the researchers involved to sample from a selection of many classrooms across the student population.

Selecting intact classes also has a practical advantage: students can be tested together in groups, in line with how they are organized in schools, reducing logistical demands. However, because of tracking (streaming), ability grouping, and other policies related to how students are assigned to classes, sampling intact classes within any one school, except in cases when all classes are selected, is not optimal in terms of providing individual schools with feedback on ILSA results. More specifically, selecting an intact class within a school does not take into account between-class variability and therefore is not optimal for reporting results at the school level.

An additional source of uncertainty in reporting results at the school level is the assessment design. Under the assessment designs used in ILSA, students are administered only a fraction of the pool of test items. As a consequence, there is uncertainty about the measurement of student proficiency, and this needs to be accounted for in the estimation procedures. Uncertainty is addressed via the use of multiple imputations, but the smaller the group of students for which the proficiency is estimated, the higher the uncertainty tends to be.

Although the text of this publication might give the impression that such reports should not be done because of the studies' designs and technical and statistical complexities, such feedback is possible under certain conditions, even if quite restricted. This publication seeks to discuss the current ILSA sampling and assessment design issues that pose limitations for reporting results at the school level. It also makes some recommendations on changes that could provide useful feedback. Additionally, the report offers an example structure and suggests what the content of such a report might be.

1.1 The Value of Feedback

In their paper on the meaning, importance, utility, and usefulness of school feedback, in general, Hattie and Timperly (2007) stress how feedback to schools, classes, and the students themselves promotes student learning: "Feedback is thus a 'consequence' of performance" (Hattie & Timperley, 2007, p. 81). However, it also has consequences for the future because "the feedback and instruction become intertwined" (p. 82).

Hattie and Timperly (2007) concluded from their meta-analyses of 196 studies that feedback generally has a positive effect on school performance (see also, in this regard, Schägen, Hutchinson, & Hammond, 2006). Hattie and Timperly furthermore identified effective feedback as feedback which provides answers to basic questions about student performance that students and their teachers tend to ask. These questions relate to how students have been performing, how they are currently performing, and where and how they need to go to next in order to meet achievement goals. "These questions," Hattie and Timperly explain, "correspond to notions of feed up, feed back, and feed forward" (p. 86). The authors also give an overview of the motivational effect that feedback has, especially in terms of helping teachers and students identify possible strategies for optimizing future learning activities. The main conclusion is that, in general, feedback has a positive effect (Hattie & Timperley, 2007; Schägen et al., 2006).

One of the main shortcomings of the practices reviewed in this publication is that most of them focus on the achievement part of the assessments, neglecting the context of the education that has important role for the outcomes. ILSA collect a large amount of background data on student, class, and school level that need to be accounted for and used in interpretation of the results. Feedback can also be used to compare schools with similar characteristics. Such comparisons, in turn, can help stakeholders identify possible reasons why schools with similar characteristics have different performance outcomes. Feedback can furthermore allow inferences to be made about schools that have similar characteristics to the ones that were studied. This information is generally useful for local educational

authorities, policymakers, and members of society who want to know how well the schools in their country or region are doing their job (Van Petegem & Vanhoof, 2004). However, as will be clarified in this publication, data from ILSA need to be handled with caution, and the limitations of the validity of information should be borne in mind by anyone attempting to use it to answer questions about school performance.

1.2 Who Needs the Reporting at School Level?

Two main parties have a vested interest in school-level feedback from ILSA—schools and educational authorities. For schools, feedback serves as a "mirror" that enables them to self-evaluate their effectiveness. Opportunity to compare the results of one's own school with other similar schools can help that school improve its work. Should the school identify its performance as satisfactory or poor, it can look for explanations, in general, identify reasons for certain results, in particular, and then take action (based on evidence). Whether or not the school scrutinizes the feedback and then acts on it depends, of course, on whether it considers taking action would be valuable, or even necessary. One thing is certain, though: the school must find the information contained in the feedback from ILSA relevant to its circumstances (Van Petegem & Vanhoof, 2004).

Governments also tend to be interested in school-level reporting for two reasons—promoting self-evaluation within schools ("health checks") and accountability purposes (ranking, using "league tables"). Governments are interested not only in schools' performance on ILSA but also in providing schools with the necessary information to, as just noted, facilitate self-evaluation and remedial action, where deemed necessary (Schägen et al., 2006).

1.3 Purpose of this Publication

In general, assessments provide stakeholders (including schools) with feedback on the various performance-related aspects of their respective education systems. Some assessments, however, are conducted specifically for the purpose of providing schools with feedback.

One such assessment is New Zealand's Assessment Tools for Teaching and Learning (asTTle), initiated by the country's Visible Learning Laboratories (Visible Learning Laboratories, n. d.). The asTTle project aims to provide schools with feedback on how well they are performing, the effectiveness of their curricula, the progress being made by their students, and so on. It also aims "to assist those involved in education with enhancing the teaching and learning opportunities" (Visible Learning Laboratories, n. d.).

In addition to conducting tests of educational achievement in key curriculum areas and providing schools with extensive reports on the results of those tests, the project endeavors to provide formative information that will have a direct effect on teaching and learning. asTTle thus provides a complete service, in terms of testing, delivering reports, and providing support to schools throughout New Zealand and, from there, the country's education system itself (Visible Learning Laboratories, n. d.).

As the previous paragraph implies, not all assessments are designed to provide feedback to schools since each assessment or type of assessment has its own goal or set of goals. The organizations conducting ILSA are mostly interested in making cross-country comparisons and identifying trends over time. Countries participating in ILSA are interested in producing national-level results and identifying patterns within certain regions or groups of schools and students in order to inform national educational policymaking.

Regardless of the level of the education system that an assessment targets (international, national, or subnational), education networks need to be able to disseminate information from assessments to stakeholders in a way that meets their various interests. Reporting also needs to be done in a manner that promotes collegial relationships among members of the educational community (Volante, 2006). However, despite ILSA having been conducted for the last 50 years, providing feedback that is timely and useful to those who can best make use of it in terms of student achievement seems to be an area that is underdeveloped. As Bos and Schwippert (2003, p. 571) advise, "some more research seems to be necessary about

the feedback strategies in general and limits and opportunities of feedback specifically."

Mindful of this advice, this publication provides an overview of the issues related to reporting results from ILSA to small groups, such as schools and the tested students within them. The publication discusses the limitations that this information holds for small groups and considers, in view of these limitations, how such information can best be presented to them. It also suggests changes in the assessment designs of ILSA that could make it possible to provide unbiased feedback to schools without interfering with the main objectives of the studies.

ILSA have complex designs, sampling strategies, and data-scaling procedures, all of which are designed to optimize measurement and reporting at national level while reducing the studies' operational procedures and costs. As a consequence, as the unit of analysis (region, district, city, school, class) becomes smaller, the risk of obtaining unreliable, or even invalid, results increases. While reporting results at the individual (student) level has never been the intention of ILSA (Rutkowski, Gonzalez, Joncas, & von Davier, 2010; von Davier, Gonzalez, & Mislevy, 2009), this publication was written on the premise, albeit expressed cautiously, that such studies could usefully consider means of reporting results to small groups such as schools and groups of students.

2 Current Practices in Providing Feedback to Schools

Discussion relating to ILSA results should not focus on the achievement data. Such an emphasis can provide a relatively narrow, biased view of the performance of tested students. Teams preparing reports for participating schools need to bear in mind Van Petegem and Vanhoof's (2004) main recommendation from their consideration of effective reporting of assessment results to schools. They advise that school background information (e.g., the socioeconomic status of a school's catchment area and its geographical placement—urban or rural), curriculum content, and the nature of the learning environment must be reported so as to promote realistic perceptions of the school factors influencing student achievement.

In Germany, reports to schools from the ILSA in which the country participates—TIMSS and PIRLS as well as the OECD's Programme for International Student Achievement (PISA)—contain both achievement information and background information particular to each school. The feedback in Germany also follows Van Petegem and Vanhoof's (2004) second main recommendation, namely to compare the results of tested students from schools and classes only with the results of other students and classes that possess similar characteristics (e.g., students' family socioeconomic status, school resources). Principals and teachers in Germany also receive assistance on how to interpret the results as well as training on how to use them (Bos & Schwippert, 2003).

In Germany, the PISA-Konsortium Deutschland and the Leibniz-Institut für Pädagogik der Naturwissenschaften (n. d.) prepare reports for the schools participating in PISA. During PISA 2006, Germany additionally sampled Grade 9 students in the selected schools, thereby supplementing the original sampling plan.

The report to each German school participating in PISA 2006 consisted of 28 pages. It included an overview of PISA (goals, description of the study), set out the purpose of the school feedback, and detailed the number of students in the school who participated in the study, along with their results on the mathematics, reading, and science tests, and the computed differences between boys' performance and girls' performance.

The report furthermore distinguished students' relative strengths in the three content domains and provided information pertaining to some student background characteristics associated with learning outcomes. The latter included students' interests and motivation relative to the content domains, students' self-perceived ability in those domains, the extent to which they were using computers and what they were using them for, and how the school was apportioning time to the content domains and other areas of the school curriculum and functioning. The report also offered an evaluation of school resources (e.g., number of students per computer).

The report additionally provided each school with a comparison of its average achievement results on the three content domains with the mean result across all other schools participating in PISA in Germany. The achievement level of each school was represented via a thermometer-like scale that ranged from 1 to 10 and that was set next to a scale that showed overall achievement for all participating schools (PISA-Konsortium Deutschland, & Leibniz-Institut für Pädagogik der Naturwissenschaften, n. d.).

In 2008, the Australian government considered adopting a school report-card system used in New York and Florida in the United States, in order to give schools feedback on student performance. Under this system, each school receives a letter grade. In her consideration of this system, Buckingham (2008) recommends (among other recommendations) that the schools' results should be publicly available. Buckingham backs up her recommendation through reference to the PISA 2006 international report: students in schools that published their results performed statistically significantly better than those that did not.

However, publishing results is akin to a chicken and egg situation in that schools whose students tend to perform well are probably more willing than schools whose students do not perform well to make their reports publicly available. In Buckingham's (2008) view, improving school performance needs incentives and poor performance needs penalties. And like the aforementioned authors, she recommends that schools take their background as well as their students' background characteristics into account when interpreting the achievement data.

The Australian Council for Educational Research (ACER) conducts the International Benchmark Tests (IBT) in English, mathematics, and science in 11 countries around the world, with tens of thousands of students, and links the results from IBT with the results on TIMSS from each country. It then provides each school and even individual students with a report that compares the IBT results with the TIMSS mathematics and science test results (ACER, n. d.). Each report is a one-page document that contains no technical information on how ACER performs this linkage and how it makes comparisons at school and class levels.

The student-level report refers to "your child" and presents the individual student's result against the average results (expressed with error-bar-like graphical elements) from 11 countries that participated in TIMSS 2003. The school-level report presents a distribution of the results within the school and compares the school's modal score against the average achievement scores of several of the TIMSS 2003 countries. Both reports contain instructions on how to read and compare the school's or student's score and make comparisons. Each report states its purpose and provides a description of TIMSS. The reports can be found on ACER's website (ACER, n. d.), but no methodological explanations are available to the public.

The "Schoolfeedbackproject" in the Flemish part of Belgium is an initiative for providing feedback to schools on their students' performance in international and national assessments—"A mirror for every school," as those associated with the project put it (*Schoolfeedbackproject*, n. d.). The initiative is a joint effort between the Center for Educational Effectiveness and Evaluation at Leuven

University, Ghent University, and the University of Antwerp. The Schoolfeedbackproject website provides information on the initiative, information on its studies, a sample report, and a reading guide that accompanies the report. At the time of writing this publication, the only feedback offered was for PIRLS 2006.

The sample report (per school) is a 20-page document which contains an overview of PIRLS, explanations on the structure and purpose of the report, and a methodological overview of the study. The results section consists of three chapters:

1. The school's reading comprehension results (raw and corrected scores) compared to the average achievement score for all of Flanders;
2. The intake characteristics of the students (personal, ethnocultural, and sociocultural); and
3. A comparison of the results across the classes within the school.

The reported data and their graphical representation are provided before and after the results were altered to take account of the students' intake characteristics. The reading guide provides a dictionary of terms and instructions on how to read and interpret the graphs. The website also provides a bibliography of school-report publications, most of which are written in Dutch (Schoolfeedbackproject, n. d.).

In Flanders, ILSA results are reported after an informal directive from the Flemish Ministry of Education. The ministry's decision on which feedback to provide to schools is based on its determination of which data are most likely to have an advantageous impact on schools' performance. One of two such reports sent out to schools focused on TIMSS 1999 data and the other on PISA 2000 data (Van Petegem & Vanhoof, 2005).

The TIMSS 1999 report used error-bars, adjusted for student intake, as well as unadjusted results in the different content domains that the tests cover (mathematics and science). Presentation of adjusted and unadjusted results allows readers to judge how effective a particular school is in terms of student performance once the results have been adjusted for student intake characteristics (Van Petegem & Vanhoof, 2005).

The TIMSS 1999 report also contains scatterplots—one per content domain. These show the relationship between achievement in the domain and a measure of intelligence for each student in the school (not collected by TIMSS 1999). The scatterplots furthermore show the overall relationship between the achievement and measure of intelligence and indicate how this relationship compares to the overall relationship shown by all other schools in Flanders. The same scatterplots furthermore show the differential effectiveness (in terms of student achievement) of the separate classes and highlight the heterogeneity of the relationship between achievement and intelligence scores (Van Petegem & Vanhoof, 2005).

The last section in the report enables readers to compare the background information of a school and its tested classes with the performance of all other schools and classes participating in the study. This information is reported for three levels: student, class, and school (Van Petegem & Vanhoof, 2005).

The PISA 2000 report was organized differently. The main figure contains scatterplots for all schools, each indicated by a different type of symbol that represents its students' achievement and their socioeconomic status (SES). The particular school for which the report is intended is marked by a red symbol. Different colors are used to mark the international and the national (Flemish) gradients on the figure (see Van Petegem & Vanhoof, 2005).

Except for the feedback system that Buckingham (2008) described, all other feedback systems described recommend that the feedback sent to each school on its students' ILSA performance remains anonymous. Also, the practices presented use different approaches and rarely present, let alone explain, methodological considerations. Nor do they provide indepth information on design, measurement, and sampling issues. An overview of these issues is the purpose of the next section of this publication.

3 Sampling and Assessment Designs and Their Implications for Providing Feedback to Schools

TIMSS and PIRLS are curriculum-based studies because they take the curriculum as the major guiding principle of teaching and learning. As the authors of the TIMSS framework state, "TIMSS uses the curriculum, broadly defined, as the major organizing concept in considering how educational opportunities are provided to students, and the factors that influence how students use these opportunities" (Mullis et al., 2005, p. 4). According to the framework authors, the rationale behind using the curriculum as the basis of the assessment can be expressed as follows: "The curriculum reflects the needs and aspirations of the students, the nature and function of learning, and the formulation of statements on what learning is important" (Mullis et al., 2005, p. 82).

TIMSS surveys two subject domains, mathematics and science, and splits each into several topics or subdomains in the assessment instruments. Each topic is represented by a list of objectives that are covered by the curricula of the majority of countries taking the assessment (Mullis et al., 2005). Three different kinds of curricula are distinguished:

1. Intended—defined by national, social, and educational contexts;
2. Implemented—by the schools' teachers; and
3. Attained—by the students (Mullis et al., 2005).

This curriculum-based approach is one of the major strengths of the study: collecting information on the intended, implemented, and attained curriculum helps policymakers and curriculum development specialists judge how well an education system is performing. In order to collect information on the intended curriculum, TIMSS asks the TIMSS national research coordinators within the participating countries to complete the study's online

curriculum questionnaire (Mullis & Martin, 2008). The TIMMS school and teacher questionnaires, which TIMSS uses to collect contextual data, gather up information on the implemented curriculum (Mullis et al., 2005).

PIRLS also uses questionnaires to collect data pertinent to learning contexts. As is the case with TIMSS, this background information pertains to students and teachers and their schools. Because TIMSS tests two subjects, the teachers who teach the sampled students in mathematics and science receive two different questionnaires, one for each subject (Mullis et al., 2005). PIRLS 2006 collected student, teacher, and school background data, and, like TIMSS, asked only those teachers of the classes selected for testing to complete a teacher questionnaire. Unlike TIMSS, however, PIRLS also collects home background data from the students' parents (Mullis et al., 2006).

The complexity of ILSA studies has consequences for reporting data for small groups, such as schools, because of the uncertainty occasioned by the sampling and assessment designs. This uncertainty then becomes the product of both the sampling and the imputation variances, which, in turn, together represent the standard error of any reported estimate. It is imperative that anyone considering a result takes the standard error into account when interpreting certain statistics.

3.1 Sampling Design and Implications for School-Level Reporting

TIMSS 2007 had two target populations of students. The study defined the Grade 4 target population as the grade where students have had four years of formal schooling, counting from the first year of ISCED Level 1, provided the average student age was 9.5 years or higher. In most countries, this target population coincided with Grade 4, or the fourth year of formal schooling. The Grade 8 student population was defined in a similar way. It took the eighth year of schooling, counting again from the first year of ISCED Level 1, provided the average student age was at least 13.5 years. This population coincided with Grade 8 in most countries (Joncas, 2008). The way in which the target population in PIRLS 2006 was defined

is similar to the approach used to define the TIMSS Grade 4 target populations (Joncas, 2007).

TIMSS and PIRLS use the same sampling strategy—a two-stage stratified cluster sampling design. During the first stage, the schools, which are the primary sampling units (PSU), are selected using systematic random sampling, with probability proportional to their size. Each country is expected to sample no fewer than 150 schools (Joncas, 2007, 2008). During the second sampling stage, one or two intact classes (depending on the required sample size and the size of the schools' target populations) are sampled at random in each selected school. The total number of selected students in PIRLS 2006 and TIMSS 2007 was at least 4,000 for each target population (Joncas, 2007, 2008).

In cases where schools are relatively small, with one or two classes, it is operationally simpler to select all classes, and hence students. This approach would facilitate reporting results for these schools because the entire target population in each school is tested. However, in larger schools, that is, those schools with more than two classes, selecting one or two classes might not yield a representative sample of the students in the school, unless we assume that all classes are equivalent to one another. When the between-class variance is relatively low, and the within-class variance is the same across the classes, randomly selecting one class within the school will generally suffice. But if there are larger differences between the classes, and therefore between the within-classroom variances, selecting one or two classes is not sufficient to obtain stable estimates of the school's performance.

A short example follows. Let us say two schools (A and B) are in the sample, and each one has five classes in the target population. On average, students within each of these schools perform about the same. One intact class is randomly selected from each school. However, through chance, the poorest performing class in School A is selected, while in School B the best performing class is selected. In both cases, the sampled classes are not representative of their schools because they do not represent the variety of their students' characteristics (i.e., abilities on the subject tested and their

background). Using a single class to report results for the schools would likely result in unstable estimates and would not give us an idea of the variability between the classes.

While sampling intact classes across the selected schools yields a representative sample of students across the country, this does not necessarily result in a representative sample of students within any one school, except in small schools where all classes, and therefore all students, have been selected. As a consequence, in large schools, inferences can be made only for the sampled and eventually tested classes, but not for all the students in the target grade in the school. School-level reporting from studies that follow the current TIMSS and PIRLS sampling designs should not be couched in terms of "reporting the school-level results," but rather in terms of reporting the "results of the sampled class(es) in the school."

3.2 Assessment Design and Implications for School-Level Reporting

When reporting the results at any level, particularly for smaller groups, it is necessary to bear in mind that ILSA attempt to measure broad subject content domains. As such, the number of items necessary to measure the domain is relatively large, making it impossible to test everyone on everything. The assessment design provides a blueprint for how the items will be allocated to students.

The assessment designs that are used in ILSA use multiple-matrix sampling. This approach means that no student takes all items, and no student receives all items. However, there have been exceptions to this practice, as was the case in TIMSS 1995 when so-called anchor items were administered to all participating students (see Martin, 1996). The multiple-matrix approach results in some uncertainty about how well the test measures the content-domain abilities of any one student, and any one reported group. This outcome is referred to as measurement error or uncertainty.

The measurement error reflects the imprecision of the measurement of the domain. Usually, the longer a test is in terms of number of items, the more precise and reliable is its measurement. ILSA are not intended to provide estimates for the individual student because the measurement errors are relatively large. As Wu (2010, p. 18) reminds

us, the "measurement error will be reduced if we are interested in tracking groups of students, such as in a class or in a school. However, the measurement error would still be relatively large for a group/class of 30 students."

In order to assess student proficiency reliably, a large number of items are required. This is a general prerequisite of efforts to minimize the measurement error (Foy, Galia, & Li, 2007, 2008). However, students cannot respond to the entire item pool, which can be quite big, because of time and financial constraints, fatigue, boredom, and the like. The instruments measuring achievement in TIMSS and PIRLS are composed of blocks of items. Each student is administered only a fraction of all items (both multiple choice and constructed response) in the study's item pool. Each text booklet is linked with the next and/or the previous booklet by a common block of items.

TIMSS 2007 had 28 item blocks (14 in mathematics and 14 in science) distributed across 14 booklets. Each booklet contains two blocks of mathematics and two blocks of science items. Each second mathematics and science block in a booklet appears in the next booklet to ensure linkage across all booklets (Mullis et al., 2005). PIRLS 2006 had 10 blocks of achievement items rotated in 12 booklets plus one separate "reader," which is the only place where two of the blocks appear (Mullis et al., 2006).

In addition to minimizing time constraints and preventing burden and fatigue in students (which usually generates poor responses), such designs result in more adequate coverage of the content domain because they distribute the items across students (Cronbach, Linn, Brennan, & Haertel, 1995). This approach, however, results in some uncertainty because each student responds only to the questions in the booklet he or she receives and not to the entire item pool. Student answers on the entire item pool can be estimated, but because these estimations are probabilistic in nature, there is some uncertainty with respect to the validity of the final achievement scores. This uncertainty—the measurement variance (also called imputation variance)—is a second component of error evident in the reporting of ILSA results (Foy et al., 2008).

The following sections briefly describe and discuss the different methods for obtaining students' proficiency scores and the issues that arise when these are made part of reports to small groups.

3.2.1 Plausible values

TIMSS and PIRLS use item response theory (IRT) models developed by the United States' Education Testing Service (ETS) for the National Assessment of Educational Progress (NAEP) in that country (Foy et al., 2008). Three latent-variable models are used. The first is a three-parameter model for the multiple-choice items, the second is a two-parameter model for the open-ended items that are dichotomously scored, and the third is a partial-credit model for the open-ended items with polytomous scoring schemes (Foy et al., 2007, 2008). The proficiency score is treated as an unknown variable for the sampled students. Multiple imputation techniques are used to generate plausible values (PVs).

The data for each student are taken from the achievement items to which the student responded as well as from his or her responses on the background questionnaire. All these variables (achievement and background) are used together in a process called "conditioning." To reduce the large number of background variables, principal component analysis (PCA) is performed, and only those components accounting for 90 percent of the common variance in the data are selected. The PCA is performed separately per country because the number of variables in each country differs. Given that a student responds to only a portion of the total number of achievement items, and presumably to all background questions, student proficiency can be estimated from his or her conditional distribution with a known mean and dispersion. PVs are generated from this multivariate normal distribution, a practice that helps to quantify the uncertainty of the imputation (Foy et al., 2007, 2008).

PVs are not intended to report individual results because of their level of uncertainty at the individual level. Also, "plausible values are not test scores for individuals in the usual sense, but rather are imputed values that may be used to estimate population characteristics correctly" (Foy et al., 2007, p. 155). PVs are deemed good group estimates because they add the right amount of variability and so

"make the distribution of PVs in the group match the distribution of the true values in the group" (von Davier et al., 2009, p. 35).

Reporting results obtained using PVs is intended for a population or groups within a population (Rutkowski et al., 2010; von Davier et al., 2009). These groups cannot be too small because the errors associated with the estimates become too big (von Davier et al., 2009). Rutkowski et al. (2010) illustrate the influence of the group size on the uncertainty of the reading achievement estimate obtained using plausible values when the level of reporting (i.e., group size) gradually decreases until it reaches the individual student. When the reporting is at country level, the largest difference in the mean estimates between the five plausible values is less than a score point. When the level of reporting is at school level, the largest difference is about four score points. When the analysis is performed at class level, the differences increase to 11 score points. And when an individual student is taken as the level of reporting, the maximum observed difference shifts to almost 58 score points (Rutkowski et al., 2010).

This example clearly shows how the uncertainty increases with decreasing group size. Someone might argue that the maximum difference between the means of the separate plausible values at school level is relatively small (a little less than four score points). However, what needs to be remembered is that selecting different units (country, school, class, student) when computing the means of the plausible values can lead to much larger differences.

What does all this mean in terms of analysis for providing feedback to schools on the performance of their tested students? The following examples illustrate the issues related to providing such feedback to each school. These examples, using data from the TIMSS 2007 international database (IEA, 2007), show the relationship between the number of elements in a sample and the reliability of the group estimate.

In the first example, the mean achievement for each school in each country was calculated along with its measurement error. The measurement error was then correlated with the number of students taking the test in each school within each single country. Tables 1 and 2 present the results of these calculations.

Table 1: Relationship between the measurement error and the number of sampled students per school (Grade 4)

Country	Correlation coefficients		Country	Correlation coefficients	
	Mathematics	*Science*		*Mathematics*	*Science*
Algeria	-0.25 **	-0.17 *	Lithuania	-0.65 **	-0.66 **
Armenia	-0.55 **	-0.5 **	Mongolia	-0.36 **	-0.22 **
Australia	-0.5 **	-0.48 **	Morocco	-0.5 **	-0.52 **
Austria	-0.53 **	-0.63 **	Netherlands	-0.47 **	-0.46 **
Chinese Taipei	-0.45 **	-0.39 **	New Zealand	-0.59 **	-0.57 **
Colombia	-0.63 **	-0.65 **	Norway	-0.58 **	-0.59 **
Czech Republic	-0.51 **	-0.59 **	Qatar	-0.56 **	-0.54 **
Denmark	-0.55 **	-0.56 **	Russian Federation	-0.57 **	-0.6 **
El Salvador	-0.58 **	-0.56 **	Scotland	-0.57 **	-0.51 **
England	-0.54 **	-0.53 **	Singapore	-0.02	-0.11
Georgia	-0.56 **	-0.64 **	Slovak Republic	-0.63 **	-0.61 **
Germany	-0.48 **	-0.34 **	Slovenia	-0.55 **	-0.55 **
Hong Kong SAR	-0.33 **	-0.36 **	Sweden	-0.59 **	-0.57 **
Hungary	-0.55 **	-0.59 **	Tunisia	-0.56 **	-0.58 **
Iran, Islamic Republic of	-0.57 **	-0.64 **	Ukraine	-0.54 **	-0.56 **
Italy	-0.38 **	-0.32 **	United Arab Emirates (Dubai)	-0.59 **	-0.59 **
Japan	-0.45 **	-0.48 **	United States	-0.46 **	-0.36 **
Kazakhstan	-0.45 **	-0.45 **	United States (Massachusetts)	-0.29 *	-0.19

Table 1: Relationship between the measurement error and the number of sampled students per school (Grade 4) *(contd.)*

Country	Correlation coefficients		Country	Correlation coefficients	
	Mathematics	*Science*		*Mathematics*	*Science*
Kuwait	-0.34 **	-0.35 **	United States (Minnesota)	-0.68 **	-0.61 **
Latvia	-0.61 **	-0.69 **	Yemen	-0.43 **	-0.52 **
Minimum	-0.68	-0.69			
Median	-0.55	-0.55			
Maximum	-0.02	-0.11			

Note: * $p < 0.05$, one-tailed, ** $p < 0.01$, one-tailed.
Source: IEA's Trends in International Mathematics and Science Study (TIMSS) 2007.

Table 2: Relationship between the measurement error and the number of sampled students per school (Grade 8)

Country	Correlation coefficients		Country	Correlation coefficients	
	Mathematics	*Science*		*Mathematics*	*Science*
Algeria	-0.28 **	-0.23 **	Kuwait	-0.36 **	-0.32 **
Armenia	-0.47 **	-0.60 **	Lebanon	-0.63 **	-0.52 **
Australia	-0.48 **	-0.43 **	Lithuania	-0.67 **	-0.64 **
Bahrain	-0.57 **	-0.52 **	Malaysia	-0.32 **	-0.28 **
Bosnia and Herzegovina	-0.46 **	-0.40 **	Malta	-0.55 **	-0.61 **
Botswana	-0.09	-0.10	Mongolia	-0.19 *	-0.34 **
Bulgaria	-0.58 **	-0.56 **	Morocco	-0.30 **	-0.21 **
Chinese Taipei	-0.25 **	-0.35 **	Norway	-0.50 **	-0.52 **
Colombia	-0.29 **	-0.38 **	Oman	-0.33 **	-0.29 **
Cyprus	-0.44 **	-0.59 **	Palestinian National Authority	-0.29 **	-0.33 **
Czech Republic	-0.41 **	-0.45 **	Qatar	-0.61 **	-0.64 **
Egypt	-0.24 **	-0.26 **	Romania	-0.59 **	-0.64 **
El Salvador	-0.49 **	-0.50 **	Russian Federation	-0.51 **	-0.52 **
England	-0.36 **	-0.53 **	Saudi Arabia	-0.48 **	-0.50 **
Georgia	-0.64 **	-0.59 **	Scotland	-0.48 **	-0.40 **
Ghana	-0.44 **	-0.48 **	Serbia	-0.50 **	-0.49 **
Hong Kong SAR	-0.27 **	-0.42 **	Singapore	0.06	-0.05
Hungary	-0.60 **	-0.57 **	Slovenia	-0.55 **	-0.54 **
Indonesia	-0.46 **	-0.40 **	Sweden	-0.42 **	-0.44 **
Iran, Islamic Republic of	-0.41 **	-0.42 **	Syria, Arab Republic of	-0.28 **	-0.28 **

Table 2: Relationship between the measurement error and the number of sampled students per school (Grade 8) *(contd.)*

Country	Correlation coefficients		Country	Correlation coefficients	
	Mathematics	*Science*		*Mathematics*	*Science*
Israel	-0.31 **	-0.22 **	Thailand	-0.51 **	-0.42 **
Italy	-0.47 **	-0.43 **	Tunisia	-0.38 **	-0.23 **
Japan	-0.34 **	-0.39 **	Turkey	-0.39 **	-0.38 **
Jordan	-0.48 **	-0.29 **	Ukraine	-0.47 **	-0.54 **
Korea, Republic of	-0.25 **	-0.25 **	United States	-0.40 **	-0.37 **
Minimum	-0.67	-0.64			
Median	-0.44	-0.42			
Maximum	0.06	-0.05			

Note: * $p < 0.05$, one-tailed, ** $p < 0.01$, one-tailed.
Source: IEA's Trends in International Mathematics and Science Study (TIMSS) 2007.

As is apparent from the Grade 4 mathematics component of Table 1, 33 out of the 40 countries listed have correlation coefficients lower than r = -0.40. For the Grade 4 science outcomes shown in Table 1, the correlation coefficients are lower than r = -0.40 in 30 countries out of 40. In Grade 8 (Table 2), the correlation coefficients in 30 of the 50 countries listed are lower than r = -0.40 for mathematics and likewise in 29 out of the 50 countries for science.

For any grade or subject, the correlation coefficient is statistically significant ($p < 0.01$, one-tailed) for nearly all countries. The negative sign in the correlation coefficient means that the lower the number of tested students per school becomes, the higher the measurement error tends to be.

Among the countries showing the strongest negative correlations in Grade 4 for both mathematics and science are the United States (r = -0.68 and r = -0.61), Lithuania (r = -0.65 and r = -0.66), Colombia (r = -0.63 and r = -0.65), the Slovak Republic (r = -0.63 and r = -0.61), and Latvia (r = -0.61 and r = -0.69). For Grade 8, the countries showing the strongest negative correlations are Lithuania (r = -0.67 and r =- 0.64), Georgia (r = -0.64 and r = -0.59), Lebanon (r = -0.63 and r = -0.52), Qatar (r = -0.61 and r = -0.64), and Hungary (r = -0.60 and r = -0.57). For these groups of countries, there is thus an inverse relationship between the sample size within the school and the amount of error, or uncertainty, of the estimate.

Of the countries included in the two tables, Singapore has the weakest correlation coefficients—close to zero—between the number of tested students per school and the measurement error for both mathematics and science in Grades 4 and 8: r = -0.02 for Grade 4 mathematics, r = -0.11 for Grade 4 science, r = 0.06 for Grade 8 mathematics, and r = -0.05 for Grade 8 science. For Grade 8 mathematics, the correlation coefficient is actually positive (i.e., the lower the number of tested students per class, the lower the measurement error).

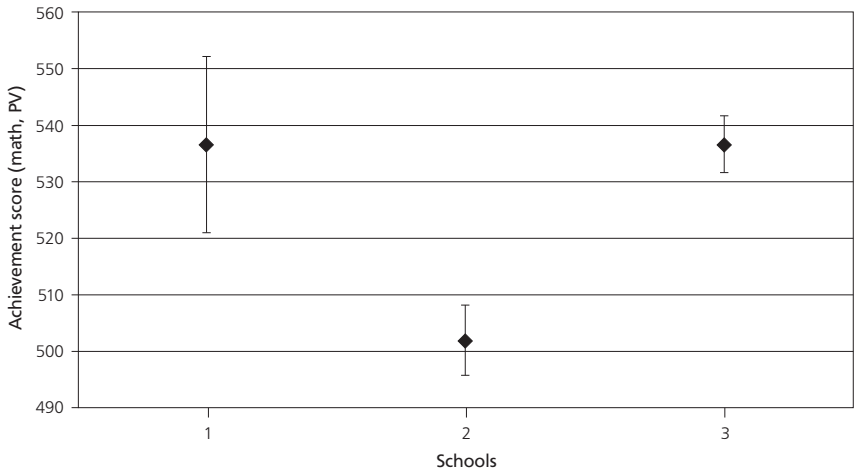
This outcome is a result of the sampling strategy that Singapore used. There, two classes were selected in each school, and 19 students were then randomly subsampled within each class (Joncas,

2008). The sample size across all schools was therefore relatively constant, placing an upper bound to the correlation coefficient.

The second example demonstrates that even when each participating school has the same number of tested students, the measurement error can still vary considerably. Figure 1 sets out the confidence intervals of the mean mathematics achievement scores of three United States schools with the same number of tested Grade 8 students ($N = 34$). Although the three schools have the same number of tested students, the first school has a confidence interval that is twice as wide as the intervals of the other two schools. Even though the first and the third school have the same number of tested students and the same mean achievement score ($X = 536$), the confidence interval of the first school is twice as wide as the confidence interval of the second school. This example is not an isolated case. The same analysis conducted with data from different countries with schools that had the same number of students showed similar results.

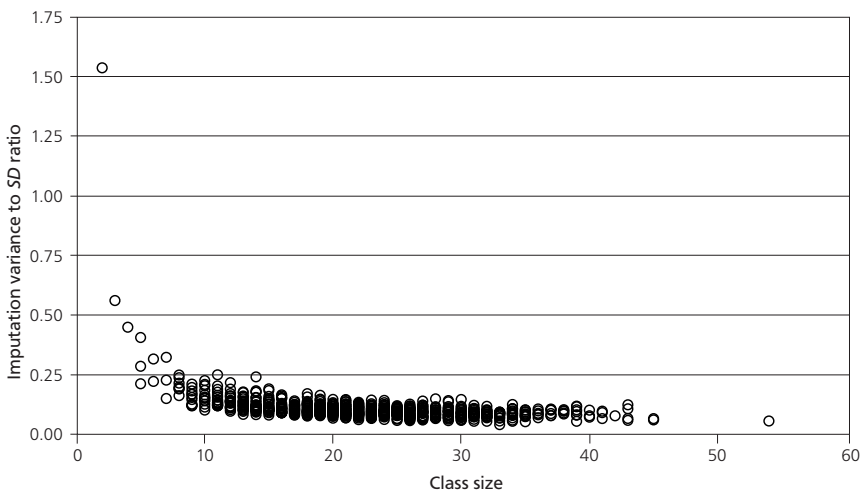
The third example demonstrates the relationship between class size and dispersion of the imputation variance of mathematics achievement scores. Data were drawn from the classes of all countries that participated at the Grade 8 level in TIMSS 2007. The horizontal axis in Figure 2 represents the class sizes across all countries, and the vertical axis shows the ratio of the imputation error and the aggregated standard deviations of the five PVs. As the figure highlights, as classes become larger, the dispersion of the imputation error becomes lower, and vice versa, thus indicating a non-linear relationship. When the class size decreases to 15 students or fewer, the dispersion of the imputation variance becomes higher and the dots become more scattered.

Figure 1: Measurement error given equal school sizes



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

Figure 2: Relationship between the class size and the ratio between the imputation variance and the achievement standard deviation



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

3.2.2 Using other scores to report school-level results

It might be tempting for someone reporting results from an ILSA to take a shortcut with respect to the plausible values and calculation of the imputations variance by reporting the students' responses to achievement items in terms of raw scores or percent correct. The advantage for that person would be the easy calculation of the raw score for each student and then the ease of calculating the mean for all tested students within a school. However, this approach would yield results that are not comparable to the international results, thus making it a less attractive approach for reporting the results of the tested students per school. Remember that each student is administered only a portion of items. A raw score or percent correct from a portion of all items is, of course, affected by the particular selection of items that the student receives. While the expectation is that all test booklets are of approximately equal difficulty, this is not always the case. Raw scores or percent correct scores are therefore neither optimal nor optional for reporting results when these assessment designs are used.

IRT-based scores could be used to overcome the differential difficulty across the test booklets. Examples of such scores include expected-a-posteriori (EAP), multi-group expected-a-posteriori (EAP-MG), maximum likelihood estimation (MLE), and the Warm's weighted likelihood estimation (WLE). However, use of these scores is not recommended when reporting ILSA results. A short overview and rationale for avoiding their use follows.

IRT provides a student-ability estimate as the score most likely to have yielded the particular response pattern observed. WLE and EAP scores derive the same value from the posterior distribution for all students exhibiting the same response pattern because the posterior distribution for these students is the same (Wu, 2005). To put this point another way, these scores are discrete point estimates that do not take into account the fact that ability in a population is a continuous variable which has some uncertainty. What we are interested in when talking about an estimate is not only a particular number as an estimate of the ability, but also the uncertainty associated with it. Wu (2005) demonstrates that although MLE, WLE,

and EAP provide unbiased estimates for the population mean, the variance of the mean ability for WLE and MLE is overestimated and for EAP is underestimated, meaning that there is a greater degree of uncertainty with these scores than with PVs. The variance estimate (bias) for WLE and MLE tends to increase for shorter tests (Wu, 2005).

Also, as von Davier et al. (2009) note, marginal maximum likelihood and EAP estimates are optimal for individuals (who are not the focus of ILSA), but not for groups of individuals because the estimates will be biased. Using simulation methods, von Davier and his colleagues showed that although the WLE, MLE, and EAP methods provide fairly accurate estimates for the means, the WLE overestimates the standard deviation while a decrease in test-item numbers leads to its underestimation. The researchers demonstrated that, in general, EAP and even EAP-MG also underestimate the standard deviations. They also demonstrated that when subgroups within the populations are analyzed, the between-group differences are captured neither by WLE nor EAP nor EAP-MG. In contrast, PVs provide unbiased estimates both for the means and the standard deviations for subgroups (von Davier et al., 2009).

Similar limitations become evident when WLE, MLE, and EAP are used to estimate statistics other than means, for example, percentiles (von Davier et al., 2009; Wu, 2005). Using TIMSS 2007 data from 49 countries, Carstens and Hastedt (2010) shed light on the implications of this use for analysis. While the mean estimates that arose out of using the WLE, MLE, EAP, and EAP-MG methods were relatively the same compared to the estimates derived from using PVs, the standard errors were completely different; in some countries, the differences were over 50 percent (Carstens & Hastedt, 2010). When Carstens and Hastedt (2010) used shorter scales instead of the overall mathematics scale, they found an even greater effect on the estimates.

3.2.3 Reporting item statistics at the school level

As described earlier, ILSA use multiple matrix sample designs to assign items to students. While it might be of interest to report results at the item level, the uncertainty pertaining to these statistics is also large.

In schools that have a large number of sampled students, a full rotation of the booklets (12+1 in PIRLS 2006 and 14 in TIMSS 2007) across students might be accomplished more than once. But what about small schools where the number of students is much smaller, in particular smaller than the number of booklets in the study?

Because ILSA use a matrix-sample design, there will be, in line with school size, different numbers of students taking the individual item blocks as well as the items that are rotated within the achievement booklets. With this booklet design, each student exhibits only a sample of the behaviors being measured, and any one item is administered to very few students. The following example shows that the distribution of items and item blocks across students can be disproportionate depending on the number of sampled students per school.

Let us take the TIMSS 2007 test booklet design as shown in Table 3. Each booklet has two mathematics and two science blocks, which means that each block of items (science or mathematics) and each item respectively appears twice in the rotation within the 14 test booklets. Assume there are four different schools. The first one has 15 sampled students in total, the second has 22, the third 28, and the fourth 33. Also assume that in each school the booklet rotation starts from the first booklet and that no students within these four schools were absent from the testing session. Then, in the first school (15 students in total), there will be one full rotation of the booklets (1 to 14), and Booklet 01 will be answered twice because of the 15th student. The second school will have the first eight booklets rotated twice—once for Students 1 to 14 and then Booklets 1 to 8 for Students 15 to 22. For the third school, all achievement booklets will be rotated twice (14 booklets for 28 students). The fourth school will experience two full rotations of the 14 booklets and then a third rotation of the first five booklets.

Table 3: TIMSS 2007 test booklet design and numbers of students taking the separate achievement booklets

Booklet	Item blocks		Number of students per booklet			
	Part 1	*Part 2*	*School 1* (*n* = 15)	*School 2* (*n* = 22)	*School 3* (*n* = 28)	*School 4* (*n* = 33)
Booklet 01	M01 M02	S01 S02	2	2	2	3
Booklet 02	S02 S03	M02 M03	1	2	2	3
Booklet 03	M03 M04	S03 S04	1	2	2	3
Booklet 04	S04 S05	M04 M05	1	2	2	3
Booklet 05	M05 M06	S05 S06	1	2	2	3
Booklet 06	S06 S07	M06 M07	1	2	2	2
Booklet 07	M07 M08	S07 S08	1	2	2	2
Booklet 08	S08 S09	M08 M09	1	2	2	2
Booklet 09	M09 M10	S09 S10	1	1	2	2
Booklet 10	S10 S11	M10 M11	1	1	2	2
Booklet 11	M11 M12	S11 S12	1	1	2	2
Booklet 12	S12 S13	M12 M13	1	1	2	2
Booklet 13	M13 M14	S13 S14	1	1	2	2
Booklet 14	S14 S01	M14 M01	1	1	2	2

Note: M = mathematics block, S = science block.
Source: IEA's Trends in International Mathematics and Science Study (TIMSS) 2007.

Because the separate blocks appear twice in two sequential booklets (e.g., Block M02 appears in Booklets 1 and 2, ensuring a link between them), different numbers of students respond to each block of items (and therefore each item). Table 4 shows the number of students given each item block for each of the four schools. As the table shows, the number of students taking each one of the test blocks/items varies across the schools. The only school where each block is taken by equal numbers of students is the third one, which has 28 students. Here, each of the booklets is taken by two students and each block by four. The highest variation is in the second and the fourth school ($n = 22$ and $n = 33$, respectively) where the rotation of the booklets is not equal to a whole number. Note that this outcome holds only if none of the sampled students was absent during the testing session. In the case of absences, the proportions of responded items would be scarcer still.

Table 4: Number of students in schools taking each item block (TIMSS 2007)

Mathematics/ science block	Number of students per item block			
	School 1 (*n* = 15)	*School 2* (*n* = 22)	*School 3* (*n* = 28)	*School 4* (*n* = 33)
M01/S01	3	3	4	5
M02/S02	3	4	4	6
M03/S03	2	4	4	6
M04/S04	2	4	4	6
M05/S05	2	4	4	6
M06/S06	2	4	4	5
M07/S07	2	4	4	4
M08/S08	2	4	4	4
M09/S09	2	3	4	4
M10/S10	2	2	4	4
M11/S11	2	2	4	4
M12/S12	2	2	4	4
M13/S13	2	2	4	4
M14/S14	2	2	4	4

Note: M = mathematics block, S = science block.
Source: IEA's Trends in International Mathematics and Science Study (TIMSS) 2007.

This example shows two important features. First, the responses missing by design have a different pattern across the schools. Remember that, because of the aggregation of items into blocks and their rotation across the booklets, there are many missing values, which means that these occur because of the design itself. Each school has a different number of students working on the individual booklets, blocks, and items. The different numbers of tested students per school mean that the number of students responding to each item will again be different (as shown in Table 4). The second feature of note is that very few students in any of these schools answer any one item. As a consequence, reporting results at the item level when reporting results to the schools is not recommended.

3.3 Other Sources of Errors

Besides the sampling and measurement issues discussed thus far, there are sources of error that are beyond the methodological and statistical ones. As Wu (2010) points out, test administration, scorer reliability, and item and test bias can contribute to the overall error of the estimates. The reliability of the scoring can introduce a large amount of error (Cronbach et al., 1995). The time of the day or the day of the week the test is administered can also affect the results: students may respond poorly on the test, not because of lack of ability but because of fatigue if the testing session is at the end of the school day or the last day of the school week. This point is consistent with what Viswanathan (2005) has to say about the idiosyncratic sources of error. Test bias occasioned by how students are grouped in their school or class is another source of error.

3.4 Reporting Standard Errors of Statistics

Researchers engaged in ILSA use both sampling and imputation variance components to calculate the standard error or uncertainty of an estimate of interest. In TIMSS and PIRLS, only the first plausible value is used to calculate the sampling variance component in order to simplify computations and reduce the time spent on them. The imputation variance is computed using all five plausible values (Foy et al., 2008; Kennedy & Throng, 2007).

When calculating mean achievement for two schools located in the same area, with the same number of sampled students, etc., a researcher might find that the tested students in School A score 75 points higher than the tested students in School B. But does this difference automatically mean that the tested students in School A performed much better than the students in School B? To answer this question, we need to take into account the precision of the estimates in each of the schools. The standard error of an estimate shows us how precise the measurement of the student abilities in each school is.

We can use the standard error to calculate a confidence interval for the tested students' results in individual schools. Such a calculation might show us that while the measured ability in School A is higher

than in School B, its dispersion is several times higher than in School B. As a consequence, the uncertainty associated with this higher estimate in School A becomes much greater and the results between the two schools are not statistically different. This is the reason why we must always take into account the standard error for the tested students in a given school when interpreting the results and making any comparisons. Both components of the standard error (sampling and imputation) must therefore be taken into account when reporting the results of the tested students back to a school.

When calculating the sampling variance within a school, we can use a bootstrapping approach. The basic idea of bootstrapping is that the sample collected is the best guess about the shape of the distribution of the population from which the sample was taken. Therefore, instead of assuming a theoretical shape for the population, we use the sample and its shape to estimate the variance of the statistic. Under the bootstrap technique, we take our original dataset of students from within each school and we make multiple new samples (called bootstrap samples) that are also of size *N*.

We then take these new samples from the original by using sampling with replacement. We can create many of these bootstrap samples (at least 1,000), and for each of them we can compute the statistic of interest. Each of these estimates is called a bootstrap estimate. We now have a distribution of the statistic of interest, and the variance of this distribution is the sampling variance of the statistic. This procedure provides an estimate of the shape of the distribution of the statistic, and that shape allows us to answer questions about how much the statistic varies.

The formula for the bootstrap sampling variance is the following:

$$V_s = \frac{\sum_{b=1}^{B} (\varepsilon_b - \bar{\varepsilon}_b)^2}{B-1}$$

where

B is the number of bootstrap samples,

$\bar{\varepsilon}_b$ is the average of the statistic across all the bootstrap samples, and

ε_b is the statistic computed from each of the bootstrap samples.

To calculate the measurement variance, we use the procedure that is recommended in the ILSA technical reports (see Foy et al., 2008, for an example). This procedure basically involves computing the statistic of interest with each of the plausible values, and then computing the measurement variance as the variance of these statistics multiplied by an expansion factor. The formula used is the following:

$$V_m = \left(1+\frac{1}{P}\right)\frac{\sum_{p=1}^{P}(\varepsilon_p - \bar{\varepsilon}_p)^2}{P-1}$$

where

P is the number of plausible values used in the analysis,

ε_p is the statistic of interest calculated via each of the plausible values, and

$\bar{\varepsilon}_p$ is the average of the statistic calculated P times, each with one of the plausible values.

The uncertainty of an estimate ε is then given by combining these two factors as:

$$SE_\varepsilon = \sqrt{V_s + V_m}\,.$$

4 Summary of Issues and Recommendations

4.1 Issues

The methodological and confidentiality issues considered in the previous sections brought to the fore several main points, a summary of which follows.

- Given the current class-sampling strategies in TIMSS and PIRLS, feedback can only be provided at the class level; no inferences to the school should therefore be made, except in cases where students from all classes in the school have been selected. As noted in Section 3.1, schools differ in the number of classes they each have, and selecting one or two of them cannot be deemed representative of the characteristics of the target-population students in the entire schools, especially the larger schools. This explains why we can talk only about the "performance of the tested class(es) in school X." The only exception to this rule is when all classes, hence students, are sampled and tested, as occurs with small schools.
- Under the matrix sampling of items, as used in TIMSS, PIRLS, and any other ILSA, the raw scores of the tested classes should not be used for reporting due to the different rotation of the test booklets. The nature of any one rotation depends on the number of students and possibly the different degrees of difficulty of the items included in the separate booklets.
- The plausible values are better estimates of student proficiency than are the percent correct and IRT scores other than the PVs (i.e., EAP, EAP-MG, MLE, WLE), and therefore should be used when giving schools feedback on the proficiency of their tested students. PVs also provide better representation of the underlying latent variable (student ability) compared to the point estimates provided by scores solely based on IRT. Plausible values also give a more precise estimate of the variance and the standard errors (Wu, 2005), provide unbiased estimates of differences between groups

(von Davier et al., 2009), which the tested students in schools are, and give the optimal statistics at group level. As von Davier et al. (2009, p. 35) note, "… from the point of view of groups, they [PVs] add exactly the right amount of variability to make the distribution of the PVs in the group match the distribution of the true values in the group."

- If the groups of tested students per school are too small, the measurement uncertainty is likely to increase substantially, as shown by the examples given in this publication and pointed out in the literature (Rutkowski et al., 2010; von Davier et al., 2009; Wu, 2010). As the example in Section 3.2 shows, the measurement error (i.e., uncertainty) for schools with 15 or fewer tested students becomes relatively high.

4.2 Recommendations

The above issues raise the most important point of this publication, namely that feedback on the performance of students selected with the current within-school sampling design should not be used for reporting student performance within a school. However, some alternative approaches can be recommended. In particular, if we are to provide reliable estimates at the school level, we need to modify the within-school sampling strategy.

One possible approach would be to sample all classes in the school. This, however, would unnecessarily increase the burden associated with implementing the study, as it would mean assessing more students than necessary, particularly within large schools, and overall across a country. Aside from the added operational burden on schools and test administrators, there is added cost in terms of increased printing and transportation of books, scoring demands, data entry requirements, and so on.

Another approach would be to sample students from across all classes in the school. However, considering that TIMSS and PIRLS ask the teacher of each class of selected students to answer the background questionnaire, the work involved in identifying and matching students with their teachers would become particularly complex and time consuming. To avoid this situation, a mixed approach might

be more desirable and easier to implement. Here, one or two intact classes would be selected from within each school, as is the current practice, but in addition a random sample of students would be selected from the remaining classes in the school.

Table 5 illustrates the different approaches, using a hypothetical case involving 150 selected schools, each of which has three classes, with 30 students in each class.

- Under the current sampling design, we would select one class from each school, resulting in an overall sample size for the country of 4,500 students. However, this approach has the limitation of not yielding representative samples within schools, as already described.
- Option 1 illustrates what would happen to the overall sample size if we then selected all classes within the schools. While we would have information about intact classes, we would still need to test a total of 13,500 students.
- Under Option 2, we would select 30 students at random from within each school, resulting in a more reasonable sample size (4,500). However, this approach would see us losing the possibility of studying intact classes, and we would increase the operational burden of tracking teachers from across all the classes.
- Under Option 3, perhaps the optimal alternative, we would need, in order to select a sample representing the students' characteristics within the school, to select an intact class, as well as students from within the rest of the classes. The intact class would be Class 2 in our example, and we would need to select a fixed number of students (10 in our example) at random from each of the remaining classes.

Table 5: Different options for sampling within schools

	Class 1 (n = 30)	Class 2 (n = 30)	Class 3 (n = 30)	School sample size	Overall sample size
Current	30			30	4,500
Option 1	30	30	30	90	13,500
Option 2	10	10	10	30	4,500
Option 3	10	30	10	50	7,500

These options, however, are not all cost-effective in terms of printing, shipping the testing materials, scoring, data entry, and data cleaning.

We also need to be mindful that sampling a proportion of students from the other available classes has to follow certain rules to ensure the desired precision is achieved. The measure of precision, under simple random sampling assumptions, would need to be the expectation of an error of the mean of a certain magnitude, given the standard deviation. Under assumptions of simple random sampling, we would expect the error of a mean to be the ratio of the standard deviation divided by the square root of the sample size, multiplied by the finite population correction. The formula looks like this:

$$SE_{\bar{x}} = \frac{SD_x}{\sqrt{n}} \times \sqrt{\frac{(N-n)}{(N-1)}}.$$

Here, *SDx* is the standard deviation within the school, *N* is the school size, and *n* is the sample size.

Table 6 presents the sample sizes we would require if we were selecting students at random from across all the classes in a school (Option 2 in Table 5) and when the desired level of precision is presented as the ratio of the standard error of the mean divided by the standard deviation. The closer that this ratio comes to zero, the more precise the results will be. Thus, a ratio of 0.05 means higher precision than a ratio of 0.10.

Note also, in particular, in regard to Table 6, the following:

- As the level of precision required becomes higher, the sample size needed becomes larger;
- Doubling the sample size more than doubles the precision;
- The relationship between sample size and precision is not linear; and
- A school double the size of another does not require double the sample size.

Table 6: Recommended within-school sample sizes using different *SE* to *SD* ratios

School size	Ratio of standard error of the mean to the standard deviation				
	0.05	0.10	0.15	0.20	0.25
20	20	18	16	13	11
25	25	22	18	14	12
30	29	25	20	15	12
35	34	28	21	16	13
40	38	30	23	17	13
45	42	33	24	18	14
50	46	35	25	18	14
55	50	37	26	19	14
60	54	39	27	19	14
65	58	41	28	20	15
70	61	43	29	20	15
75	65	45	30	20	15
80	68	46	30	21	15
85	72	48	31	21	15
90	75	49	31	21	15
95	78	50	32	21	15
100	82	52	32	22	15
105	85	53	33	22	16
110	88	54	33	22	16
115	91	55	34	22	16
120	94	56	34	22	16
125	97	57	34	22	16
130	100	58	35	23	16
135	103	59	35	23	16
140	105	60	35	23	16
145	108	61	36	23	16
150	111	62	36	23	16
155	113	63	36	23	16
160	116	63	36	23	16
165	119	64	37	23	16
170	121	65	37	23	16
175	123	65	37	23	16
180	126	66	37	24	16
185	128	67	37	24	16
190	131	67	38	24	16
195	133	68	38	24	16
200	135	68	38	24	16

Table 6 essentially serves as a guideline for sample-size requirements. For example, notice that for schools with fewer than 40 students, we would need to select most (30) students in the school when the desired precision is 0.10 of the standard deviation. Therefore, it might be practically more feasible and operationally less complex for us to simply take all students in the school when there are fewer than 40 of them, and to select as many as the table recommends from schools with more than 40 students. We would also need to calculate sampling weights accordingly within each school to take into account the selection of students, and to make the corresponding adjustment for nonparticipation, if this is deemed necessary.

Table 6 also provides us with a guideline for sample selection should we wish to select one or more intact classes from within each school (Option 3 in Table 5). When selecting an intact class within a school, we would need to select the remaining students from the remaining classes, but only after taking out the contribution of the students from the selected class. For example, if we were looking for a precision of 0.10 of the standard deviation in a school that has 200 students, the recommendation in Table 6 would be for us to select 68 students from that school. If the school had 10 classes, of approximately 20 students each, we would then select one intact class from that school, and would then select 61 students from the remaining nine classes.

Notice here that even though the selected class is contributing approximately 20 students, we would have to treat these 20 students as contributing only 7 students to the overall sample (calculated as 68 students/10 classes in the school). This is because these 20 students were selected as an intact class. We would need to select the remaining 61 students from across the not-selected classes in the school. If we selected two intact classes, these would contribute 14 students (6.8 each), and the remaining 54 students would come from the remaining classes.

A few other points have salience here. The first is that because we are selecting one or more intact classes, the procedure used will yield a somewhat more precise estimate within the school than if we simply selected students at random across the classes. Second, the procedure assumes that each class has approximately the same number of

students. Third, the selection of students across the remaining classes should preferably be done using a systematic random sampling procedure whereby all students from the remaining classes are selected from a list sorted by class and other relevant implicit stratification variables. Fourth, unless we can expect a 100 percent participation rate within each school, we would need to select a larger sample to account for non-participation. In this instance, a 5 to 10 percent nonparticipation could be built into the selection.

In addition to taking into account the points mentioned above, we need to pay attention to how best to collect the teacher data, especially in terms of the consequences for international analysis and reporting of that data. Two options that we could usefully explore are these:

1. Use the intact classes for international reporting, but the across-school samples for reporting results to the schools. In this case, we would need to calculate two sets of weights and to allocate the contribution of each of the sampled students accordingly.
2. Use all the student data from each of the schools. In this case, we would need only one set of weights, but we would then have the operational burden of identifying and administering the teacher background questionnaires to the teachers of the selected students in the school. The burden of doing this would be directly related to the number of classes in the school.

Some alternatives that could simplify the implementation of the study in the field might be establishing shortcut alternatives. One, mentioned earlier, involves selecting all the students in schools in cases when there are fewer than 40 students. Another is to select all students in schools with one or two classes, but then to select one intact class in schools with three or more classes, and the rest of the students from the remaining classes, using the procedure described above.

Although all of the above approaches might facilitate administration of the assessment in the field, they will also result in somewhat larger sample sizes and, from there, additional costs in terms of preparing and handling the assessment materials (printing, distribution, scoring, data entry, etc.), as well as the additional burden associated

with processing and cleaning the data. But these approaches would also yield data that can be used to report results at the school level, thus increasing participation rates from within each school, as well as the usefulness of the information collected.

5 Reporting Results to Schools

5.1 Reporting on Student and School Backgrounds and Performance

An important issue related to school reporting is how to report the results in terms of student characteristics. Here, the issue of small groups again needs to be taken into account. While choosing the variables to report is up to the national staff conducting the study, care needs to be exercised when choosing those variables. Reported variables can include, among many others, the students' socioeconomic status (SES) aggregated at school level, availability of school resources, school composition, the type of school funding, school location, and teacher qualifications.

Some information on the teaching practices (e.g., time devoted to different topics, use of demonstrations, use of computers, homework assignments), as reported in the questionnaires completed by the teachers teaching the sampled classes, could also be reported. However, we could not consider this information as valid for all teachers of the subject in the target grade in the school because only teachers teaching the sampled classes would have been selected. Choosing these variables must take into account relevant theory and research and align with the contextual features of the respective countries' education systems. Thus, the variables considered must be ones whose relationship with the educational outcomes is proven.

One of the most important factors seemingly associated with school performance is student intake. So far, studies have shown that student intake has a "compositional effect" on the outcomes of education. This term refers to different aspects of the students' social background, such as immigration status, SES, and ethnic origin, as well as school and class characteristics that include, amongst others, school and class learning resources, the SES of the school, and the percentage of nonnative speakers in the school (Bellin, Dunge, & Gunzenhauser, 2010).

As early as 1966, the Coleman Report (Coleman et al., 1966) from the United States drew attention to the influence of the compositional effect on the outcomes of education, and research over the decades since has both supported and strengthened our knowledge and understanding of that effect. As such, it cannot be ignored when reporting school and student performance. To give an example of a more recent study focusing on the compositional effect, Dumay and Dupriez (2008) found that in Belgium (French community), this effect influenced students' reading achievement even after the researchers had controlled for students' initial performance and sociocultural backgrounds.

As Harker and Tymms (2004) point out, students in some schools come predominantly from families with relatively limited resources, and these families might also have low motivation with respect to achievement. Some schools have student intakes where children have well-developed learning abilities at the age of school entrance. There may also be schools where student intake is affected by segregation based on gender, ethnicity, or religion. Then there are schools where the student intake is relatively heterogeneous in terms of learning skills and background characteristics. Harker and Tymms (2004) also draw attention to how student composition in association with school resources and peer influence can affect the overall performance of any one school. The work of these two researchers again makes evident the importance of setting reported results within the context of school-level background characteristics.

Besides detailing variables on student achievement and background characteristics, feedback to schools should provide particularly valuable information for school principals and teachers. The variables include student motivation, student attitudes toward the subjects tested, and students' views on the importance that the studied subjects hold for their (the students') future. The results for these variables can be presented as simple univariate statistics.

5.2 Information and Level of Detail in the Feedback

The scaling of overall Grade 4 mathematics achievement in TIMSS 2007 used 177 items. Comparable scaling for overall science used 170. For Grade 8, these numbers were 214 and 210, respectively. These two

achievement scales for each grade incorporated information drawn from large numbers of items. However, the separate content domain subscales in mathematics (Grade 4: number, geometric shapes and measure, and data display; Grade 8: number, algebra, geometry, data and chance, and guidelines for calculator use) and science (Grade 4: life science, physical science, and earth science; Grade 8: biology, chemistry, physics, and earth science) were produced using much smaller numbers of items—between 26 and 71 for Grade 4 and between 40 and 75 for Grade 8. The three cognitive domain subscales (knowing, applying, and reasoning), which are the same across grades and subjects, are also produced by a small number of items. As Wu (2010) emphasizes, the shorter a scale is (i.e., the lower the number of items), the greater is the degree of uncertainty associated with it.

The fact that the number of tested students per school in TIMSS 2007 was also not big (one or two classes) makes for another serious source of measurement error. As such, reporting achievement results should ideally focus only on the overall mathematics and science scales, and not on the subscales. Due to the small number of both sampled students and achievement items per subscale, reporting achievement focused on subscale data should be avoided no matter what the sampling design of the study is. The only exception to this rule is census sampling (i.e., when all target students within a country are selected).

One of the most important features of the report should be that its intended readers can easily comprehend the information it includes. As Van Petegem and Vanhoof (2004) maintain, we cannot expect or guarantee that teachers and school principals possess deep knowledge and understanding of the statistical concepts and complexities. The report must therefore also include some means of helping these individuals read, understand, and interpret the results. Even so, the information should still be presented in a way that people unversed in research methodology and statistics can easily understand. To avoid confusion and misinterpretation, those preparing the reports need to make sure that not only the language used but also the tabular and graphical representation of the data are kept as simple as possible.

The information in the following panel provides advice on the components of reports to schools and what to include in each section.

- *Introduction:* This should contain a short description of the report and its objectives and a note on confidentiality. It should also contain general information about the study, the organization conducting it, the study's objectives, and the number of participating countries in each population (e.g., Grades 4 and 8 in TIMSS).
- *A general overview of the country sample and its characteristics:* The information included here should cover the number of participating schools and tested students, the average number of students per school, and the percentages of boys and girls per school and overall.
- *An outline of the structure of the report:* This should include brief explanations of the content of each section of the report. This is also the place where the members of the national study team can express their gratitude to the school for taking part in the study.
- *Information on the sample particular to the school receiving feedback*.
- *Descriptions of the background variables* used to classify the school and to compare its students' achievement with the achievement of students from the group of similar schools.
- *Information on the average mathematics and science achievement of the school's students*, with comparisons of these students' performance with the performance of students from the group of schools with similar background characteristics and spread of results (heterogeneity).
- *Concluding remarks*.

5.3 Uses of the Feedback

As discussed in Section 3 of this publication, sampling and measurement issues affect the precision of the estimates, and the school composition can affect the performance of the tested students in the participating schools (see also Section 5.1). These design-related issues of ILSA limit the extent and type of use that schools can make of the feedback sent to them. In terms of school characteristics, the limitations stem not only from factors associated with the student intake, such as SES and learning motivation, but also from the volatility of the year-to-year test results of a school or class.

As Wu (2010) points out, although teachers and the instruction they provide are important with respect to education outcomes, teachers have limited influence in terms of the variance in achievement accounted for by student characteristics. The variation in student proficiency within schools is higher than it is between schools. Even if teachers do not change their instruction, the outcomes from ILSA from one student cohort to another will vary due to the influence of random factors. As Wu (2009, p. 19) cautions, "Any suggestion that teacher or school performance should be determined by student test results is of serious concern." In short, the results from ILSA should not be used to compare individual schools, rank them, and make judgments about the job a school or a teacher does; nor should they be used as a means of rewarding or penalizing teachers and schools.

One final important point is the necessity of keeping the reports for individual schools confidential. The information contained in each report should reflect only the performance of the tested students within the school, should keep the results of other individual schools anonymous to avoid comparisons, and should not be made public (Van Petegem & Vanhoof, 2004). In similar vein, the feedback should provide information only on the results of the tested students, and it should not contain judgmental inferences about the quality of the education in the school or make recommendations for improvement. To do otherwise risks, as Bos and Schwippert (2003) put it, feedback easily turning from "use to abuse."

References

Australian Council for Educational Research (ACER). (n. d.). *International Benchmark Tests* (IBT). Retrieved from http://www.acer.edu.au/tests/ibt/overview-ibt

Bellin, N., Dunge, O., & Gunzenhauser, C. (2010). The importance of class composition for reading achievement: Migration background, social composition, and instructional practices. An analysis of the German 2006 PIRLS data. *IERI Monograph Series*, *3*, 9–34.

Bos, W., & Schwippert, K. (2003). The use and abuse of international comparative research on student achievement. *European Educational Research Journal*, *2*(4), 559–573.

Buckingham, J. (2008). Making the grade: School report cards and league tables. *Centre for Independent Studies*, *103*, 1–8.

Carstens, R., & Hastedt, D. (2010). The effect of not using plausible values when they should be: An illustration using TIMSS 2007 Grade 8 mathematics data. In *Proceedings of the 4th IEA International Research Conference*. Retrieved from http://www.iea-irc.org/fileadmin/ IRC_2010_papers/TIMSS/Carstens_Hastedt.zip

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

Cronbach, L. J., Linn, R. L., Brennan, R. L., & Haertel, E. (1995). *Generalizability analysis for educational assessments*. Los Angeles, CA: CRESST, University of California at Los Angeles.

Dumay, X., & Dupriez, V. (2008). Does the school composition effect matter? Evidence from Belgian data. *British Journal of Educational Studies*, *56*, 440–477.

Foshay, A. W., Thorndike, R. L., Hotyat, P., Pidgeon, D. A., & Walker, D. A. (1962). *Educational achievement of thirteen-year-olds in twelve countries: Results of an international research project, 1959–1961*. Hamburg, Germany: UNESCO Institute for Education.

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

Foy, P., Galia, J., & Li, I. (2008). Scaling the data from the TIMSS 2007 mathematics and science assessments. In J. F. Olson, M. O. Martin, & I. V. Mullis (Eds.), *TIMSS 2007 technical report* (pp. 225–280). Chestnut Hill, MA: Boston College.

Harker, R., & Tymms, P. (2004). The effects of student composition on school outcomes. *School Effectiveness and School Improvement*, *15*, 177–199.

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, *77*(1), 81–112.

Husén, T. (Ed.). (1967). *International study of achievement in mathematics: A comparison of twelve countries* (Vols. 1–2). Stockholm, Sweden: Almqvist & Wiksell.

International Association for the Evaluation of Educational Achievement (IEA). (2007). *Trends in International Mathematics and Science Study* [Data files and codebook]. Retrieved from http://rms.iea-dpc.org/

Joncas, M. (2007). PIRLS 2006 sampling design. In M. O. Martin, I. V. Mullis, & A. M. Kennedy (Eds.), *PIRLS 2006 technical report* (pp. 35–48). Chestnut Hill, MA: Boston College.

Joncas, M. (2008). TIMSS 2007 sample design. In J. F. Olson, M. O. Martin, & I. V. Mullis (Eds.), *TIMSS 2007 technical report* (pp. 77–92). Chestnut Hill, MA: Boston College.

Kennedy, A. M., & Throng, K. L. (2007). Reporting student achievement in reading. In M. O. Martin, I. V. Mullis, & A. M. Kennedy (Eds.), *PIRLS 2006 technical report* (pp. 195–222). Chestnut Hill, MA: Boston College.

Martin, M. O. (1996). Third International Mathematics and Science Study: An overview. In *Third International Mathematics and Science Study technical report: Vol. I. Design and development* (pp. 1-1–1-20). Chestnut Hill, MA: Boston College.

Martin, M. O., Mullis, I. V., Foy, P., & Stanco, G. M. (2012). *PIRLS 2011 international results in science*. Chestnut Hill, MA: Boston College.

Mullis, I. V., Kennedy, A. M., Martin, M. O., & Sainsbury, M. (2006). *PIRLS 2006 assessment framework and specifications* (2nd ed.). Chestnut Hill, MA: Boston College.

Mullis, I. V., & Martin, M. O. (2008). Overview of TIMSS 2007. In J. F. Olson, M. O. Martin, & I. V. Mullis (Eds.), *TIMSS 2007 technical report* (pp. 1–12). Chestnut Hill, MA: Boston College.

Mullis, I. V., Martin, M. O., Foy, P., & Arora, A. (2012). *PIRLS 2011 international results in mathematics*. Chestnut Hill, MA: Boston College.

Mullis, I. V., Martin, M. O., Foy, P., & Drucker, K. T. (2012). *PIRLS 2011 international results in reading*. Chestnut Hill, MA: Boston College.

Mullis, I. V., Martin, M. O., Ruddock, G. J., O'Sullivan, C. Y., Arora, A., & Erberber, E. (2005). *TIMSS 2007 assessment framework*. Chestnut Hill, MA: Boston College.

PISA-Konsortium Deutschland, & Leibniz-Institut für Pädagogik der Naturwissenschaften. (n. d.). *PISA 2006 Schülerleistungen im internationalen Vergleich: Realschule im Schulzentrum Sandesneben Schulrückmeldung* [PISA 2006 student performance in international comparison: Feedback to secondary schools]. Kiel, Germany: Author. Retrieved from http://www.gemeinschaftsschule-sandesneben.de/PDF/Realschule%20im%20Schulzentrum%20Sandesneben.pdf

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.

Schägen, I., Hutchinson, D., & Hammond, P. (2006). League tables and health checks: The use of statistical data for school accountability and self-evaluation. In P. Dobbelstein & T. Neidhardt (Eds.), *Schools for quality: What data-based approaches can contribute* (pp. 57–75). Brussels, Belgium: CIDREE/DVO.

Schoolfeedbackproject [website]. (n. d.). Retrieved from http://www.schoolfeedback.be/index.php?/het_project/

Van Petegem, P., & Vanhoof, J. (2004). Feedback on indicators to schools. *European Educational Research Journal*, *3*(1), 246–277.

Van Petegem, P., & Vanhoof, J. (2005). Feedback on performance indicators: A tool for school improvement? Flemish case studies as a starting point for constructing a model for school feedback. *REICE—Revista Electrónica Iberoamericana sobre Calidad, Eficacia y Cambio en Educación*, *3*(1), 222–234.

Visible Learning Laboratories. (n. d.). *asTTle project*. Auckland, New Zealand: Author. Retrieved from http://www.visiblelearning.biz/pageloader.aspx?page=534d96d0d0

Viswanathan, M. (2005). *Measurement error and research design*. Thousand Oaks, CA: Sage Publications.

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

von Davier, M., Gonzalez, E., & Mislevy, R. (2009). What are plausible values and why are they useful? *IERI Monograph Series: Issues and Methodologies in Large-Scale Assessments*, *2*, 9–36.

Wu, M. (2005). The role of plausible values in large-scale surveys. *Studies in Educational Evaluation*, *31*, 114–128.

Wu, M. (2010). Measurement, sampling, and equating errors in large-scale assessments. *Educational Measurement: Issues and Practice*, *29*(4), 15–27.