Education Policy Analysis Archives

Volume 8 Number 43 **August 21, 2000** **ISSN 1068-2341**

A peer-reviewed scholarly electronic journal
Editor: Gene V Glass, College of Education
Arizona State University

Articles appearing in **EPAA** are abstracted in the *Current Index to Journals in Education* by the ERIC Clearinghouse on Assessment and Evaluation and are permanently archived in *Resources in Education*.

Consistency of Findings Across International Surveys of Mathematics and Science Achievement: A Comparison of IAEP2 and TIMSS

Michael O'Leary
Thomas Kellaghan
St Patrick's College, Dublin

George F. Madaus
Albert E. Beaton
Boston College

Abstract
The investigation reported in here was prompted by discrepancies between the performance of Irish students on two international tests of science achievement: the Second International Assessment of Educational Progress (IAEP2) administered in 1991 and the Third International Mathematics and Science Study (TIMSS) administered in 1995. While average science achievement for Irish 13-year-olds was reported to be at the low end of the distribution representing the 20 participating countries in IAEP2, it was around the middle of the distribution representing the 40 or so countries that participated in

TIMSS at grades 7 and 8. An examination of the effect sizes associated with mean differences in performance on IAEP2 and TIMSS indicated that the largest differences are associated with the performance of students in France, Ireland and Switzerland. Five hypotheses are proposed to account for the differences.

Introduction

International comparative studies of student achievement have become part of the educational landscape over the past four decades. In these studies, a number of countries (usually represented by research organizations) agree on an instrument to assess achievement in a curriculum area, the instrument is administered to a representative sample of students at a particular age or grade level in each country, and comparative analyses of the data obtained are carried out. The most frequently assessed areas have been reading, mathematics, and science at ages 9 or 10 and 13 or 14. The number of participating countries has grown from 12 in a pilot project conducted between 1959 and 1961 to over 40 for a survey of mathematics and science achievements in 1995 (see Goldstein, 1996; Husén & Postlethwaite, 1996; Kellaghan, 1996).

The potential of international studies to contribute to policy formation was made clear from the earliest studies (Husén, 1967; Lambin, 1995). Over the years, a range of purposes to which information derived from such studies might be put has been suggested. These include the pursuit of equity goals, setting priorities, assessing the effectiveness and efficiency of the educational enterprise and the appropriateness of curricula, evaluating instructional methods and the organization of the school systems, and providing a mechanism for accountability (Kellaghan & Grisay, 1995; Plomp, 1992). While we have relatively little information on the extent to which the findings of international studies have in fact been utilized, there is no doubt that they attract considerable media and public attention.

A variety of factors can affect the extent to which data obtained in an international study accurately reflects what students have learned in the participating countries, something that is necessary if valid comparisons between countries are to be made (see Brown, 1996, 1998; Goldstein, 1996; Kellaghan, 1996; Kellaghan & Grisay, 1995; Murphy, 1996; Nuttall, 1994). One relates to the adequacy of a uniformly administered assessment procedure to measure the outcomes of a variety of curricula. Since curricula differ from country to country, an assessment instrument will not reflect the curricula of all countries participating in an international study to the same degree.

The second factor relates to the extent that the populations and samples of pupils for whom data are obtained can be regarded as equivalent. Defined target populations may not be comparable across countries since exclusion practices may differ (e.g., relating to students with handicapping conditions/learning problems or when the language of the assessment instrument differs from the language of the school). Differences in participation rates of selected samples (due to lack of co-operation from schools, student absenteeism) will make matters worse.

Many commentators have considered how these problems impact on comparisons based on a single study. Additional problems arise when the findings of two different surveys are being compared. In the case of IAEP2 and TIMSS, instruments used to measure achievement differed in form and content sampled, age-based versus grade-based populations definitions were used, and different methods of data

manipulation were utilized.

The investigation reported here was prompted by discrepancies between the performance of Irish students on tests of science in the Second International Assessment of Educational Progress in Mathematics and Science (IAEP2) (Lapointe, Askew & Mead, 1992) in 1991 and, four years later in the Third International Mathematics and Science Study (TIMSS) (Beaton, Mullis, Martin, Gonzalez, Kelly & Smith, 1996a; Beaton, Martin, Mullis, Gonzalez, Smith, & Kelly, 1996b). Initially, the intention was to focus on the Irish problem but, as the investigation proceeded, it became clear that discrepancies in performance between the two surveys were not confined to Irish students.

In this article, we first present brief descriptions of IAEP2 and TIMSS. We then select 12 countries that participated in both surveys for further analyses: Canada, England, France, Hungary, Ireland, Korea, Portugal, Scotland, Slovenia, Spain, Switzerland, and the United States. Our approach to assessing the consistency of countries' performances is based on an examination of the performance of each country relative to the performance of other countries in both surveys. If results are stable, differences in performance between countries should not vary very much from one survey to the next. To the extent that they do, findings may be regarded as unstable. Change in effect sizes between pairs of means on the two assessments were calculated to obtain an estimate of the magnitude of differences between performance on the two occasions.

IAEP2 and TIMMS

In IAEP2, representative samples of 9 and 13-year-olds in 20 countries were assessed in mathematics and science in 1991 (Lapointe, Askew & Mead, 1992). In TIMSS, the mathematics and science achievements of students in grades 3, 4, 7, 8, and in the final grade of secondary education were assessed in 1995 (Beaton et al., 1996). Data are reported in our article for 13- year-olds in IAEP2 and for grades 7 and 8 students in TIMSS. However, the main focus is on grade 7 performance, since in all countries that had participated in both assessments, except Scotland, more 13 year-olds were in grade 7 than in grade 8 (Beaton et al, 1996a, p. A12).

The IAEP2 tests for 13-year olds were contained in two separate booklets, each of which had to be completed by students in four 15-minute segments (one hour testing time in all). The mathematics booklet contained 76 items and covered four content areas: Measurement, Geometry, Data Analysis/Statistics/Probability, and Algebra/Functions. The science test consisted of 72 items and covered four content areas: Life Sciences, Physical Sciences, Earth/Space Sciences, and the Nature of Science. Students completed either a mathematics or science test and were administered all items on the test.

Unlike IAEP2, the TIMSS test booklets contained both mathematics and science items. At grades 7 and 8, the mathematics test comprised 151 items and the science test 135 items. The TIMSS mathematics items covered six content areas: Fractions/Number Sense, Geometry, Algebra, Data Representations/Analysis/Probability, Measurement, and Proportionality. The science content areas were: Earth Science, Life Science, Physics, Chemistry, and Environmental Issues/Nature of Science. Items were rotated across eight test booklets and student performance was matrix-sampled using a modified Balanced-Incomplete-Block (BIB) spiraling design (Martin & Kelly, 1997). One and a half hours were allocated for the completion of each booklet. In both studies, performance on both tests was reported in the form of an average percentage correct score. In the case of TIMSS, an average scale score for each country was also reported.

While scale scores were calculated for the IAEP2 study, they were not included in the published reports.

The Consistency of IAEP2 and TIMSS Science Results

In 1991, the average science performance of Irish 13-year-olds is significantly below the average performance of students in all but two of the 'common' countries (Portugal and the US) and also significantly below the international mean (Lapointe, Askew, & Mead, 1992). However, in 1995, the average performance of Irish students on the TIMSS test at grades 7 & 8 compares much more favorably with the 'common' countries and with the overall TIMSS means (Beaton et al, 1996b). This change of fortune is clearly evident in Table 1, in which countries are listed from highest achieving to lowest achieving, and are categorized according to whether their means were statistically significantly above, below, or did not differ from, the Irish mean.

Table 1
Science and Mathematics Means of Countries that Participated in IAEP2 and TIMSS
(Categorised in Terms of the Significance of Difference of Each Mean from the Irish Mean)[a, b]

IAEP2 13-year-olds			**TIMSS Grade 7**			**TIMSS Grade 8**		
Science								
	M	SE		M	SE		M	SE
Overall[c]	66.9			49.8	(0.1)		55.5	(0.1)
Kor	77.5	(0.5)	Kor	61.4	(0.4)	Kor	65.5	(0.3)
Swi	73.7	(0.9)	Slo	57.2	(0.5)	Slo	61.7	(0.5)
Hun	73.4	(0.5)	Hun	55.5	(0.6)	**Eng**	**61.3**	(0.6)
Slo	70.3	(0.5)	Eng	55.6	(0.6)	**Hun**	**60.7**	(0.6)
Can	68.8	(0.4)	**US**	**54.0**	(1.1)	**Can**	**58.7**	(0.5)
Eng	68.7	(1.2)	**Can**	**54.0**	(0.5)	**Ire**	**58.4**	(0.9)
Fra	68.6	(0.6)	**Ire**	**52.0**	(0.7)	**US**	**58.3**	(1.0)
Sco	67.9	(0.6)	**Swi**	**50.1**	(0.4)	**Swi**	**56.3**	(0.5)
Spa	67.5	(0.6)	Spa	49.3	(0.4)	Spa	55.6	(0.4)
US	**67.0**	(1.0)	Sco	48.2	(0.8)	**Sco**	**55.3**	(1.0)
Ire	**63.3**	(0.6)	Fra	46.1	(0.6)	Fra	53.7	(0.6)
Por	**62.6**	(0.8)	Por	41.3	(0.5)	Por	49.9	(0.6)
IAEP2 13-year-olds			**TIMSS Grade 7**			**TIMSS Grade 8**		
Mathematics								
	M	SE		M	SE		M	SE
Overall	58.3			49.3	(0.1)		55.1	(0.1)

Kor	73.4	(0.6)	Kor	67.0	(0.6)	Kor	71.7	(0.5)
Swi	70.8	(1.3)	**Hun**	**53.8**	(0.8)	**Swi**	**62.0**	(0.6)
Hun	68.4	(0.8)	**Swi**	**53.1**	(0.5)	**Hun**	**61.5**	(0.7)
Fra	**64.2**	(0.8)	**Ire**	**53.3**	(1.0)	**Fra**	**61.3**	(0.8)
Can	**62.0**	(0.6)	**Slo**	**52.5**	(0.7)	**Slo**	**61.2**	(0.7)
Eng	**60.6**	(2.2)	**Can**	**51.6**	(0.5)	**Ire**	**58.7**	(1.2)
Sco	**60.6**	(0.9)	**Fra**	**51.0**	(0.8)	**Can**	**58.7**	(0.5)
Ire	**60.5**	(0.9)	US	47.7	(1.2)	Eng	53.1	(0.7)
Slo	**57.1**	(0.8)	Eng	47.2	(0.9)	US	53.0	(1.1)
Spa	55.4	(0.8)	Sco	44.3	(0.9)	Sco	51.6	(1.3)
US	55.3	(1.0)	Spa	42.4	(0.6)	Spa	51.0	(0.5)
Por	48.3	(0.8)	Por	36.6	(0.6)	Por	42.9	(0.7)

[a]In TIMSS, overall scale scores rather than overall average percents correct were used to report the outcomes of statistical tests.

[b] Average performance in countries whose data appear in bolded type is not statistically significantly different from that in Ireland. Average performance in countries above the bolded entires is statistically significantly above that in Ireland. Average performance in countries below the bolded entries is statistically significantly below that in Ireland.

[c] The international averages in the table are for all participating countries and educational systems in each of the studies. The standard errors for the IAEP averages were not published.
Source. For IAEP2: Lapointe, Askew, & Mead (1992), Lapointe, Mead, & Askew (1992), ETS, (1992). For TIMSS: Beaton et al. (1996a; b), Center for the Study of Testing, Evaluation and Public Policy (n.d.).

Compared to their performance on the IAEP2 science assessment, four countries maintain their superiority over Ireland on the TIMSS assessment at grade 7 (Korea, Slovenia, Hungary, England). Two, having performed at a superior level on IAEP2, achieve at levels comparable to Ireland in TIMSS (Canada, Switzerland), while three that were superior on IAEP2 record a significantly poorer performance on TIMSS (France, Scotland, Spain). Comparisons between IAEP2 performance and performance at grade 8 on TIMSS reveal a somewhat similar pattern in which only two countries (Korea and Slovenia) maintain their superior position.

It is apparent that the relative performances of countries other than Ireland also change between IAEP2 and TIMSS (e.g., France and Switzerland). It could be argued that the same phenomenon occurs in mathematics (compare, for example, English and Scottish performances in the two surveys). However, changes in position are less frequent in mathematics, a finding that is reflected in the magnitude of the correlations between scores in the two assessments (Table 2).

Table 2
Correlations Between the Performances of Countries that Participated in Both IAEP2 and TIMSS (n=12)

	TIMSS Grade 7 Mean Scale Score	**TIMSS Grade 7 Mean Percent Correct**

Mathematics		
IAEP2 Mean Scale Score	.83	
IAEP2 Mean Percent Correct		.83
Science		
IAEP2 Mean Scale Score	.55	
IAEP2 Mean Percent Correct		.66

In considering the consistency of scores from one assessment to another, data on statistical significance from the published reports could have been used (as they were in Table 1). However, since our interest is in the extent to which the size of differences between pairs of country means changed across the assessments, we chose to use an effect-size index.

Effect Size Differences

The effect size is a measure of the magnitude in numerical terms of a difference of interest (in the present case, mean differences between countries) (Hair, Anderson, & Black, 1995; Wolf, 1986). The measure chosen for the present analysis is Cohen's *d* which is a measure of standardized differences between means, expressed in terms of standard deviation units (Cohen, 1977). The measure provides a scale-invariant estimate of the magnitude of an effect and involves dividing the value of the difference between two group means by the pooled standard deviation, using the formula,

$$d = (M_1 - M_2)/s_{pooled}$$ in which,

d is the effect size index for differences between means in standard units;
M_1 and M_2 are the sample means in original measurement units; and
s_{pooled} is the pooled standard deviation for both samples and is calculated as
$$[(n_1 - 1)s_1 + (n_2 - 1)s_2]^{1/2} (n_1 + (n_2 - 2)^{-1/2}$$

The effect size measure is now in the common metric of standard deviation units. Thus, an effect size of 0.3 indicates that one country scored 0.3 of a standard deviation higher (or lower) than the comparison country. Guidance for interpreting effect sizes is equivocal. It has been suggested that effect sizes around 0.2 are small, those around 0.5 are medium, and those around or above 0.8 are large (Cohen, 1977). However, the significance of an effect size will depend on the context in which it is obtained (Durlak, 1995).

Table 3

Effect Sizes Observed in Science for IAEP2

	Can	Eng	Fra	Hun	Irl	Kor	Por	Sco	Slo	Spa	Swi	US
Can	.00	+.01	+.04	-.27	+.39	-.54	+.45	+.08	-.03	+.15	-.31	+.16
Eng	-.01	.00	+.03	-.27	+.34	-.53	+.41	+.06	-.04	+.12	-.28	+.15
Fra	-.04	-.03	.00	-.30	+.32	-.57	+.39	+.03	-.07	+.09	-.32	+.12
Hun	+.27	+.27	+.30	-.00	+.60	-.26	+.67	+.32	+.23	+.42	-.01	+.43
Ire	-.39	-.34	-.32	-.60	.00	-.89	+.07	+.28	-.39	-.26	-.65	-.21
Kor	+.54	+.53	+.57	+.26	+.89	.00	+.96	+.60	+.50	+.69	+.25	+.69
Por	-.45	-.41	-.39	-.67	-.07	-.96	.00	-.35	-.45	-.33	-.70	-.28
Sco	-.08	-.06	-.03	-.32	+.29	-.60	+.35	.00	-.10	+.06	-.36	+.09
Slo	+.03	+.04	+.07	-.23	+.39	-.50	+.45	+.10	.00	+.18	-.27	+.19
Spa	-.15	-.12	-.09	-.42	+.26	-.69	+.33	-.06	-.18	.00	-.46	+.03
Swi	+.31	+.28	+.32	+.01	+.66	-.25	+.70	+.36	+.27	+.46	.00	+.44
US	-.16	-.15	-.12	-.43	+.21	-.69	+.28	-.09	-.19	-.03	-.44	.00

Note: Reading across the row and comparing performance with country listed in heading: Positive effect sizes reflect higher average performance; negative effect sizes reflect lower average performance.

Table 4
Effect Sizes Observed in Science for TIMSS Lower Grade

	Can	Eng	Fra	Hun	Irl	Kor	Por	Sco	Slo	Spa	Swi	US
Can	.00	-.14	+.61	-.21	+.04	-.39	+.83	+.34	-.35	+.26	+.17	-.09
Eng	+.14	.00	+.72	-.06	+.17	-.24	+.89	+.44	-.18	+.39	+.28	+.04
Fra	-.61	-.72	.00	-.88	+.58	-1.01	+.31	-.23	-1.06	-.34	-.44	-.57
Hun	+.21	+.06	+.88	.00	+.25	-.19	+1.12	+.54	-.13	+.50	+.39	+.10
Ire	-.04	-.17	+.58	-.25	.00	-.44	+.86	+.29	-.39	+.22	+.13	-.12
Kor	+.39	+.24	+1.01	+.19	+.44	.00	+1.20	+.73	+.05	+.66	+.56	+.26
Por	-.83	-.89	-.31	-1.12	-.86	-1.20	.00	-.51	-1.39	-.63	-.75	-.77
Sco	-.34	-.44	+.23	-.54	-.29	-.73	+.51	.00	-.68	-.11	-.18	-.38
Slo	+.35	+.18	+1.06	+.13	+.39	-.05	+1.39	+.68	-.00	+.66	+.55	+.21
Spa	-.26	-.39	+.34	-.50	-.22	-.66	+.63	+.11	-.66	.00	-.09	-.30
Swi	-.17	-.28	+.44	-.39	-.13	-.56	+.75	+.18	-.55	+.09	.00	-.23
US	+.09	-.04	+.57	-.10	+.12	-.26	+.77	+.38	-.21	+.30	+.23	.00

Note: Reading across the row and comparing performance with country listed in heading: Positive effect sizes reflect higher average performance; Negative effect

sizes reflect lower average performance.

The effect sizes associated with country differences in the IAEP2 and TIMSS surveys are contained in Tables 3 and 4 respectively and are based on the weighted *n*s, scale scores, and standard deviations (see Appendix A and B). Scale scores for IAEP2 were taken from the public use data file. Changes in effect sizes between pairs of means on the assessments are the absolute values of the difference between the effect size for the IAEP2 assessment and the effect size for TIMSS, i.e.,

$$d_{change} = |d_{IAEP2} - d_{TIMSS}|.$$

These absolute values are presented in Table 5.

Table 5
Absolute Value of the Differences Between the Effect Sizes Observed in Science for IAEP2 and TIMSS Lower Grade

	Can	Eng	Fra	Hun	Ire	Kor	Por	Sco	Slo	Spa	Swi	US
Can	.00	.15	.56	.06	.35	.15	.38	.26	.31	.11	.48	.25
Eng	.15	.00	.69	.21	.17	.29	.48	.38	.14	.27	.57	.11
Fra	.56	.69	.00	.58	.90	.44	.08	.25	.99	.43	.12	.69
Hun	.06	.21	.58	.00	.35	.08	.45	.22	.36	.08	.40	.33
Ire	.35	.17	.90	.34	.00	.46	.79	.58	.01	.48	.77	.08
Kor	.15	.29	.44	.08	.46	.00	.24	.12	.45	.02	.31	.43
Por	.38	.48	.08	.45	.79	.24	.00	.16	.93	.30	.05	.49
Sco	.26	.38	.25	.22	.58	.12	.16	.00	.57	.17	.18	.47
Slo	.31	.14	.99	.36	.00	.45	.93	.57	.00	.48	.82	.02
Spa	.11	.27	.43	.08	.48	.02	.30	.17	.48	.00	.37	.33
Swi	.48	.57	.12	.40	.78	.31	.05	.18	.82	.37	.00	.67
US	.25	.11	.69	.33	.08	.43	.49	.47	.02	.33	.67	.00

Note: Slight differences between the absolute values in this table and the values in Tables 3 and 4 on which they are based result from rounding error.

Reading across the columns or down the rows gives the effect size differences for a country compared to all other countries. For example, the difference between the effect sizes for Canada and England in the two assessments is 0.15 standard deviation units – a small difference reflecting the fact that the mean achievement in both countries is not significantly different in either assessment.

Most of the largest effect size differences are associated with France, Ireland, and Switzerland (Table 5). Large effect size differences are evident at the intersection of France and Ireland (0.90) and at the intersection of Ireland and Switzerland (0.77). This

is a reflection of the fact that while Ireland's standing relative to these countries was poor in IAEP2, Ireland scored higher than these countries in TIMSS. The intersection of France and Switzerland shows a small effect size difference (0.12) and confirms that these countries maintained their position relative to each other on both occasions. However, effect sizes at the intersection of France and countries such as England (0.69), Hungary (0.58), Slovenia (0.99) and the US (0.69) are large. The Swiss change of fortune is clearly reflected in the effect size differences between it and England (0.57), Slovenia (0.82), and the US (0.67).

Moderate to large effect sizes are also associated with comparisons involving Portugal, Scotland, Slovenia, and the US. For example, the effect size difference at the intersection of Portugal and Slovenia is 0.93. In both assessments, Portugal scored significantly lower than Slovenia. However, the large value results from the fact that while the effect size was in the order of 0.45 in IAEP2, it increased to 1.39 in TIMSS. Indeed, most of the other large effect sizes associated with Portugal reflect that country's very poor performance in TIMSS. Other moderately large effect sizes worth noting are those at the intersections of Scotland and Slovenia (0.57), Scotland and the US (0.47), Korea and Slovenia (0.45), Slovenia and Spain (0.48), and Korea and the US (0.43). Other analyses, not reported here, show that the absolute value of differences between effect sizes observed for mathematics, though large in some cases, are generally much smaller than for science (O'Leary, 1999).

Conclusion

The dilemma that our findings give rise to for policy makers seems straightforward enough. Do the findings (for more countries at any rate) indicate a change in level of science achievement over time? And if not, which results are to be taken as a 'true' reflection of its nation's achievement? Careful consideration now needs to be given to the task of trying to explain why performance in the two assessments seems to be so different for some countries. At least five hypotheses can be suggested (see Beaton et al., 1990 for a description of efforts to disentangle the 1985/86 reading anomaly in the National Assessment of Educational Progress in the United States). These, each of which will be briefly considered, relate to population definitions, survey implementation, approaches to data analysis, the possibility of real gains or losses in the achievement of students in some countries during the period between the two surveys and measuring instrument issues.

Firstly, differences in population definitions might account for differences in the relative performance of students in IAEP2 and TIMSS science. In IAEP2 a sample of students who were 13 years old was tested. In TIMSS the students were in grades 7 and 8. While there is some overlap between these two populations, there are differences between them that need to be taken into account when comparing performance. For example, it is noteworthy that for TIMSS science the estimated median scale score for Irish 13-year olds (486) is lower than the mean scale score for Irish seventh graders (495) and that the median score for Swiss 13-year-olds is exactly equivalent to the Irish mean at the seventh grade (see, Beaton et al., 1996b, pp. 26 and 37).

(A median scale score rather than a mean scale score was calculated for 13-year-olds in TIMSS due to the fact that students were sampled by grade and not by age. Not all 13- year-olds were in the grades sampled and, as a consequence, an estimate of the median was thought to be more reliable.) Ramseier (1997, personal communication) claims that a large part of the change in Swiss performance between IAEP2 and TIMSS can be explained by the fact that 44% of Swiss 13-year olds are in

grade 8. He argues that comparing Swiss grade 8 performance to the performance of grade 7 students in Ireland (where most 13-year olds are) provides evidence that Swiss IAEP2 and TIMSS performances may not be all that different. However, taking the sampling variability of both medians into account, it must still be argued that, as the scores for both sets of 13-year olds suggest, Switzerland did not perform significantly better than Ireland in TIMSS. (The standard errors of the Irish and Swiss medians were 3.1 and 2.2 respectively).

Secondly, populations with exclusions and low participation rates in some countries may also account for some of the differences in outcomes across the two studies. Exclusions were caused by countries modifying the internationally agreed definition of the population to be tested. Low participation rates were caused by having combined school and student participation rates below an agreed cut-off mark (70% in IAEP2 and 75% in TIMSS). A few examples will suffice to illustrate the point. In IAEP2, Spain excluded students in Cataluna but included them in TIMSS. In IAEP2, Switzerland tested in only 15 of the 26 Cantons whereas 22 Cantons were involved in TIMSS. In IAEP2, England had a final participation rate of only 48% while in TIMSS it was closer to 80% after replacement. Indeed, a particularly vexing question in international assessments (or any large- scale assessment for that matter) is the extent to which exclusions and participation rates affect overall performance (see Linn & Baker, 1995).

Thirdly, differences in approaches to data analysis may account for differences in the relative performance of students in IAEP2 and TIMMS science. Both IAEP2 and TIMSS use complex procedures for estimating average percentage correct and average proficiency scale scores. Technical reports that were published in conjunction with the assessments indicate that the technologies differed for the two surveys. For example, approaches to handling missing data when calculating average percents for items differed across the two studies (not reached items were treated as not administered in IAEP2 while they were treated as incorrect in TIMSS). Moreover, in IAEP2, average scale scores were calculated using a 3-parameter Item Response Theory model, while in TIMSS a modified Rasch model was used (see Adams, Wilson & Wang, 1997). The fact that TIMSS items were matrix sampled (using a BIB design) and that a plausible values technology was used makes it a very different kind of survey to the more straightforward IAEP2.

Fourthly, between 1991 and 1995, levels of science achievement for students around 13 years of age may have increased or decreased, accounting for differences in the relative performance of students in IAEP2 and TIMSS science. We do not, however, have any evidence to support the view that substantial change occurred in the achievement of Irish 13- year old students during the four years between IAEP2 and TIMSS. Comparing outcomes from the two assessments, all we can say is that, in a normative sense, Irish performance in TIMSS improved. Comparison with the Swiss is important here. Ramseier (1997, personal communication) suggests that age, instruction time and curriculum issues affected Swiss performance in TIMSS. Was Ireland's favorable comparison with the Swiss in TIMSS merely an artifact of poor Swiss performance? Of course Ireland's performance relative to more than one country improved and this suggests that achievement in a real sense may have improved. But we cannot say for sure. While the time-span between the two assessments is probably not long enough to allow for the kind of gains that might help explain the improved relative performance in TIMSS, the matter of how performance in IAEP2 can be equated with performance in TIMSS in an absolute sense is a substantial matter and one that is of the utmost importance to an accurate interpretation of national performance in the two

surveys.

Fifthly, differences in measuring instruments might account for differences in the relative program of students in IAEP2 and TIMSS science. As noted above, there were differences in the content areas of the IAEP2 and TIMSS tests. TIMSS had a section entitled Environmental Issues which IAEP2 did not. There were also differences in the proportion of items assigned to common content areas. For example, while 17% of the IAEP2 items were devoted to the Nature of Science, the figure for TIMSS was 6%. In addition, more of the TIMSS test (5%) was devoted to Physics. Hence, differences in performance may be a function of differences in the nature of the achievement that was assessed. However, an interesting issue arising in this context is worth raising here. The fact is that while the instruments measuring mathematics achievement also differed in content coverage, the mathematics performance of countries across the two studies was more consistent. The question arises: In international studies do particular factors impinge much more strongly on science achievement than mathematics achievement?

Finally, and as an extension of the last point, what seems reasonably clear is that underlying the reporting of results of international studies in the popular media and in many reports emanating from government ministries is an assumption that 'science,' 'mathematics,' 'reading' and the like are clearly understood. But is this the case? Can we say that there is real consensus about the nature of these domains and the underlying psychological constructs implied by "achievement" in these subjects? Or could it be that at the international level an understanding of what constitutes achievement in mathematics, for example, is at a more advanced level than the understanding of what constitutes science achievement? It is noteworthy that some support for this hypothesis is contained in our finding that country rank orderings were more stable in mathematics than in science across two distinct international assessments. Moreover, in the United States the analysis by Hamilton and her colleagues (1995) of a large scale national test (NELS:88) provides further food for thought in suggesting that "achievement patterns in science were much more heterogeneous than in math" and that "[i]n science, a far greater number of factors was required to account for student performance differences" (p. 577). Such findings raise critical questions about the science tests used in international comparative studies.

Note

The poor performance of Irish students in science was also a feature of the First International Assessment of Educational Progress in Mathematics and Science (IAEP1) test in 1988 (Lapointe, Meade, & Phillips, 1989).

References

Adams, R. J., Wilson, M., & Wang, W-C. (1997) The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement,* 21, pp. 1-23.

Beaton, A. E., Mullis, I. V. S., Martin, M. O., Gonzalez, E. J., Kelly, D. L., & Smith, T. A. (1996a) *Mathematics achievement in the middle school years: IEA's Third International Mathematics and Science Study*. Chestnut Hill, MA: Center for the Study of Testing, Evaluation, and Educational Policy, Boston College.

Beaton, A. E., Martin, M. O., Mullis, I. V. S., Gonzalez, E. J., Smith, T. A., & Kelly, D. L (1996b) *Science Achievement in the Middle School Years: IEA's Third International*

Mathematics and Science Study. Chestnut Hill, MA: Center for the Study of Testing, Evaluation, and Educational Policy, Boston College.

Beaton, A. E., Zwick, R., Yamamoto, K., Mislevy, R. J., Johnson, E. G., & Rust, K. F. (1990) *Disentangling the NAEP 1985-86 Reading Anomaly.* Princeton, NJ: Educational Testing Service.

Brown, M. (1996) FIMS and SIMS: the first two IEA international mathematics surveys, *Assessment in Education: Principles, Policy & Practice*, 3, pp. 193- 212.

Brown, M. (1998) The tyranny of the international horse race. In R. Slee & G. Weiner with S. Tomlinson (Eds) *School Effectiveness for Whom? Challenges to the School Effectiveness and School Improvement Movements*, pp. 33- 47 (London, Falmer Press).

Cohen, P. (1977) *Statistical Power Analysis for the Behavioural Sciences.* New York: Academic Press.

Durlak, J. A. (1995) Understanding meta analysis. In L. G. G. P. R. Yarnold (Eds.), *Reading and Understanding Multivariate Statistics*, pp. 319-352 Washington, D.C.: American Psychological Association.

ETS (1992) *IAEP technical report: Volume 1*. Princeton, NJ: Educational Testing Service.

Goldstein, H. (1996) Introduction. *Assessment in Education: Principles, Policy and Practice*, 3, pp. 125- 128.

Hair, J. F., Anderson, R. E., & Black, R. L. T. C. (1995) *Multivariate Data Analysis (4th ed)* Englewood Cliffs, NJ: Prentice Hall.

Hamilton, L. S., Nussbaum, E. M, Kupermintz, H., Kerkhoven, J. I. M., & Snow, R. S. (1995) Enhancing the validity and usefulness of large-scale educational assessments: NELS:88 science achievement. *American Educational Research Journal*, 32, pp. 555-581.

Husén, T. (1967) *International Study of Achievement in Mathematics*, 2 vols. Stockholm: Almquist & Wiksell.

Husén, T., & Postlethwaite, T. N. (1996) A brief history of the International Association for the Evaluation of Educational Achievement (IEA). *Assessment in Education: Principles, Policy and Practice*, 3, pp. 129- 141.

Kellaghan, T. (1996) IEA studies and educational policy. *Assessment in Education: Principles, Policy and Practice*, 3, pp. 143-160.

Kellaghan, T., & Grisay, A. (1995) International comparisons of student achievement: Problems and prospects. In *Measuring What Students Learn*, pp. 41-61 Paris: Organization for Economic Co-Operation and Development.

Lambin, R. (1995) What can planners expect from international quantitative studies?, in W. Bos & R. H. Lehman (Eds) *Reflections on Educational Achievement. Papers in honour of T. Neville Postlewaithe*, pp. 169-182 New York: Waxman.

Lapointe, A. E., Askew, J. M., & Mead, N. A. (1992) *Learning Science.* Princeton, NJ: Educational Testing Service.

Lapointe, A. E., Mead, N. A. & Phillips, G. W. (1989) *A World of Differences: An International Assessment of Mathematics and Science.* Princeton, NJ: Educational Testing Service.

Linn, R. L., & Baker, E. L. (1995). What do international assessments imply for world-class standards? Implications of international assessments. *Educational Evaluation and Policy Analysis,* 17*, 4*, pp. 405-418.

Martin, J. O., Hickey, B. L., & Murchan, D. P. (1992) The second international assessment of educational progress: mathematics and science findings in Ireland. *Irish Journal of Education*, 26, pp. 3-146.

Martin, M. O., & Kelly, D. L. (1996) *Third International Mathematics and Science Study. Technical report. Volume 1: Design and Development* (Chestnut Hill, MA, Center for the Study of Testing, Evaluation, and Educational Policy, Boston College).

Murphy, P. (1996) The IEA assessment of science achievement, *Assessment in Education: Principles, Policy & Practice*, 3, pp. 213-232.

Nuttall, D. (1994) Choosing indicators, In *Making Education Count. Developing and using international indicators*, pp. 79-96 (Paris, Organisation for Economic Co-operation and Development)

O'Leary, M. (1999) *The validity and consistency of findings from international comparative studies of student achievement: A comparison of outcomes from IAEP2 and TIMSS.* (Unpublished doctoral dissertation, Boston College, Massachusetts).

Plomp, T. (1992) Conceptualizing a comparative educational research framework. *Prospects*, 22, pp. 278-288.

Winer, B. J., Brown, D. R., & Michels, K. M. (1991) *Statistical Principles in Experimental Design.* New York: McGraw Hill.

Wolf, F. M. (1986) *Meta Analysis: Quantitative Methods for Research Synthesis.* Newbury Park, CA: SAGE.

About the Authors

Michael O'Leary
St. Patrick's College
Drumcondra, Dublin 9, Ireland
Telephone +353-1-8842000
Fax +353-1-8376197
Email: Michael.OLeary@spd.ie
Michael O'Leary is a member of the Education Department at St.Patrick's College, Dublin, Ireland. He holds a Ph.D. from Boston College in the area of educational

research, measurement and evaluation. He has served as Ireland's representative on the Board of Participating Countries for the OECD's Programme for International Student Assessment (PISA) project.

Thomas Kellaghan
Thomas Kellaghan is Director of the Educational Research Centre at St. Patrick's College, Dublin. He is a member of Academia Europaea and a Fellow of the International Academy of Education. He is currently serving as President of the International Association for Educational Assessment.

George Madaus
George Madaus is Boisi Professor of Education and Public Policy at Boston College. He has served as Director of the National Commission on Testing and Public Policy, and as Vice President of AERA Division D, and as President of NCME. He is a a member of the National Academy of Education and a Senior Research Associate of the National Board on Educational Testing and Public Policy.

Albert Beaton
Albert Beaton is a professor in Boston College's Graduate School of Education. He was international study director of the Third International Mathematics and Science Study (TIMSS) and is a former director of design, research, and data analysis for the National Assessment of Educational Progress (NAEP). He is a member of the International Academy of Education and an honorary member of the International Association for the Evaluation of Educational Achievement (IEA).

Appendix A
Average Science Scale Scores for 13-year-olds in IAEP2

	n	Weighted n	Scale Score	se	sd
Can	4980	182312	534	1.5	61
Eng	929	504590	533	3.9	71
Fra	1787	672764	531	2.5	69
Hun	1623	149647	552	2.3	72
Ire	1657	63791	509	2.5	72
Kor	1635	671867	570	2.3	68
Por	1520	149228	504	3.8	72
Sco	1584	55398	529	2.8	69
Slo	1598	26640	536	2.2	65
Spa	1609	440322	525	2.3	61
Swi	3653	52726	553	3.4	63
US	1404	3028386	523	4.4	68

Source: International Assessment of Educational Progress (IAEP2), 1991-1992.

Appendix B
Average Science Scale Scores at Grade 7 in TIMSS

	n	Weighted n	Scale Score	se	sd
Can	8219	377731	499	2.3	90
Eng	1803	465457	512	3.5	101
Fra	3016	860657	451	2.6	74
Hun	3066	118727	518	3.2	91
Ire	3127	68477	495	3.5	91
Kor	2907	798409	535	2.1	92
Por	3362	146882	428	2.1	71
Sco	2913	62917	468	3.8	94
Slo	3600	28049	530	2.4	86
Spa	3741	549032	477	2.1	80
Swi	4085	66681	484	2.5	82
US	3886	3156847	508	5.5	105

Source: IEA's Third International Mathematics and Science Study (TIMSS), 1994-1995.

EPAA Editorial Board