DOCUMENT RESUME

ED 454 252 TM 032 855

AUTHOR	O'Leary, Michael
TITLE	Item Format as a Factor Affecting the Relative Standing of Countries in the Third International Mathematics and Science Study (TIMSS).
PUB DATE	2001-04-00
NOTE	28p.; Paper presented at the Annual Meeting of the American Educational Research Association (Seattle, WA, April 10-14, 2001).
PUB TYPE	Reports - Evaluative (142) -- Speeches/Meeting Papers (150)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	Comparative Analysis; Constructed Response; Cross Cultural Studies; Foreign Countries; *International Education; *Mathematics Achievement; Multiple Choice Tests; Secondary Education; *Secondary School Students; Test Construction; Test Format; *Test Items
IDENTIFIERS	*Ireland; Rank Order; *Third International Mathematics and Science Study

ABSTRACT

Data from the Third International Mathematics and Science Study (TIMSS) were examined to determine the extent to which the rank ordering of countries based on pupil test performance was consistent across three different item formats: multiple-choice, short-answer, and extended-response. Findings from the analysis are used to make the case that international comparative studies are very complex and that the data they generate cannot be taken at face value but need close examination before firm conclusions can be drawn about a country's relative performance. The focus was the science performance of Irish second year secondary school students (Grade 8) in TIMSS across different item types, comparing this with the performance of similar cohorts in 11 other countries. Irish student performance was close to the international averages for short-answer and multiple-choice items, but performance on extended-response items was significantly above the international average. An examination of the match of these test items to the Irish curriculum was not good, and the Irish curriculum was judged to encourage higher-order thinking less than in other countries. Both of these factors made the good performance on extended-response items surprising. In many respects, these findings confirm the suspicion of W. Cooley and G. Leinhart (1980) that frequent exposure to test format will make a difference in performance. In Ireland there is a tradition of more open-ended essay type tests, and this may account for students' success with extended-response items. These findings also demonstrate the difficulties involved in making international comparisons of academic performance. An appendix contains a table of science averages for TIMSS participants. (Contains 63 references.) (SLD)

Item Format as a Factor Affecting the Relative Standing of Countries in the Third International Mathematics and Science Study (TIMSS).

Michael O'Leary
St Patrick's College
Drumcondra
Dublin 9
Ireland
Michael.oleary@spd.ie

Paper presented at the meeting of the American Educational Research Association Seattle, April 10-14, 2001

Item Format as a Factor Affecting the Relative Standing of Countries in the Third International Mathematics and Science Study (TIMSS).

Since the early 1960s international comparative studies have been designed to provide policy makers, educators, researchers and others with information about pupil achievement and the functioning of different educational systems. While it is not always clear exactly how the results of these studies are used in the various countries, there is evidence to suggest that they often attract a good deal of attention especially when a country's results are poor (see, Colvin, 1996; Innerst, 1996; Lally, 1997). A case in point is the number of references in the popular media to the low levels of literacy among many Irish participants in the International Association for the Evaluation of Educational Achievement (IEA) study of Adult Literacy Study (Morgan, Hickey, & Kellaghan, 1997). Two tasks associated with a recent international study are undertaken in this paper. First, data from the Third International Mathematics and Science Study (TIMSS) are examined to determine the extent to which the rank ordering of countries based on pupil test performance is consistent across three different item formats – multiple-choice, short answer and extended-response. Second, the findings from this analysis are used to make the case that international comparative studies are very complex undertakings and that the data they generate cannot be taken at face value but need to be closely examined before firm conclusions about the relative performance of a nation can be reached.

The paper begins with a review of the literature on the effects of item format on performance in international surveys and other tests. The TIMSS survey is then described briefly. Following that, the performance of Irish second year secondary school pupils on multiple-choice, short-answer, and extended-response item sets from the TIMSS science test is compared and contrasted in an effort to determine if Ireland's relative position in the rank ordering of participating countries remains stable. The paper concludes with a discussion about the implications of the findings for how international survey data are interpreted, reported and used.

Background.

Issues of efficiency and appropriateness are usually the criteria used by test constructors to choose item types. For example, multiple-choice items are considered to be an efficient way of measuring knowledge and of ensuring that large amounts of curriculum content are covered by the test. On the other hand, extended-response items are considered to be more appropriate for assessing process skills and higher-order thinking. In international tests, as Lapointe, Askew and Mead (1992) note, there is the added concern that "the testing format ... is not equally familiar to students from all countries" (p. 11). Indeed, item format might even be regarded as another aspect of opportunity to learn (OTL). As Cooley and Leinhart (1980) have shown, "students are more likely to answer correctly if they have been taught the specific material covered by the test, and if they have been frequently exposed to the test format" (quoted in Winfield, 1993, p. 290). Wolf (1998) points to the well known fact that students in the United States are well used to answering multiple-choice questions, while students in many European countries are more often tested using free- or extended-response questions. This is the case in Ireland, where the two major examinations at the secondary school level (the Junior and Leaving Certificates) are dominated by short-answer and essay-type questions. Indeed, Wolf (1998), also notes that constructed-response items were used in the first IEA mathematics study not because they were the best or most efficient method but because of the need to appease countries less familiar with alternative item types. Again, this raises the issue of fairness across educational systems. In the past, international assessments of science achievement have relied predominantly on multiple-choice items even though many commentators point out that open-ended items fit more closely with science teaching around the world and provide the better option for testing process skills (Kjoernsli & Jorde, 1992; Stedman, 1994). In addition, some commentators have argued that multiple-choice should not be the predominant testing mode when evaluating the output of schools and educational systems (e.g., Madaus & Kellaghan, 1992). In general, multiple-choice items have been criticised for failing to measure significant learning outcomes and complex abilities thinking (e.g., Aschbacker, 1991) and provide little information about students' understanding or quality of thinking (e.g., Gipps & Murphy, 1994; Darling-Hammond, 1994). However, others disagree with such views and contend that multiple-choice items are capable of measuring more than just basic curriculum facts or simple recall (e.g., Airasian, 1997).

In TIMSS, 102 of the 135 science items were multiple-choice. The remaining items consisted of 22 short-answer and 11 extended-response items (Beaton, Martin, Mullis, Gonzalez, Smith, & Kelly, 1996). This is in sharp contrast to other large-scale assessments of science. For example, in the US's National Assessment of Educational Progress (NAEP), only 20% of the science tests were multiple-choice and many questions required the assessment of students' actual performance of tasks (Atkin & Black, 1997). In the view of Atkin and Black (1997), the TIMSS test by virtue of its format does not fit well with efforts to reform mathematics and science curricula in many countries where the emphasis is directed "towards applications, toward practical work, toward increasing students' capacity to see real-world relevance, and toward enhancing students' enthusiasm for further study of the subject" (p. 25). However, the not inconsiderable problems associated with administering open-ended and/or performance items which include extra cost in time and effort, rater effects, inconsistent scoring and lower generalisability, have been well documented in the literature (see, for example, Gipps, 1995; Huchison & Schagen, 1994; Madaus & Kellaghan, 1993) and may help to explain the practice/policy mismatch.

In addition to the issue of student familiarity with item type, there is the issue of whether or not multiple-choice tests and open-ended response tests are psychometrically equivalent. According to Perkhounkova, Hoover and Ankemann (1997) "if the goal is to make relative comparisons among students, the psychometric equivalence of tests of different formats can be established by showing that the tests rank-order examinees in the same way, after adjustments for test unreliability are made" (p. 2). Messick (1989) uses the term discriminant validity to refer to evidence that shows consistency across different methods of measurement. The literature on this topic though large, (e.g., Bennet and Ward, 1993), is equivocal. Studies by Bridgeman (1992), Bennet, Rock and Wang (1991), Lukhele, Thissen, and Wainer (1994), Thissen, Wainer and Wang (1994), and Perkhounkova et al. (1997) found that multiple-choice and open-ended items measured the same basic trait or proficiency. However, Birenbaum and Tatsuoka (1987) and Birenbaum, Tatsuoka, and Gutvirtz (1992) found that format did make a difference when the purpose of the assessment was diagnostic. In general, it was found that constructed-response items provided better information in so far as they helped to provide more in-depth information on the test taker. Messick (1993) also noted that format effects tended to be dependent on the purposes of the assessment and varied across content areas and the degree of structure in the format. In terms of large scale studies, Hamilton (1997) found that in the National Educational Longitudinal Study (NELS:88) in the US the total score for a student masked differences among item types within a test. For example, Hamilton concluded that while males achieved higher than females when compared on total score, this was, in fact, "due to performance differences on one type of item and not to overall superiority in science" (p.22). The item type in this case was multiple-choice. A similar outcome resulted from an earlier study conducted in Ireland by Bolger and Kellaghan (1990) in which males were found to perform relatively better than females on multiple-choice tests compared with free-response tests.

In the realm of international comparative studies, findings with respect to the effect of item format on test performance also vary across studies. Using pilot data from the IEA's reading literacy study, Elley and Mangubhai (1992) compared outcomes on multiple-choice and open-ended items (based on similar reading passages) for a cross section of 9- year-olds from one Australian city and one New Zealand city. Three conclusions from this study are worth noting. The first is that item format had no significant impact on the outcomes of the study. As Elley and Mangubhai report it: "Those students who did well on one test were the same ones who did well on the other, regardless, of item format" (p. 196). The second was that the multiple-choice format produced higher scores on average due in part to the fact that the open-ended items produced many more omissions or "don't know" responses. The third was that most students (88%) preferred the multiple-choice format. Elley and Mangubhai point out that their findings (with respect to the effects of item format on 9-year-olds) were consistent with the outcomes of studies undertaken by Vernon (1962) of UK and US college students, by Choppin and Purves (1969) of UK and US 14- and 18-year-olds, by Traub and Fisher (1977) of Canadian eight graders and by Van der Berg (1988) of Dutch 15- year-olds. In another study of IEA reading literacy data, a range of questions were raised by Kapinus and Atash (1995) about the use of multiple-choice and constructed-response items and whether the latter were worth the time and cost required to answer and score them. Among the questions they addressed were:

- What is the relationship between scores on multiple-choice test items and scores on constructed-response items?
- What can be said about the psychometric qualities (e.g., reliabilities) of constructed-response items?

With respect to the first question, Kapinus and Atash's analysis lead them to the conclusion that:

> while there [was] a significant relationship between the two variables, nevertheless, based on [the] coefficient of determination, the variance in common between the two variables was at best 33 percent. While some of the variation not common between the two measures (i.e., unique variation) may be due to measurement error (i.e., measurement error tends to attenuate the relationship between the two measures), it appears that the two variables [were] measuring different aspects of reading proficiency. (p. 127)

With respect to the second question, the researchers found that the estimated reliability for the constructed-response items was lower than for multiple-choice items (which was not surprising given that there were many more multiple-choice items). They also warned that the number of systematic and random error components associated with constructed-response items (e.g., scoring or ambiguities in the scoring guides) was much larger than with multiple-choice items. With that in mind, the findings of two studies using TIMSS data present an interesting contrast. Mullis and Smith (1996) report that generalisability analyses resulted in high degrees of reliability in the relative ranking of countries based on data from the TIMSS free-response mathematics and science items (generalisability coefficients were slightly lower for science). On the other hand, Jakwerth and her colleagues (1997) contend they found great instability in country ranks across the item formats in TIMSS (p. 26). Unfortunately, they did not report the specifics of their analyses or results.

Other studies at the international level relate more to comparisons using performance items, but the findings from these studies are worth noting. In the First International Science Study (FISS), England and Japan administered a "practical test" as well as the main multiple-choice format test at the ninth grade level. The practical tests consisted of five tasks that required the use of simple apparatus and simple laboratory facilities. The multiple-choice section also contained a set of items designed to assess practical work in science. Comber and Keeves (1973) concluded that "the evidence from these suggests that such practical tests measure quite different abilities from those assessed by the more traditional tests, even those designed to assess practical skills as far as possible without resort to actual apparatus" (p. 288). In the Second International Science Study (SISS), five countries administered a practical test at ages 10 and 14 (Hungary, Israel, Japan, Singapore, and the US). The findings indicated a weak correlation between the practical science tests and the main SISS test (Kjoernsli & Jorde, 1992). The question of whether international tests rank order countries differently on the basis of different item formats is an interesting one and one that can be addressed with the TIMSS data, given the variety of formats utilised.

The TIMSS Survey

In 1995 TIMSS was conducted in 45 countries around the world. The principal focus of TIMSS was on the mathematics and science achievements of pupils in the grades containing most 9-year-olds (equivalent to 3rd and 4th class in Ireland), most 13-year-olds (equivalent to First and Second Year in Ireland) and in the final year of secondary education (Martin & Kelly, 1996).

The TIMSS test booklets contained both mathematics and science items. At the seventh and eight grades the mathematics test was comprised of 151 items and the science test was comprised of 135 items. All items were rotated across eight test booklets and student performance on these booklets were matrix sampled using a modified Balanced-Incomplete-Block spiraling (BIB) design (Beaton, Martin, & Mullis. 1997).[1] Each booklet was completed by students in two timed blocks of 44 and 46 minutes -- a total of one and one half-hours of testing time in all.

Three item formats were used in the main TIMSS test (a performance assessment using performance tasks was also administered in some countries but not in Ireland). Figure 1 presents an example of each item type used in TIMSS to measure science performance in Chemistry.

[1] In TIMSS, clusters or blocks of items were rotated within eight test booklets and these booklets were randomly assigned to sampled pupils.

Figure 1. *Item Formats Used in TIMSS.*

Multiple-Choice	Short-Answer	Extended-Response
If a neutron atom loses an electron, what is formed/ A. A gas B. An ion C. An acid D. A molecule	Carbon dioxide is the active material in some fire extinguishers. How does carbon dioxide extinguish a fire?	It takes 10 painters 2 years to paint a steel bridge from one side to the other. The paint that is used lasts about 2 years, so when the painters have finished painting at one end of the bridge, they go back to the other end and start painting again. a. Why must steel bridges be painted b. A new paint that lasts 4 years has been developed and costs the same as the old paint. Describe 2 consequences of using the new paint.

Source: TIMSS, 1994.

As should be evident from the figure, the multiple-choice format required pupils to select a correct answer from four choices. The free-response formats required that pupils supply a short answer or provide a longer response showing their work and/or providing explanations for their answers (Beaton et al., p. A8).

Analysis

The principal focus of the analysis undertaken for this paper is on the science performance of Irish second year secondary school pupils (Grade 8) in TIMSS across different item types. The performances of similar cohorts of students in 11 other countries are also considered. These countries are Canada, England, France, Hungary, Korea, Portugal, Scotland, Slovenia, Spain, Switzerland, and the US. These countries were chosen to represent a range in the distribution of performance levels in TIMSS (above average, average, below average) and were the focus of another study by the author that compared the performance of countries that had participated in both the second International Assessment of Educational Progress (IAEP2) (Lapointe, Askew, & Meade, 1992) and TIMSS (see, O'Leary, 1999). The full set of overall average percents correct in science for all countries that participated in TIMSS at the first and second year levels is contained in Appendix A.

When the results of international comparative studies are being discussed it is always tempting to talk in terms of rank ordering based on overall averages because this is the simplest and most straightforward way in which to present country differences. However, the reality is that ranks have limited meaning and this is especially true in situations where country averages are "statistically indistinguishable from one another" (Baker, 1997, p. 296). Simply discussing Ireland's performance in terms of rank ordering would do little justice to the complex business of making *meaningful* comparisons between educational achievement in different countries. Moreover, there is agreement with Beaton (1998) that "[i]t is unwise to treat rankings as critical when the means on which they are based differ by less than could be expected by sampling and measurement error" (p. 539). Therefore, in this paper a change in rank order is considered important only if it implies a concomitant change in the statistical relationship between two or more country means. Because the focus of interest was on comparing the Irish mean to the means of the other common countries, a suitable Bonferroni critical value was set to guard against the probability of Type 1 errors (Winer, Brown, & Michels, 1991). This critical value was based on the alpha level (.05) adjusted for 11 comparisons.[2]

[2] In this study a Type 1 error would occur if the researcher concluded that there was a statistically significant difference between two country means when, in fact, there wasn't (rejecting the null hypothesis when it is true). Because the likelihood of finding a significant difference between two means by chance increases when many means are compared, the researcher needs to make the criteria for finding a statistically significant difference more stringent. The Bonferroni procedure allows the researcher to set this criteria in light of the number of comparisons being made.

Results

Table 1 presents data with respect to the average percents correct in TIMSS for the three item types by grade for students in Second Year. Analyses pertinent to First Years (Grade 7) are discussed in O'Leary (1999) and reveal similar findings. To aid analysis, the overall average percent correct for each country is included and countries are categorised in terms of the significance of the difference between each average and the Irish average.

Table 1

Average Percents Correct at Grade Eight[a] for 12 Countries Across Different Item Sets in TIMSS (Categorised in Terms of the Significance of Difference of Each Average from the Irish Average)[b]

Overall 135 Items 146 Score Points[c]			Multiple-Choice 102 Items 102 Score Points			Short-Answer 22 items 25 Score Points			Extended-Response 11 items 19 Score Points		
	$\bar{x}$	se		$\bar{x}$	se		$\bar{x}$	se		$\bar{x}$	se
Kor	65.5	0.3	Kor	70.2	0.4	Kor	62.1	0.9	**Eng**	**54.6**	0.9
Slo	61.7	0.5	Slo	66.5	0.5	Eng	61.9	1.0	**Ire**	**52.8**	1.2
Eng	**61.3**	0.6	Hun	65.6	0.5	**Hun**	**59.0**	1.1	**Kor**	**52.6**	0.7
Hun	**60.7**	0.6	**Eng**	**63.7**	0.6	**Slo**	**58.0**	0.9	Can	48.7	0.7
Can	**58.7**	0.5	**Can**	**62.2**	0.5	**Can**	**57.3**	0.6	Swi	48.4	0.8
Ire	**58.4**	0.9	**US**	**61.9**	0.9	**Spa**	**56.0**	0.8	Slo	47.7	1.1
US	**58.3**	1.0	**Ire**	**61.3**	0.9	**Ire**	**55.7**	1.2	Sco	47.6	1.2
Swi	**56.3**	0.5	**Swi**	**59.8**	0.5	**US**	**54.5**	1.2	US	47.1	1.3
Spa	**55.6**	0.4	**Spa**	**59.2**	0.4	**Swi**	**52.7**	0.7	Hun	43.6	1.0
Sco	**55.3**	1.0	**Sco**	**58.7**	1.0	**Sco**	**52.4**	1.3	Spa	41.8	0.6
Fra	53.7	0.6	Fra	57.9	0.6	Fra	49.9	1.0	Fra	40.7	0.9
Por	49.9	0.6	Por	55.5	0.6	Por	43.7	0.9	Por	34.1	0.7
Int'l[d]	58.0			61.9			55.3			46.6	

[a] Grade 8 in most countries.

[b] Average performance in countries within the shaded area is not statistically significantly different to that in Ireland. Average performance in countries above the shaded area is statistically significantly above that in Ireland. Average performance in countries below the shaded area is statistically significantly below that in Ireland. Statistically significant at the 0.05 level, adjusted for 11 comparisons.

[c] Some of the TIMSS science items had more than one part and this resulted in a total of 146 score points in all.

[d] The average of the 12 country averages.

Source: IEA (1997).

In terms of Irish performance, what becomes readily apparent in the table is that Irish students rank higher on extended-response items than they do on multiple-choice or short-answer items. Indeed, while Irish average percents correct for multiple choice (61.3%) and short-answer (55.7) are close to the international averages for these item sets (61.9 % and 55.3% respectively), the Irish average for extended response (52.8%) is significantly above the international average (46.6%). It is noticeable in the table that Irish performance on the 11 extended-response items is either comparable to, or significantly better than, the performance of countries achieving much higher averages overall (e.g., Korea and Slovenia). On the same item set, Irish pupils also performed significantly better than pupils from Canada, Hungary, Scotland, Slovenia, Spain, Switzerland and the US even though there was no significant difference between performances when the overall test was considered. Given that the TIMSS test consisted of multiple-choice items predominantly (70%), it is not surprising to find that the ranking of countries for the multiple-choice item set reflects the overall rankings fairly closely. What is harder to explain is the fact that in comparison to the overall rankings, the relative standings of countries is also fairly stable for the short-answer item set but unstable for the extended-response items even though the weightings for these item types in the overall test were equally small (17% and 13% respectively).[3]

At this point it may be prudent to examine the extended-response item set in more detail to determine if the strong Irish performance here could have been helped by a good test-curriculum match rather than the format of the test questions *per se*. Table 2 provides details about the content category and performance expectations for the 11 items.

[3] These percentages are derived using the score points rather than the number of items.

Table 2

Classification of the Extended-Response Items in TIMSS by Content Category and Performance Expectation

Item ID	Content Category	Performance Expectation
L04	Physics	Applying and Investigating Scientific Principles
M11	Life Science	Understanding Complex Information
O14	Earth Science	Applying and Investigating Scientific Principles
W01	Earth Science	Applying and Investigating Scientific Principles
W02	Earth Science	Applying and Investigating Scientific Principles
X01	Life Science	Applying and Investigating Scientific Principles
X02	Life Science	Applying and Investigating Scientific Principles
Y01	Physics	Applying and Investigating Scientific Principles
Y02	Physics	Applying and Investigating Scientific Principles
Z01	Chemistry	Applying and Investigating Scientific Principles
Z02	Environmental Issues	Applying and Investigating Scientific Principles

Source: Beaton et al (1996); Ramseier (1997).

The TIMSS science test was comprised of six content areas: Physics, Chemistry, Earth Science, Life Science, Environmental Issues and the Nature of Science (the latter two were combined for reporting purposes due to the small number of items involved). A feature of the TIMSS reporting of content area performance were profiles designed to show whether participating countries performed better or worse in some content areas than they did on the test as a whole (see Beaton et al., 1996, pp. 40-44). Irish second years were shown to have performed better in Earth Science and Environmental Issues/Nature of Science, worse in Life Science and Physics, and about the same in Chemistry. It can be seen from Table 2 that only 5 of the 11 extended-response items came from the content areas in which Irish pupils performed relatively well. In addition, Robitaille et al (1993) defined a series of "performance expectations" for all 135 items, which Ramseier (1997) condensed into *Understanding Simple Information, Understanding Complex Information and Applying and Investigating Scientific Principles*. As the names suggest, increasingly sophisticated cognitive functioning is meant to be required to complete items within the categories. What is clear from Table 2 is that most of the extended-response items in TIMSS were classified as cognitively complex. Given the perception that the curriculum in Irish schools encourages higher-order thinking less than in other countries, the relatively strong Irish performance on the TIMSS extended-response item set might be considered surprising. It is also surprising to find that the curriculum-test match for these 11 items was judged to be quite poor (see Table 3).

Table 3

The Test-Curriculum Match for Extended-Response Items in TIMSS

Item ID	Can	Eng	Fra	Hun	Ire	Kor	Por	Sco	Slo	Spa	Swi	US
L04	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes
M11	Yes	Yes	No	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes
O14	Yes	Yes	Yes	Yes	No	No	Yes	Yes	Yes	Yes	Yes	Yes
W01	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
W02	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes
X01	Yes	Yes	No	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes
X02	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	No	Yes
Y01	Yes	Yes	No	Yes	No	No	Yes	Yes	Yes	Yes	No	Yes
Y02	Yes	Yes	No	Yes	No	No	Yes	Yes	Yes	Yes	No	Yes
Z01	Yes	Yes	Yes	Yes	No	Yes	Yes	No	Yes	Yes	Yes	Yes
Z02	Yes	Yes	No	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes
Total Yes	11	11	5	11	6	3	11	10	11	11	8	11

Source: IEA (1997).

In TIMSS a measure of the appropriateness of the science items for each country (or opportunity-to-learn) was achieved from ratings carried out by personnel from each country (Beaton & Gonzalez, 1997).[4] The TIMSS country coordinators were then required to report on whether or not an item was in the country's intended curriculum. A judgement of an item's appropriateness was made on the basis of answers to two questions: 1) is the item topic in the

[4] In TIMSS this process was not documented at the international level but anecdotal evidence suggests that the ratings were carried out by subject specialists (Albert Beaton, personal communication).

intended curriculum for more than 50% of pupils at the grade level? and, 2) is the item topic likely to be encountered by the pupils prior to TIMSS testing? (TIMSS, 1995).

Results in Table 3 show that only 6 of the 11 extended-response items were judged to be curriculum appropriate for Irish pupils. Only France and Korea had a poorer test-curriculum match. In most countries all 11 items were judged to be curriculum appropriate. Again, this provides further evidence to suggest that the item format may be an important factor underlying Irish performance on this item set.

One other issue associated with the extended-response items in TIMSS that may not be readily apparent is that they were placed at the end of answer booklets (or booklet sections). A difficulty that arises in this case is that approaches to test taking can differ across countries and pupils may not reach items at the end of a test due to time constraints or may deliberately omit them due to low motivation. This problem arose in the IEA reading literacy study when "an unusual level of non-completion of the test in some countries" was found (Elley, 1992, p. 99). In the literature on large-scale surveys of achievement questions have been raised about the motivation of students to perform well on tests that have little consequences for them personally (see Kiplinger & Linn, 1995/6; Mislevy, 1995; O'Neill, Sugrue, & Baker, 1995/6). In the case of TIMSS, it could be hypothesised that the relatively strong performance of Irish pupils on the extended-response items was helped by the poor motivation of pupils in other countries (e.g. Korea) to complete these items. In previous research studies the percents of omitted and not-reached item responses have been used as a proxy measure of motivation (see for example, Swinton, 1996). Utilising a similar approach, data on the combined percentages of omitted and not-reached item responses for five extended-response items placed at the end of TIMSS booklets were analysed. These data are presented in Table 4.

Table 4

Combined Percentages Omitting and Not Reaching Extended-Response Items placed at the end of the TIMSS Booklets

Item ID	Can	Eng	Fra	Hun	Ire	Kor	Por	Sco	Slo	Spa	Swi	US
W02	11	12	23	26	10	25	31	13	12	26	14	11
X02(part B)	8	9	25	25	10	10	20	14	11	17	13	9
Y02	9	5	16	18	5	10	19	10	17	16	8	10
Z01(part C)	33	28	60	na	25	9	53	37	47	42	45	32
Z02 (part B)	19	21	42	37	19	19	54	28	45	25	27	20
Average	16	15	33	21	14	15	35	20	26	25	21	16

Source: TIMSS (1996).

These data indicate that while Ireland had the smallest average proportion of pupils not completing items (14%), most other countries performing significantly below Ireland on the extended-response item set had similar proportions of omitted and not-reached responses. Examples in this case include Canada and the US. In fact the data also show that this issue cannot be used to explain why Irish pupils did as well as their Korean counterparts as the latter country had just 1% more pupils on average not attempting these items. The only countries where there seems to be a much larger proportion of pupils not attempting the items are France and Portugal. However, it will be noted from Table 1 that Irish pupils performed significantly better than pupils from both of these countries across all item formats and on the test as a whole. Analyses conducted to determine the effects of omitted and not-reached item responses on the overall average percent correct for individual countries and described in detail in O'Leary (1999) revealed that the overall impact of missing responses did not affect the average percents correct to an extent that would alter the country rankings on the extended-response item set.

Conclusions and Implications

In many respects these findings confirm the suspicion of Cooley and Leinhart (1980) that frequent exposure to test format will make a difference to performance. Given the fact that multiple-choice standardized tests are more prevalent in the US, while in Ireland there is a tradition of more open-ended essay type tests, these results make sense. In addition, given the British experience with performance testing over the past decade or so, it is not surprising to find that English students outperform their counterparts in other parts of the world in supply items requiring an extended-response. It will be particularly interesting to see how Irish 15-year-olds perform in the upcoming Programme for International Student assessment (PISA) survey where a greater proportion of extended-response items are being used than in any previous international survey (OECD, 1999).

In the context of future international surveys such as PISA there may be an even more important implications of the findings presented here. When the results of international comparative studies of pupil achievement are published the principal focus of attention is often on the rank ordering of countries based on overall mean scores. However, the reality is that sometimes overall mean scores are not the best yardstick for judging a country's performance. According to Mislevy (1995) "the fundamental law of data aggregation is that collapsing information simultaneously (a) highlights the common pattern and (b) obscures patterns that are unique" (p. 426). In Goldstein's (1997) view, the emphasis given to aggregated scores has two principal drawbacks. It disguises interesting patterns of achievement and it reflects the weightings of topics chosen by those who constructed the test. Another issue is that in international assessments a test in a given subject area is composed of sets of items weighted differently by topic or sub-domain. In the TIMSS test of science achievement, for example, 60% of the items were devoted to Life Science and Physics, two areas in which Irish pupils did relatively badly. The two areas in which Irish pupils did relatively well, Earth Science and Environmental issues/Nature of Science, contributed just 26% of the total item set. Again, it could be argued that Ireland's overall performance would have benefited had the latter topic areas received greater emphasis. Kellaghan and Grisay (1995) make a similar point about Ireland's mathematics ranking in the first IAEP study in arguing that it would have been improved had the proportion of number items been even greater than it was. The point stressed by Mislevy (1995) is that while comparisons and rankings can be essentially the same within a set of items, they can differ substantially across a set of items. In the words of Airasian and Madaus (1983) "the situation is analogous to scoring in the decathlon: two contestants may have the same total score across the ten events, but perform very differently in each specific event" (p. 106). In this paper we have seen that in comparison with many other TIMSS' contestants, the extended-response "event" in science seemed to pose less of a problem for Irish pupils. Trying to disentangle why this occurred involved further complexities of content emphases, performance expectations, test-curriculum overlap and the motivation of pupils to do well on international tests. So what is to be done about the rankings if factors such as item format play a role in pupil performance? What is to be done if the full complexity of international survey data is to be acknowledged in initial reportage and subsequent use by policy makers and others? Mislevy (1995) is in no doubt about what he would do:

> ... my answer to people who want comparative standings is to give them comparative standings - lots of them: in different topics, at different ages, with different kinds of

> tasks; unweighted, weighted by national curriculum guidelines, weighted by surveyed opportunity-to-learn; unadjusted results for the full sample, for students in selected courses of study, for students at or above selected percentiles on within-nation performance. (p. 427)

While it must be acknowledged that the rationale for providing such an array of comparisons is commendable, the practicalities of such a plethora of comparisons may be confusing for policy makers and other consumers of international survey data. A more reasoned approach for individual countries like Ireland may lie in a careful consideration of how best to highlight not just the differences but also the similarities between countries and educational systems. After all, rank orderings, in and of themselves, are useful only to the extent to which they facilitate a greater understanding of why some countries perform better on international tests than others. This understanding will follow only from a close examination of the data in all its complexity. It follows then that in each participating country a national report highlighting the unique aspects of the country's performance should receive even more attention than the international reports produced by those responsible for conducting the surveys. A lacuna in terms of Ireland's involvement in TIMSS was that an Irish report was never produced. Ultimately, the real value of international surveys such as TIMSS and PISA will only be derived once the complexity of they data they generate is acknowledged, carefully considered in the national context and used to make informed judgements in the policy arena.

REFERENCES

Airasian, P. W. (1997). *Classroom Assessment*. New York: McGraw-Hill Companies, Inc.

Airasian, P. W., & Madaus, G. F. (1983). Linking testing and instruction: Policy issues. *Journal of Educational Measurement, 20*, 103-118.

Aschbacher, P. R. (1991). Performance assessment: State activity, interest, and concerns. *Applied Measurement in Education, 4*, 275-278.

Atkin, J. M., & Black, P. (1997). Policy perils of international comparisons. *Phi Delta Kappan, 79, 1*, 22-28.

Baker, D. P. (1997). Surviving TIMSS: Or everything you blissfully forgot about international comparisons. *Phi Delta Kappan, 79*, 295-300.

Beaton, A. E. (1998). Comparing cross-national student performance on TIMSS using different test items. *International Journal of Educational Research, 29*, 529-542.

Beaton, A. E., & Gonzalez, E. J. (1997). TIMSS test-curriculum matching analysis. In M. O. Martin & D. L. Kelly (Eds.), *Third International Mathematics and Science Study Technical Report Volume 2: Design and development* (pp. 187-193). 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. (1997). Providing data for educational policy in an international context: The Third International Mathematics and Science Study (TIMSS). *European Journal of Psychological Assessment,, 13*(49-58).

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.

Bennet, R. E., Rock, D. A., & Wang, M. (1991). Equivalence of free response and multiple-choice items. *Journal of Educational Measurement, 28*, 77-92.

Bennet, R. E., & Ward, W. C. (Eds.). (1993). *Construction versus choice in cognitive measurement: Issues in constructed response, performance testing, and portfolio assessments*. Hillsdale, NJ: Erlbaum.

Birenbaum, M., & Tatsuoka, K. K. (1987). Open-ended versus multiple-choice response formats -- it does make a difference for diagnostic assessment. *Applied Psychological Measurement, 11*, 385-395.

Birenbaum, M., Tatsuoka, K. K., & Gutvirtz, Y. (1992). Effects of response format on diagnostic assessment of scholastic achievement. *Applied Psychological Measurement, 16*, 353-363.

Bolger, N., & Kellaghan, T. (1990). Method of measurement and gender differences in scholastic achievement. *Journal of Educational Measurement, 27*, 165-174.

Bridgeman, B. (1992). A comparison of quantitative questions in open-ended and multiple-choice format. *Journal of Educational Measurement, 29*, 253-271.

Choppin, B. F., & Purves, A. L. (1969). A comparison of open-ended and multiple-choice items dealing with literacy understanding. *Research in the Teaching of English, 3*, 15-24.

Colvin, R. L. (1996, November 21). Global study finds U.S. students weak in math. *Los Angeles Times,* pp. 1.

Comber, L. C., & Keeves, J. P. (1973). *Science education in nineteen countries: An empirical study*. New York: John Wiley.

Cooley, W. W., & Leinhardt, G. (1980). The instructional dimensions study. *Educational Evaluation and Policy Analysis, 2*, 7-25.

Darling-Hammond, L. (1994). Performance assessment and educational equity. *Harvard Educational Review, 64*, 5-30.

Elley, W. B. (Ed.). (1992). *How in the world do students read?* The Hague, Netherlands: International Association for the Evaluation of Educational Achievement.

Elley, W. B., & Mangubhai, F. (1992). Multiple-choice and open-ended items in reading tests: Same or different? *Studies in Educational Evaluation, 18*, 191-199.

Gipps, C. V. (1995). Reliability, validity, and manageability in large-scale assessment. In H. Torrance (Ed.), *Evaluating authentic assessment* (pp. 105-123). Buckingham, England: Open University Press.

Gipps, C. V., & Murphy, P. (1994). *A fair test?: Assessment, achievement, and equality.* Buckingham, England: Open University Press.

Goldstein, H. (1997). International comparisons of student achievement. In A. Little & A. Wolf (Eds.), *Assessment in transition: Learning, monitoring and selection in international perspective* (pp. 58-87). Oxford, England: Pergamon.

Huchison, D., & Schagen, I. (Eds.). (1994). *How reliable is National Curriculum Assessment*. Slough, England: NFER-Nelson.

IEA. (1997). *TIMSS International Database: Middle School Years 1995 Assessment: CD ROM*. Chestnut Hill, MA: TIMSS International Study Center, Boston College.

Innerst, C. (1996, November 21). U.S. Students lag in math, science. *Washington Times,* pp. 1.

Jakwerth, P. M., Bianchi, L. J., Houang, R. T., Schmidt, W. H., Valverde, G. A., Wolfe, R. G., & Yang, W. (1997, March). *Validity in cross-national assessments: Pitfalls and possibilities*. Paper presented at the meeting of the American Educational Research Association, Chicago, IL.

Kapinus, B., & Atash, N. (1995). Exploring the possibilities of constructed-response items. In M. Binkley, K. Rust, & M. Winglee (Eds.), *Methodological issues in comparative educational studies* (pp. 105-129). Washington, DC: U.S. Department of Education, National Center for Education Statistics.

Kellaghan, T., & Grisay, A. (1995). International comparisons of student achievement: Problems and prospects. In OECD (Ed.), *Measuring what students learn*. Paris: OECD.

Kiplinger, V. L., & Linn, R. L. (1995/1996). Raising the stakes of test administration: The impact on student performance on the National Assessment of Educational Progress. *Educational Assessment, 3*, 111-133.

Kjoernsli, M., & Jorde, D. (1992, April). *Evaluation in science: Content or process.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Lally, C. (1997, August 31). Poor literacy linked to course attendance. *The Irish Times,* pp. 8.

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

Lukhele, R., Thissen, D., & Wainer, H. (1994). On the relative value of multiple-choice, constructed response and examinee selected items on two achievement tests. *Journal of Educational Measurement, 31*, 234-250.

Madaus, G. F., & Kellaghan, T. (1992). Curriculum evaluation and assessment. In P. W. Jackson (Ed.), *Handbook of curriculum research* . New York: Macmillan.

Madaus, G. F., & Kellaghan, T. (1993). The British experience with authentic assessment. *Phi Delta Kappan, 74*, 485-469.

Martin, M. O., & Kelly, D. L. (Eds.). (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.

Martin, M. O., & Kelly, D. L. (Eds.). (1997). *Third International Mathematics and Science Study Technical Report Volume 2: Implementation and analysis*. Chestnut Hill, MA: Center for the Study of Testing, Evaluation, and Educational Policy, Boston College.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement: Third edition* (pp. 13-103). New York: American Council on Education, Macmillan Publishing Company.

Messick, S. (1993). Trait equivalence as construct validity of score interpretation across multiple methods of measurement. In R. E. Bennet & W. C. Ward (Eds.), *Construction versus choice in cognitive measurement: Issues in constructed response, performance testing, and portfolio assessments* (pp. 61-73). Hillsdale, NJ: Erlbaum.

Mislevy, R. J. (1995). What can we learn from international assessments. *Educational Evaluation and Policy Analysis, 4*, 419-437.

Morgan, M., Hickey, B., & Kellaghan, T. (1997). *International Adult Literacy Survey: Results for Ireland*. Dublin: Stationery Office.

OECD. (1999). *Measuring student knowledge and skills: A new framework for assessment.* Paris: OECD.

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).

O'Neill, H. F., Sugrue, B., & Baker, E. L. (1995/1996). Effects of motivational interventions on the National Assessment of Educational Progress mathematics performance. *Educational Assessment, 3*, 135-157.

Perkhounkova, Y., Hoover, H. D., & Ankemann, R. D. (1997, March). *An examination of construct validity of multiple choice versus constructed response tests.* Paper presented at the meeting of the American Educational Research Association, Chicago, IL.

Ramseier, E. (1997). *Task difficulty and curricular priorities in Science: Analysis of typical features of the Swiss performance in TIMSS.* Paper presented at the annual meeting of the European Conference on Educational Research, Frankfurt, Germany, September, 1997.

Robitaille, D. F., Schmidt, W. H., Raizen, S., McKnight, C., Britton, E., & Nicol, C. (1993). *Curriculum frameworks for mathematics and science: TIMSS Monograph No. 1.* Vancouver, Canada: Pacific Educational Press.

Shiel, G., Forde, P., & Morgan, M. (1996). International studies of achievement: What do they tell us? *Oideas, 44*, 80-97.

Stedman, L. C. (1994). Incomplete explanations: The case of U.S. performance in the international assessments of education. *Educational Researcher, 23,7*, 24-32.

Swinton, S. S. (1997, March). *Block position effects as indirect indicators of motivation in the 1996 NAEP science assessment.* Paper presented at the annual meeting of the American Educational Research Association, Chicago, Il.

Thissen, D., Wainer, H., & Wang, X. B. (1994). Are tests comprising both multiple-choice and free-response items necessarily less unidimensional than multiple-choice tests: An analysis of two tests. *Journal of Educational Measurement, 31*, 113-123.

TIMSS. (1995). *Main Survey Test Items.* Chestnut Hill, MA: TIMSS international Study Center, Boston College.

TIMSS. (1996). *P-value almanac for cognitive items. Science population 2: Upper grade.* Chestnut Hill, MA: TIMSS International Study center, Boston College.

Traub, R. E., & Fisher, C. W. (1977). On the equivalence of constructed-response and multiple-choice tests. *Applied Psychological Measurement, 1*, 355-369.

Van den Berg, H. (1988). *Examens Geexamineerd.* The Hague, Netherlands: S.V.O.

Vernon, P. E. (1962). The determinants of reading comprehension. *Educational and Psychological Measurement, 22*, 269-286.

Winer, B. J., Brown, D. R., & Michels, K. M. (1991). *Statistical principles in experimental design*. New York: McGraw Hill.

Winfield, L. F. (1993). Investigating test content and curriculum content overlap to assess opportunity to learn. *Journal of Negro Education, 62*, 288-310.

Wolf, R. (1977). *Achievement in America.* New York: Teachers College Press.

Wolf, R. M. (1998). Validity issues in international assessments. *International Journal of Educational Research, 29*, 491-501.

Appendix A

Science Average Percents Correct for First and Second Years in TIMSS

First Years			**Second Years**		
Int'l $\overline{X}$	50	(0.1)	Int'l $\overline{X}$	56	(0.1)
Singapore	61	(1.2)	Singapore	70	(1.0)
Korea	61	(0.4)	Korea	66	(0.3)
Japan	59	(0.3)	Japan	65	(0.3)
Czech Republic	58	(0.8)	Czech Republic	64	(0.8)
Belgium (Fl)	57	(0.5)	England	61	(0.6)
England	56	(0.6)	Hungary	61	(0.6)
Hungary	56	(0.6)	Belgium (Fl)	60	(1.1)
Slovak Republic	54	(0.6)	Slovak Republic	59	(0.6)
United States	54	(1.1)	Sweden	59	(0.6)
Canada	54	(0.5)	Canada	59	(0.5)
Hong Kong	53	(1.2)	**Ireland**	**58**	**(0.9)**
Ireland	**52**	**(0.7)**	United States	58	(1.0)
Sweden	51	(0.5)	Russian Federation	58	(0.8)
New Zealand	50	(0.7)	New Zealand	58	(0.8)
Norway	50	(0.6)	Norway	58	(0.4)
Switzerland	50	(0.4)	Hong Kong	58	(1.0)
Russian Feder.	50	(0.8)	Switzerland	56	(0.5)
Spain	49	(0.4)	Spain	56	(0.4)
Scotland	48	(0.8)	France	54	(0.6)
Iceland	46	(0.6)	Iceland	52	(0.9)
France	46	(0.6)	Latvia (LSS)	50	(0.6)
Belgium (Fr)	45	(0.7)	Portugal	50	(0.6)
Iran, Islamic Rep	42	(0.6)	Lithuania	49	(0.7)
Latvia (LSS)	42	(0.5)	Iran, Islamic Rep	47	(0.6)
Portugal	41	(0.5)	Cyprus	47	(0.4)
Cyprus	40	(0.4)	*Countries Not Satisfying Guidelines for Sample Participation*		
Lithuania	38	(0.7)			
			Australia	60	(0.7)
Countries Not Satisfying Guidelines for Sample Participation			Austria	61	(0.7)
			Belgium (Fr)	50	(0.7)
Australia	54	(0.7)	Bulgaria	62	(1.0)
Austria	55	(0.6)	Netherlands	62	(1.0)
Bulgaria	56	(1.0)	Scotland	55	(1.0)
Netherlands	56	(0.7)			
Countries Not Meeting Age/Grade Specifications					
Colombia	35	(0.7)	Colombia	39	(0.8)
Germany	53	(0.8)	Germany	58	(1.0)
Romania	45	(0.7)	Romania	50	(0.8)
Slovenia	57	(0.5)	Slovenia	62	(0.5)
Countries with Unapproved Sampling Procedures at Classroom Level					
Denmark	44	(0.4)	Denmark	51	(0.6)
Greece	45	(0.5)	Greece	52	(0.5)
South Africa	26	(1.0)	Thailand	57	(0.9)
Thailand	53	(0.8)	*Countries with Unapproved Sampling Procedures at Classroom Level and Not Meeting Other Guidelines*		
			Israel	57	(1.1)
			Kuwait	43	(0.9)
			South Africa	27	(1.3)

Standard errors in parentheses.
Source: Beaton et al. 1996.

TM032855

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

ERIC

Reproduction Release

(Specific Document)

I. DOCUMENT IDENTIFICATION:

Title: Item format as a factor affecting the relative standing of countries in TIMSS.	
Author(s): Michael O'Leary	
Corporate Source: St. Patrick's College	Publication Date: 2001

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, Resources in Education (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic media, and sold through the ERIC Document Reproduction Service (EDRS). Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce and disseminate the identified document, please CHECK ONE of the following three options and sign in the indicated space following.

The sample sticker shown below will be affixed to all Level 1 documents	The sample sticker shown below will be affixed to all Level 2A documents	The sample sticker shown below will be affixed to all Level 2B documents
PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY SAMPLE TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)	PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE, AND IN ELECTRONIC MEDIA FOR ERIC COLLECTION SUBSCRIBERS ONLY, HAS BEEN GRANTED BY SAMPLE TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)	PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY SAMPLE TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
Level 1	Level 2A	Level 2B
☑ ✓	☐	☐
Check here for Level 1 release, permitting reproduction and dissemination in microfiche or other ERIC archival media (e.g. electronic) *and* paper copy.	Check here for Level 2A release, permitting reproduction and dissemination in microfiche and in electronic media for ERIC archival collection subscribers only	Check here for Level 2B release, permitting reproduction and dissemination in microfiche only

Documents will be processed as indicated provided reproduction quality permits.
If permission to reproduce is granted, but no box is checked, documents will be processed at Level 1.

I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce and disseminate this document as indicated above. Reproduction from the ERIC microfiche, or electronic media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries.

Signature: Michael O'Leary	Printed Name/Position/Title: MICHAEL O'LEARY	
Organization/Address: St. Patrick's College Drumcondra Dublin 9 Ireland	Telephone: 353-1-8842065	Fax: 353-1-8376197
	E-mail Address: Michael.oleary@spd.ie	Date: MAR 24/2001