DOCUMENT RESUME

ED 358 125 TM 019 904

AUTHOR	Medrich, Elliott A.
TITLE	Improving Data Quality in IEA Studies: Looking Backward and Thinking Forward.
PUB DATE	Apr 93
NOTE	8p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (Atlanta, GA, April 13-15, 1993).
PUB TYPE	Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)
EDRS PRICE	MF01/PC01 Plus Postage.
DESCRIPTORS	*Academic Achievement; *Comparative Analysis; Cross Cultural Studies; *Data Collection; Elementary Secondary Education; Foreign Countries; Futures (of Society); International Studies; National Surveys; Quality Control; Response Rates (Questionnaires); *Scores; *Test Results
IDENTIFIERS	Educational Information; Evaluation Standards; *International Assn Evaluation Educ Achievement; Target Populations; *Third International Mathematics and Science Study

ABSTRACT

Since studies conducted by the International Association for the Evaluation of Educational Achievement (IEA) have had a dramatic impact on the way in which officials in the United States and the American public think about the performance of our students, it is essential that IEA surveys accurately measure real differences in student performance across comparable populations in participating countries. Although data quality in past IEA studies has sometimes been problematic, the upcoming Third International Mathematics and Science Study (TIMSS) affords the opportunity to develop methods of data presentation that achieve reliable cross-national comparisons. Two issues in particular merit consideration. The first issue is ensuring that field outcomes in participating countries are comparable and representative of a defined target population. A second aspect concerns survey response rates. It will also be necessary to determine how to deal with data when certain standards are not achieved. One chart lists the number of participating systems in the various IEA studies. (SLD)

IMPROVING DATA QUALITY IN IEA STUDIES: LOOKING BACKWARD AND THINKING FORWARD

Elliott A. Medrich
MPR Associates, Inc.
Berkeley, California

Improving Data Quality in IEA Studies: Looking Backward and Thinking Forward

Elliott A. Medrich
MPR Associates, Inc.
Berkeley, California

While cross-national studies of mathematics and science achievement have always been of interest to the American education community, recent comparative studies conducted by the IEA are attracting considerable attention from a broad range of audiences. The IEA surveys, particularly country by country rankings of student achievement in mathematics and science, receive substantial press, and as a consequence, the results have become the subject of intense public debate and discussion. In the U.S., the survey findings have been closely scrutinized, and they have affected the highest levels of government policy, witness the National Education Goals adopted in 1990. So much for the comfortable early days, when these studies were viewed as experimental, "works in progress" so to speak. Today's international achievement studies have succeeded in capturing the public spotlight, something which I am sure the founding fathers had never envisioned, even in their wildest dreams. It is this public face of the IEA surveys that serves as a backdrop to my remarks today.

Since the IEA studies apparently have had a dramatic impact on the way in which officials in the U.S. and the American public generally thinks about the performance of our students, the quality of our school curriculum, and the effectiveness of our teaching practice vis a vis other nations, *it is essential that the surveys accurately measure real differences in student performance across comparable populations in participating countries*. The interest in scores and rankings demands that these data meet high technical standards and achieve statistical reliability.

Looking back over the IEA history, and considering some of the daunting problems associated with these massive data collection efforts, it seems extraordinary how much excellent work has been accomplished. At the same time, it also seems appropriate to consider how TIMSS and beyond will assure two of the most important user groups—policymakers and the public—that the data achieve a level of statistical precision that is necessary to the schooling outcomes debate. To meet the need, stringent data collection standards must be established and achieved by all participating systems. If data quality on past IEA studies has been sometimes problematic, TIMSS and beyond affords and opportunity to rewrite the book, sharpen procedures, and develop methods of data presentation that achieve reliable cross-national comparisons.

The size and breadth of TIMSS is remarkable. Without regard to the number of countries ultimately participating in the study, the undertaking is extremely ambitious. I think we all recognize that the TIMSS technical advisory committee, and others charged with designing, implementing, and refereeing the field study must contend with great variation in survey and data collection capabilities from system to system, and real differences in human and financial resources to support the effort. That said, however, TIMSS also offers an opportunity to address a number of data quality questions that have been identified subsequent to the several previous IEA mathematics and science studies.

In this brief presentation, I will highlight just two issues, one pertaining to data collection, and the other concerning data presentation. I must leave to those more qualified than I the difficult task of assuring that appropriate data collection instruments and standards are derived, and that field methods are adequately tested, and universally implemented. Here I can only reflect on what it is we are trying to collect—and by inference, who it is we are trying to compare.

I think it is safe to say that almost every public official, press representative, and other "secondary" user I speak with— that is those who are reading published IEA results and using these data without further analysis—work under the assumption that all the participating systems surveyed the same populations, in the same ways, and that, therefore, the populations are "comparable." We know, unfortunately, that this has not been the case in the past. It is important to recognize that some problems of comparability can be solved, while other problems can not. An example of a problem that is virtually impossible to solve concerns surveys at the last year secondary level. From system to system the proportion of the age cohort still in-school varies considerably (from 80 or 90 percent or more in some countries to under 50 percent in others). Comparing samples of students from systems with dramatically different participation patterns will always be problematic, even if it can be argued that, strictly speaking, "in-school populations at the last year secondary" are the sample reference group. An example of a problem that is possible to solve concerns surveys at the pre-secondary level. Theoretically, at least, nearly comparable age-grade cohorts could be sampled for all participating systems. As a practical matter, however, a variety of resource and survey administration issues can make it almost as difficult to achieve sample comparability at pre-secondary as it has been at the last year secondary level.

But we need somewhere to start, and one appropriate place would be to describe and contrast, from system to system, who exactly has been surveyed. In fact is the implications of different systems having different samples, has not been discussed with much enthusiasm in the IEA survey reports. The previous IEA surveys have not succeeded in implementing uniform sampling strategies, and the result is that national targets have not held up well against an international standard. The result has been a fair amount of confusion, and worse, real concern that fair comparisons of populations could not be derived. The issue is not just which sampling design is selected by the IEA, but whether the field outcomes, from system to system, are comparable and representative of a defined target population. It is essential, at the least, that future surveys enable those using and reporting the data be able to ascertain the degree to which samples represent targets, and and that there be incentives to encourage participants to achieve samples that meet international targets. Part of the process of validating the survey data in the public forum requires that the following kinds of information be reported and clearly discussed in research reports and public documents:

1. What was the international sampling frame? *This should be the standard against which we should measure comparability and representativeness of each systems samples.*

2. What was the sampling frame used by each system, and how did that compare with the international sampling frame?

3. What were the system field outcomes in comparison with its national sampling frame?

4. What were the system field outcomes in comparison with the international frame?

To assure appropriate use and interpretation of these data, only systems that do well against the international sampling frame and that achieve high response rates , a point I will discuss below, should comprise the "main table" data set. Other systems should be part of an "appended" data set. For analytical purposes, systems should be sorted and reported against the international frame—this not to exclude the hard working national teams, who may have only modest success against the international standard. Rather, the standard should be applied to make clear that those systems compared in the main table reports designed and executed their field work to a similar standard of comparability, and achieved a similar standard of field outcomes. I understand that IEA has a strategy such as this in mind for TIMSS, and I trust that they will follow through.

A second aspect of data quality concerns survey response rates. While this was not addressed as a serious problem in past studies, it is now receiving considerable attention from the IEA. While there is no universally agreed upon statistical basis for defining the adequacy of response rates, it is surely safe to say that an 85 percent rate at each stage of sampling (which is the NCES benchmark) is healthy, and that as the rate declines, confidence in the data must decline as well. As Jeanne Griffith and I have pointed out in a number of papers, few educational systems participating in previous IEA studies have achieved high response rates against national much less international targets, so the challenge here is significant.[1] I might add that data from the U.S. has been problematic in this regard, as has data from many of the highly developed countries. I am pleased to see that the preliminary version of the TIMSS sampling manual has adopted the 85 percent standard. I also gather that response rates will be evaluated against original samples, not replacements. I regard this as an important development and I trust it will help increase confidence in the TIMSS data set. I hope that this standard will also be applied when the IEA selects systems for main table presentation.

I close by noting that effective evaluations of data quality require that we have a common set of criteria against which to judge field outcomes. The examples I have posed here—defining a common target; determining a metric against which to evaluate response rates; and facing up to the problem of how data are handled when certain standards are not achieved—represent important issues that we must try to address in TIMSS and beyond.

Forthcoming surveys will do a great service by speaking plainly about differences in the nature of the field experience from system to system, and by providing clear roadmaps that will enable us to understand how comparable, or not, the results of the surveys are, across educational systems. To accurately interpret findings, we need improvements in data collection standards and outcomes, but we also need a fresh and open approach to quality of data questions. A perceived shortcoming of international achievement surveys in the past has been that there is not sufficient discussion of what did and did not happen in the field. Nor was there a strategy for addressing the consequences of differences in data collection outcomes. This need not be the case in the future, and we are all looking to TIMSS to set the stage for more informed discussion of these concerns.

[1] See Elliott A. Medrich and Jeanne E. Griffith, International Mathematics and Science Assessments: What Have We Learned? (Washington, D.C.: U.S. National Center for Education Statistics, 1992); and Jeanne E. Griffith and Elliott A. Medrich, "What Does the United States Want to Learn from International Comparative Studies in Education?" *UNESCO Prospects*, Fall 1992.

Number of Participating Systems Known to Achieve 85 Percent Response Rate at Each Level of Sampling

	Age 10		Age 13		Age 14		Last-year secondary	
	Total participating	Known to achieve 85% criterion	Total participating	Known to achieve 85% criterion	Total participating	Known to achieve 85% criterion	Total participating	Known to achieve 85% criterion
First Mathematics Study	—	—	12	0	—	—	12 10	0 (Math students) 0 (Non-Math students)
Second Mathematics Study	—	—	20	4	—	—	15	6
First Science Study	17	6	—	—	19	7	19	4
Second Science Study	15	8	—	—	17	10	14 14 14	1 (Biology) 1 (Chemistry) 2 (Physics)

— Not sampled.