The Validity of Comparative Educational Studies

Timo M. Bechger, *National Institute of Educational Measurement*
Godfried van den Wittenboer, *University of Amsterdam*
Joop J. Hox, *Utrecht University*
C. De Glopper, *University of Amsterdam*

What are the conditions that will allow for appropriate comparisons among groups? Should the concept of comparative validity be added to the lexicon of the psychometrician?

Various branches of the social sciences are concerned primarily with the comparison of existing groups or populations. The focus on the comparison of existing groups is the reason that these branches are labeled the *comparative social sciences* (Easthope, 1974). Examples are cross-cultural psychology (Bleichrodt & Drenth, 1990; Poortinga, 1976), and comparative education (Bos & Lehman, 1995; Keeves, 1992).

In most methodological papers, it is concluded that the extent to which the measurement permits comparison is the main methodological problem of the comparative social science. Although similar problems arise in the various branches of the comparative social science, we focus on comparative education. This article is intended to provide a clear explanation of the conditions required for valid comparison of educational achievement among existing groups and to stimulate discussion about the validity of international educational comparisons.

Throughout the article, the reading literacy study that was recently conducted by the International Association for the Evaluation of Educational Achievement (IEA) serves as a specific example. The results of the reading literacy study were reported in detail by Elley (1992, 1994), Postlethwaite and Ross (1992), Lundberg and Linnakylä (1993), Binkley and Trevor (1996), and Wagemaker, Taube, Munck, Kontogiannopoulou-Polydorides, and Martin (1996). Technical reports were published by the IEA (Wolf, 1995) and the U.S. Department of Education (1995). We will occasionally refer to a similar study that was conducted by the Organisation for Economic Co-operation and Development (OECD, 1995, 1997). Since it is our objective to discuss methodological issues, the IEA study is discussed briefly. A more detailed evaluation of this study is published elsewhere (Bechger, Van Schooten, De Glopper, and Hox, 1998) followed by a comment from Warwick B. Elley (Elley, 1998) who was, at the time, chairman of the Steering Committee of the IEA study.

To illustrate the methodological problems involved in comparative education, the article starts with a brief description of the reading literacy study. This sets the stage for a description of the conditions required for valid comparison, which are then discussed in more detail.

A Brief Description of the IEA Reading Literacy Study

In the period 1989–1992, the IEA investigated reading literacy in about 32 systems of education in different nations in order to assess the relative level of reading literacy in each society and to examine whether differences in reading literacy are associated with differences in a number of characteristics of nations, schools, and teachers. The study focused on two levels in each educational system: the grade level where most 9-year-olds were to be found and the grade level where most 14-year-olds were to be found. Pupils in separate special education schools were not included in the defined populations.

The IEA defined reading literacy as the ability to understand and use those written language forms that are required by society and/or valued by the individual (Elley, 1994, pp. 5–6). For various reasons, the IEA used short passages of prose followed by multiple-choice questions to assess reading literacy (Elley, 1998; Kapinus & Atash, 1995), while the OECD study made use of real life tasks, such as filling in

Timo M. Bechger is a Postdoctoral Fellow at the National Institute of Educational Measurement (CITO), P.O. Box 1034, 6801MG, Arnheim, The Netherlands. His specializations are structural equation models and psychometrics.

Godfried van den Wittenboer is an Associate Professor, Faculty of Educational Sciences, University of Amsterdam, Wibautstraat 4, 1091 GM, Amsterdam, The Netherlands. His specialization is mathematical psychology.

Joop J. Hox is a Professor, Department of Methods and Statistics, FSW Utrecht University, P.O. Box 80140, 3508 TC, Utrecht, The Netherlands. His specializations are multilevel modeling, structural equation modeling, and meta-analysis.

Cees De Glopper is a Professor, Faculty of Educational Sciences, University of Amsterdam, Wibautstraat 4, 1091 GM, Amsterdam, The Netherlands. His specializations are language learning and international comparative research.

deposit slips or following a bus timetable. Subjects from each of the nations involved were administered translated versions of the questionnaire. (The original tests were written in English.) Their scores were used to calculate mean reading literacy scores, which were then used to compare the reading literacy of various populations: nations, boys versus girls (Wagemaker et al., 1996), rural versus urban, and so forth. To be able to compare the mean values, it is essential that reading literacy be measured on the same scale in each population. The IEA claims that this is true in spite of cultural differences among subjects from different populations and the fact that subjects were administered different versions of the tests.

To obtain comparability of the reading literacy scores across nations, the IEA standardized the measurement procedures among nations as much as possible. A large pilot test was conducted to detect items that might not be suitable for comparison. Representatives from each nation were given various opportunities to express their opinions about the instruments and to provide suggestions for translation. In addition, the Rasch model (Fischer & Molenaar, 1995; Rasch, 1960) was fitted to the data and used to detect incomparable items and standardize the scale of measurement across nations.

Establishing the rank order among groups was not the only purpose of the reading literacy study. To understand why some educational systems were more effective than others, the IEA compared mean differences among high and low scoring nations and used regression techniques to relate differences in reading literacy to numerous variables measuring characteristics of students, their teachers, their schools, and homes. This revealed among other things that girls read better than boys in nearly all countries, that female teachers tend to produce better readers, and that large school libraries are beneficial. Some of these findings were used to formulate policy recommendations. It was recommended, for example, that educational budgets be prepared to establish adequate book supplies (Postlethwaite & Ross, 1992, p. 43).

Comparative Validity

Most contributions to comparative education deal with *comparability*, a term that is used informally to indicate that a particular comparison is believed to be valid. We propose to use the more precise term *comparative validity*. According to the 1985 standards for educational and psychological testing, validity refers to:

> The appropriateness, meaningfulness, and usefulness of the specific inference made from test scores. . . . A variety of inferences may be made from scores produced by a given test, and there are many ways of accumulating evidence to support any particular inference. Validity, however, is a unitary concept. Although evidence may be accumulated in many ways, validity always refers to the degree to which that evidence supports the inferences that are made from the scores. The inferences regarding specific uses of the test are validated, not the test itself. (American Educational Research Association, American Psychological Association, and National Council of Measurement on Education, 1985, p. 9)

Following this definition, we define *comparative validity* as the appropriateness, meaningfulness, and usefulness of comparative inferences made from test scores. We believe that comparative researchers have the responsibility to provide support for the validity of the comparative statements they make.

The literature mentions many kinds of *comparability*. Some refer to the meaningfulness of the comparison, while others deal with aspects of the appropriateness of comparative statements made from measurements. In Section 3, we discuss both these aspects separately. In Section 4 and Section 5, we discuss several issues that we believe to be important in relation to the appropriateness and meaningfulness of comparative inference.

The Appropriateness of Comparative Statements

The methodological literature is decidedly nonstandard. Numerous synonymous or similar concepts are used throughout the literature on comparative research to express the importance of comparability of measurement across groups. Among these: equivalence of measurement (Hulin, Drasgow, & Parsons, 1982), conceptual equivalence (Drenth & Groenendijk, 1992, p. 9), comparability of measurement (Cook & Campbell, 1979), comparability (van der Flier, 1977; van der Flier & Drenth, 1980), psychometric comparability (Novy, Nelson, Goodwin, & Rowzee, 1993), psychometric invariance (Bejar, 1980), construct invariance (Marsh & Grayson, 1994), factorial invariance (Mulaik, 1972, chap. 14), cross-cultural transportability (McCauley & Colberg, 1983), translation equivalence (Albert, 1986), functional equivalence, equivalence, statistical equivalence (Alwin, Braun, Harkness, & Scott, 1994; Goldstein, 1993, 10.1), score equivalence and item equivalence (Elley, 1994, p. 27; Poortinga, 1976), metric equivalence (Berry, 1980, p. 10), cross-ethnic equivalence (Knight, Yun Tein, Shell, & Roosa, 1992), construct equivalence (Turban, Francis, Orburn, & Sanders, 1989), and conceptual invariance (Armer, 1973). We believe that there are three basic ideas underpinning this myriad of terms. We propose the following definitions for these three basic ideas.

Construct Equivalence

The term *construct equivalence* indicates that the instruments (such as multiple-choice items) used in separate groups measure the same construct regardless of whether the instruments are identical. Without this assumption, we would be comparing apples and oranges. For example, if foreign speakers are administered a test for mathematical ability, the test is likely to measure reading ability, and the scores of these subjects are not comparable to those obtained by native speakers. Construct equivalence may also be violated when different subjects use different solution strategies to solve the same items. This may be investigated, for example, by collecting verbal protocols during testing (e.g., Farr, Pritchard, & Smitten, 1990, or Kapinus & Atash, 1995), but, as far as we know, verbal protocols have never been collected to compare the response behavior of subjects from different nations.

In the context of a comparison of educational achievement, we propose to distinguish between two aspects of construct equivalence (cf. Embretson, 1983). First, we require that the same theoretical mechanisms underlie task performance in each group. Second, we require that the pattern of relations of a test to other measures be similar across groups. Clearly, the more we know about the construct of interest, the better we are able to investigate construct equivalence.

Scale Equivalence

In general, construct equivalence is necessary but not sufficient for comparative validity because the actual numbers, or labels, that are assigned to subjects may not be comparable even though the same construct has been measured. For example, translated items may measure the same construct, but the responses may not be comparable without transforming them to a common scale (Sireci, 1997). *Scale equivalence* indicates that we measure on a common scale across groups.

If we compare the temperature of two objects, we measure temperature on the same scale. We can be sure that the scale of measurement is equivalent because every thermometer is calibrated against a carefully guarded standard. All this is normal yet quite different from the practice of international comparative surveys where subjects are assigned arbitrary numbers, which are then treated as if they are on the same scale. In absence of a golden standard, there is, however, no way to be sure that translated questionnaires provide numbers on the same scale.

Assuming that construct equivalence holds, the scores of subjects from different nations may be placed on the same scale by means of a measurement model, such as the Rasch model, which was used in the IEA study. Vice versa, the model can also be used to provide evidence for construct equivalence, which is what we mean by measurement equivalence.

Measurement Equivalence

According to the standard definition of validity, there are many ways of accumulating evidence to support any particular inference. A widely accepted way to investigate comparative validity is the use of a formal measurement model, which can be tested empirically. The IEA, for instance, used the Rasch model to gather support for construct equivalence. According to the IEA, "an item that displays different characteristics across countries is likely to have measured something different in some of these countries" (Wagemaker et al., 1996, p. 21). When the conditions required for comparative validity are formulated and tested within the context of a formal measurement theory, we refer to it as *measurement equivalence.*

Our definition of measurement equivalence does not presuppose any particular kind of measurement theory. In practice, various models are used even within the same study. The IEA, for instance, used the Rasch model to analyze the literacy measures, while principal component analysis (e.g., Flury & Riedwyl, 1988) was used to analyze the explanatory variables (Lundberg & Hastedt, 1995). We do believe that researchers should explain how construct equivalence and scale equivalence are represented in their model. Fischer (1974, sect 13.4), for example, explains clearly why the item parameters in the Rasch model should be equal across groups.

The Meaningfulness of the Comparison

Being able to compare observations across groups does not imply that we learn anything from the comparison that might improve the educational system. To focus the attention on this aspect of comparative research, we define *the meaningfulness of the comparison* as the degree to which the data support a substantive interpretation of the observed differences.

In comparative surveys, it is often difficult to explain the observed differences because these studies lack the theory and/or the design to determine why the groups are different. At the same time, causal inferences were made in all the studies we came across, as well as in the popular press (Bracey, 1996). The main reason why causal inference may be impossible is that the data do not contain information about the processes of allocation of individuals into groups (classes, schools, nations, etc.), and the antecedents of observed differences remain essentially unknown. Controlling for background variables may suggest directions for further research, but the explanations for the group differences remain uncertain and may involve variables not measured in the study (Cook & Campbell, 1979). For example, it is difficult to explain the positive association between the number of female teachers at the school and reading ability that was reported by the IEA. To our surprise, we found that the issue of causality is almost completely ignored in comparative studies, as judged from the subject indexes, and progress in this area (e.g., Rosenbaum, 1995) appears to have gone unnoticed.

In view of the difficulty of making valid causal inference in the best of situations, it would be useful if the findings of comparative surveys could be placed in an appropriate substantive theory. Unfortunately, the literature on school or class effects is not sufficiently crystallized to identify a small set of theories as relevant. While it is possible to identify a broad set of variables as reasonably comprehensive, it is hard to go much further in theoretical specification (e.g., Raudenbush, 1995). Therefore, cross-national comparative studies like the IEA study have often produced results that are difficult to explain.

The Usefulness of the Comparison

Many international comparisons are conducted because governments are curious to know how their students do in comparison to other countries and where there is room for improvement. It can be argued that, although such studies are unlikely to produce many new insights, they may provide information that is desired by policymakers, curriculum specialists, or teachers (Elley, 1998). Comparative surveys may, for example, demonstrate the effectiveness of a certain kind of reading instruction in practice (e.g., Binkley, Phillips, & Norris, 1995) or reveal the association between illiteracy and unemployment (OECD, 1995).

We define the usefulness of comparative studies as the degree to which they contribute to the actual improvement of education or the quality of future studies. The usefulness of a comparative study depends foremost on the quality of the data and the validity of inferences made from the data. In addition, it requires that the findings be effectively communicated to subjects both inside and outside the scientific community. To this aim, the IEA and the OECD have published several reports for different types of readers. In addition, the IEA has made its international data set available, so that the data can be used for further research.[1] International data sets have the advantage that one can investigate the influence of variables that vary over nations, such as the length of the school year.

A detailed discussion of the properties that large-scale educational assessments should exhibit to be useful for educational policy is provided by Messick (1987). As an aside, we note here that the results of international studies are often useless to policymakers because the variables that are found to be associated with students' achievement cannot be manipulated—for example, gender or gross national product (Lambin, 1995).

Issues Related to the Appropriateness of a Comparison

Comparative Validity, Measurement Procedures, and Data Collection Methods

In general, measurement procedures and data collection methods should be standardized and made as similar as possible across all comparisons. Otherwise, the differences of interest become confounded with differences in the measurement procedure.

In most international comparative surveys, a serious effort is made to standardize the tests and the conditions under which the responses are elicited. At the same time, the literature describes numerous instances where a particular measurement procedure did not work for some groups and had to be adapted (Armer, 1973; Cole, 1977). Elder (1973), for instance, found that Indian subjects refused to answer direct questions, and he had to reformulate his items to obtain (valid) responses. A similar situation occurs when blind subjects are to be compared to seeing subjects. In situations such as these, it is difficult to achieve scale equivalence unless the measurement is unaffected by differences in the measurement procedure.

In general, the reason why measurement procedures are adapted in practice is that, even when the measurement procedures are physically identical, subjects from different groups may respond differently to them. In these cases, the desire to obtain valid measures conflicts with the need to have equivalent scales.

Translation

Cross-lingual assessments are difficult because differences in ability are completely confounded with differences in the measurement procedure due to translation. To avoid invalid differences due to translation, special procedures have been developed, which can be classified into two broad categories.

With *forward translation,* a single translator, or preferably a group of translators, translates the test from the source language to a target language. Then, another group of translators judges the equivalence of the two versions of the instrument. Sometimes examinees are asked to provide translators with their interpretation of the material on the test. This will point out ambiguous items and instructions and provide information on the validity of the items (e.g., Mehan, 1973).

Backward translation proceeds differently. In one variety, a group of translators translates the instruments from the source language to the target language. A second group of translators takes the translated instrument and translates it back to the source language. Then, the original version of the instrument and the back-translated version are compared, and judgments are made about their equivalence.

It depends on the context of the study whether systematic translation is convincing as a criterion for comparative validity. General guidelines are given by Brislin (1976, 1986), Geisinger (1994), and Hambleton and Kanjee (1994, 1995), and the International Test Commission (ITC) is currently preparing guidelines (Hambleton, 1994). Sireci (1997) discusses issues related to the scale equivalence of translated tests. These authors consider systematic translation necessary, but far from sufficient for comparative validity, and recommend combining the subjective procedures with an investigation of measurement equivalence.

Componential Item Response Models

Knowledge about the cognitive processes that underlie the responses may be used to formulate a *componential item response model,* which may then be used to test measurement equivalence. In the IEA study, for instance, items were classified according to the things students must do to solve the item (Elley, 1994, pp. 10–15; Wagemaker et al., 1996)—for example, whether subjects would have to go beyond the information given and make inferences in arriving at the correct answer. This information was not used in the data analyses. The Rasch model treats each item as a single task and has no implications for the nature of the solution process. To obtain a more powerful test of measurement equivalence, the researchers involved in the IEA study could, for instance, have used the Linear Logistic Test Model (LLTM; e.g., Fischer, 1974, 1995, and references therein). The LLTM is an extension of the Rasch model, which recognizes that different subtasks may be involved in solving different items. Embretson and Wetzel (1987) report an application of this model to reading comprehension items.

There are now several componential item response models available (e.g., Embretson, 1985; Maris, 1995), most of which have never been applied in comparative studies. We believe that comparative researchers should use them to test measurement invariance, when such models are appropriate to their data.

Issues Related to the Meaningfulness of a Comparison

The Comparability of the Target Populations

Usually, only those subjects are included in the comparison who hold certain specified attributes or fulfill certain specified criteria (e.g., Schleicher, 1995, pp. 221–223). In comparing students across nations, it is desirable to assess students of similar maturity; that is, students should be of the same age. At the same time, it is useful to compare students with similar amounts of formal education. That is, students should be in the same grade. Thirdly, it is important that students have received exposure to a comparable breadth and depth of material in the subject being assessed; that is, the curricula to which students have been exposed should be similar from country to country (Rust, 1995).

In practice, it is often difficult to obtain samples across countries that are grade-based and yet age-comparable. The IEA attempted to solve the problem by sampling adjacent pairs of grades that best cover the age population of interest. In this manner, a high proportion of students defined by a 12-month age span would be included, and thus comparisons across countries of students of the same age could be carried out. Still, some countries deliberately decided to test either the grade above (Indonesia) or the grade below (Canada, BC) the one that fitted the description of the target population. There were other nations that did not know the age/grade distribution in their system of education because no official information was available. Defining populations with comparable curricula turns out to be an even bigger problem (Goldstein, 1993, sects 4.2, 8.1).

General guidelines are difficult to develop. Even samples that are similar in age and grade may be inadequate for a fair comparison. For example, in the second international mathematics study, advanced algebra and calculus tests were administered to students of similar age who attended advanced science and algebra classes. However, the percentage of students admitted to these classes differed across nations (e.g., Robitaille & Garden, 1989). As a consequence, the nations with the most selective policy received the highest ranks.

Kish (1994) notes that there is an important difference between survey aspects (definition of concepts, variables and populations, methods of measurement) and sampling aspects (sample size, weighting procedures). While survey aspects must be closely controlled for the comparison, sampling methods may be free and flexible as long as they are probability samples. For example, although convenient, sample size (or fraction) need not be balanced across groups, and the possibility of comparing results across countries does not depend on the surveys having similar designs. The use of a formal sampling design is required, however, because convenience samples do not allow a fair comparison (Armer, 1973; Little, 1982). Unfortunately, it is often difficult to obtain probability samples, and, even if probability samples could be achieved, one might have to offer explicit incentives to avoid incomparability due to nonresponse (OECD, 1996, chap. 5).

Mean Differences Should Be Invariant Under Arbitrary Changes of Scale

When means are calculated, the variables are treated as if they are on an interval scale, while they are normally on an ordinal scale only. In this situation, the meaningfulness of comparative studies is questionable if the observed differences are not invariant under order-preserving transformations so that the rankings may be artifacts of the particular metric that has been selected. An example is given by Zwick (1993), who demonstrates that mean differences can reverse if the data are transformed by an exponential function. Invariance under order-preserving transformations is especially important in comparative education where scale information is typically regarded as arbitrary or of little interest and the instruments are translated and therefore different across groups.

Lehmann (1955) proves that a difference between means will be invariant under any monotone transformation of the data if and only if the distributions are *stochastically ordered*—that is, if and only if the ordering of the respective cumulative distribution functions is constant and the graphs of the cumulative distribution functions never cross. A significant t test will then establish the direction of the ordering and hence establish not only a significant difference between the means but also between all the percentile points on the two distributions, including the median. The distributions are stochastically ordered if they have the same shape but differ in location, which is a standard assumption in many parametric and nonparametric tests for group differences. The t test, for instance, assumes that the distributions are normal with equal variance in each group. Hence, there are good reasons to test this assumption.

Darlington (1973) discusses a simple graphical procedure to test whether distributions are stochastically ordered. This procedure also incorporates many of the standard statistical tests. A test for stochastically ordered distributions is given by Jonckheere (1954). (See also Kendall & Stuart, 1967, Vol. 2, p. 505.) For further information, we refer the reader to Davidson and Sharma (1988) or to Maxwell and Delaney (1985).

The Differences Between Means Should Be Large Enough

Comparative studies frequently report differences between means. As a rule, only the statistically significant differences are subject to causal interpretation. If the number of subjects is large, mean differences may well be statistically significant although they are trivial from a substantive point of view. Usually, the scale of measurement is not well understood, and comparative researchers resort to *effect size measures* to determine whether differences are large enough to be of interest. For example, the IEA reports significant differences between boys and girls in reading literacy (Wagemaker et al., 1996, Tables 5, 6). However, Cohen's d for Population A was 0.009, and 0.04 for Population B, which indicates that the differences are trivial according to the usual standard. If the results are reported for individual countries, the advantage of girls over boys is large in many of them (d > 0.7), especially in the narrative domain.

Note that there are many measures of effect size (e.g., McGraw & Wong, 1992; Mishra, Shah, & Lefante, 1986). A careful choice of effect size measure may fascilitate the interpretation of mean differences.

Discussion

Recent decades have seen a rapid increase in the number of large-scaled multinational comparative studies launched both by official international agencies and by academic researchers (e.g., Kish, 1994).[2] As the number of participating countries increases, the range of cultural differences becomes larger, and it becomes increasingly difficult to obtain internationally comparable data. As soon as more than one language is involved, it is no longer possible to use the same instrument in each group. This article was intended to stimulate discussion about the current popularity of international educational comparison.

We have proposed the term *comparative validity* to indicate the appropriateness, meaningfulness, and usefulness of comparative inferences made from test scores. This definition was based on the AERA/APA/NCME standards from 1985. New standards are about to come out, and we hope that members of the committee that is currently working on the definite version of the new standards find this article helpful.[3] The stance of this article is, we believe, consistent with recent writings of many scholars (e.g., Messick, 1989; Wainer & Braun, 1988).

We have stated that, in general, appropriate comparative statements are premised on evidence that the same attribute is measured on the same scale under the different conditions over which differences are assumed to occur. Comparative inference belongs primarily to the subject-matter domain, but comparative validity can be supported by statistical analysis of studies with probability sampling. To this end, the assumptions involved in a particular comparison must be expressed in a formal measurement model. This was called measurement equivalence. We noted that, if the measurement instruments are constructed with a well-defined and limited set of cognitive operations in mind, this information might possibly be used to formulate a componential item response model.

The meaningfulness of the comparison was defined as the degree to which the data support a causal interpretation of the observed differences. This aspect of comparative validity is frequently overlooked, which suggests that many comparative surveys have established differences among groups without providing an explanation. Although this may be useful as an incentive for further study, it seems unlikely that the knowledge of the differences among groups per se is sufficient to improve the education of pupils in any of the populations in the study. To improve the meaningfulness of comparative educational studies, the first requirement is that the study be supported by educational theory. The meaningfulness may also be improved by gathering longitudinal data or data on relatives, which permit stronger conclusions regarding the direction of causality (Neale, Eaves, Kendler, Heath, & Kessler, 1994; Wadsworth, DeFries, Fulker, Olson, & Pennington, 1995). When mean differences are reported, researchers may use an effect size measure to understand the size of these differences. The bottom line is that comparative researchers should be able to explain the size of the observed mean differences, before these differences are subject to causal explanation. We have also found that researchers should demonstrate that the distributions are stochastically ordered. Otherwise, the observed mean differences are dependent on the scale of measurement.

The idea of comparative validity applies to comparative statements of various kinds. Issues of equivalence and meaningfulness arise in qualitative—or *case-oriented studies*, as well as in studies such as the IEA study, that are called quantitative—or *variable-oriented studies* (Ragin, 1984). Significant differences between counts or proportions between countries, for instance, need not reflect true differences but may be due to the informant's culture dependent thresholds for perceiving and reporting behavior. Similarly, nominal measurements, such as the existence of a democracy, may not be comparable due to idiosyncrasies in the definition of these phenomena. In this context, the comparison requires the existence of a single nominal scale, and we may conduct an experiment with paired comparison to test measurement equivalence (Steyer & Eid, 1993). An experiment of this kind was conducted by Davies and Gorbett (1997), who investigated the similarity of color categories across different languages.

We have used the IEA reading literacy study to illustrate the difficulties involved in a cross-national comparison of reading abilities. We admit that the IEA study is among the best studies in the field and that an international comparison of reading literacy is among the most difficult. Surely, it is easier to obtain comparable measurements in studies involving so-called "unobtrusive measures" (Webb, Campbell, Schwartz, & Sechrest, 1971) that are less susceptible to cultural differences and easier to standardize.

Finally, we have defined the usefulness of comparative studies as the degree to which they contribute to the actual improvement of education, or to the quality of future studies. We do not believe that comparative studies can be useful unless the comparison is both appropriate and meaningful. We would certainly hesitate to base a policy decision on results for which we have no plausible explanation. If comparative studies are to serve as instruments for educational policy, they should be designed as such, following the guidelines given by Messick (1987). To this purpose, control theory might, for instance, provide a framework (e.g., Holly & Hallett, 1989).

Next to being useless, comparative studies may even be harmful when they give rise to discrimination against lower scoring groups. At present, the validity of international educational assessments is not subject to the same scrutiny as studies involving, for example, a comparison in intelligence (e.g., Jensen, 1969). This is harmful because erroneous educational policy affects us all.

Notes

Timo M. Bechger's postdoctorate was financed by the Dutch organization for scientific research (NWO). We thank Cor Sluijter and two anonymous reviewers for their helpful comments. The authors bear sole responsibility for the content of this article.

[1] Information about IEA's data enhancement project can be found at the

following internet address: http:\\utou2.to.utwente.nl\dep\iea-dep.htm.

[2] Survey of studies in comparative education are (at the time of writing) maintained at the following internet addresses: http:\\www.nap.edu/reading room/books/icse/index.html and http://nces.ed.gov/pubs/eiip/eiipsrc.html.

[3] For more information the reader is referred to the following internet adress: http://www.apa.org/science/standards.html.

References

Albert, S. (1986, June). *On translation equivalence.* Paper presented at the International Conference on Translation in Leipzig, Germany.

Alwin, D. F., Braun, M., Harkness, J., & Scott, J.(1994). Measurement in multi-national surveys. In I. Borg & P. Mohler (Eds.), *Trends and perspectives in empirical social research* (pp. 45–60). Berlin: De Gruyter.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1985). *Standards for educational and psychological testing.* Washington, DC: American Psychological Association.

Armer, M. (1973). Methodological problems and possibilities in comparative research. In M. Armer & A. D. Grimshaw (Eds.), *Comparative social research: methodological problems and strategies* (pp. 10–25). New York: Wiley.

Armer, M., & Grimshaw, A. D. (1973). *Comparative social research: methodological problems and strategies.* New York: Wiley.

Bechger, T. M., Van Schooten, E. J., DeGlopper, C., & Hox, J. J. (1998). The validity of international surveys of reading literacy: The case of the IEA reading literacy study. *Studies in Educational Evaluation, 24*(2), 99–125.

Bejar, I. I. (1980). Biased assessment of program impact due to psychometric artifacts. *Psychological Bulletin, 87,* 513–524.

Berry, J. W. (1980). Introduction to methodology. In H. C. Triandis & J. W. Berry (Eds.), *Handbook of cross-cultural psychology. Methodology Vol. 2* (pp. 32–40). London: Allyn & Bacon.

Binkley, M. R., Phillips, L. M., & Norris, S. P. (1995). Creating a measure of reading instruction. In M. Binkley, K. Rust, & M. Winglee (Eds.), *Methodological Issues in comparative educational studies: The case of the IEA reading literacy study* (pp. 193–221). Washington, DC: U.S. Department of Education, National Center for Education Statistics.

Binkley, M. R., & Trevor, M. (1996). *Reading literacy in the United States: Findings from the IEA reading literacy study* (NCES 96-258). Washington, DC: U.S. Department of Education, National Center for Education Statistics.

Bleichrodt, J. P., & Drenth, P. J. D. (1990). *Contemporary issues in cross-cultural psychology.* Amsterdam, The Netherlands: Swets & Zeitlinger.

Bos, W., & Lehmann, R. H. (Eds.). (1995). *Reflections on educational achievement: Papers in honour of T. Neville Postlethwaite.* Münster, Germany: Waxmann.

Bracey, G. W. (1996). International comparisons and the condition of American education. *Educational Researcher, 25*(1), 5–11.

Brislin, R. W. (1976). *Translation: Applications and research.* New York: Goudner Press.

Brislin, R. W. (1986). The wording and translation of research instruments. In W. J. Lonner & J. W. Berry (Eds.), *Field methods in cross-cultural research* (pp. 137-164). California: Sage.

Cole, N. (1977). *An ethnographic psychology of cognition.* In P. N. Johnson-Laird & P. C. Wason (Eds.), *Thinking* (pp. 340–356). London: Cambridge University Press.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-Experimentation: Design & analysis issues for field settings.* Boston: Houghton-Mifflin.

Darlington, R. B. (1973). Comparing two groups by simple graphs. *Psychological Bulletin, 79,* 110–116.

Davidson, M. L., & Sharma, A. R. (1988). Parametric statistics and levels of measurement. *Psychological Bulletin, 104,* 137–144.

Davies, I. R. L., & Gorbett, G. G. (1997). A cross-cultural study of colour grouping: Evidence for weak linguistic relativity. *British Journal of Psychology, 88,* 493–517.

Drenth, P. J. D., & Groenendijk, B. (1992). Organisatiepsychologie in cross-cultureel perspectief [Industrial psychology in cross-cultural perspective]. In P. J. D. Drenth, H. Thierry, & C. de Wolff (Eds.), *Nieuw handboek arbeids-en organisatie psychologie* [The new handbook of industrial psychology] (pp. 121–134). Deventer: Van Loghum Slaterus.

Easthope, G. (1974). *A history of social research methods.* London: Longmann.

Elder, J. W. (1973). Problems of cross-cultural methodology: Instrumentation and interviewing in India. In M. Armer & A. D. Grimshaw (Eds.), *Comparative social research: methodological problems and strategies* (pp. 234–250). New York: Wiley.

Elley, W. B. (1992). *How in the world do students read?* The Hague, The Netherlands: IEA.

Elley, W. B. (1994). *The IEA Study of Reading Literacy: Achievement and instruction in thirty-two school systems.* New York: Pergamon Press.

Elley, W. B. (1998). An insider's view of the IEA Reading Literacy Study. *Studies in Educational Evaluation, 24*(2), 127–136.

Embretson, S. E. (1983). Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin, 93,* 179–197.

Embretson, S. E. (Ed.). (1985). *Test design: Developments in psychology and psychometrics.* New York: Academic Press.

Embretson, S. E., & Wetzel, C. D. (1987). Component models for paragraph comprehension tests. *Applied Psychological Measurement, 11,* 175–193.

Farr, R., Pritchard, R., & Smitten, B. (1990). A description of what happens when an examinee takes a multiple-choice reading comprehension test. *Journal of Educational Measurement, 27,* 209–226.

Fischer, G. H. (1974). *Einfurung in die theorie psychologischer tests: Grundlagen und anwendungen* [Introduction to the theory of psychological test: Foundations and applications]. Bern, Switzerland: Verlag-Huber.

Fischer, G. H. (1995). The linear logistic test model. In G. H. Fischer & I. W. Molenaar (Eds.), *Rasch model: Foundations, recent developments, and applications* (pp. 131–156). Berlin: Springer-Verlag.

Fischer, G. H., & Molenaar, I. W. (Eds.). (1995). *Rasch model: Foundations, recent developments, and applications.* Berlin: Springer-Verlag.

Flier, H., van den. (1977). Environmental factors and deviant response patterns. In Y. H. Poortinga (Ed.), *Basic problems in cross-cultural psychology* (pp. 230–293). Amsterdam: Swets & Zeitlinger.

Flier, H., van den, & Drenth, P. J. D. (1980). Fair selection and comparability of test scores. In L. J. T. van der Kamp, W. F. Langerak, & D. N. M. de Gruijter (Eds.), *Psychometrics for educational debates* (pp. 34–38). New York: Wiley.

Flury, B., & Riedwyl, H. (1988). *Multivariate statistics: A practical approach.* London: Chapman & Hall.

Geisinger, K. F. (1994). Cross-cultural normative assessment: translation and adaptation issues influencing the normative interpretation of assessment instruments. *Psychological Assessment, 6,* 304–312.

Goldstein, H. (1993). *Interpreting international comparisons of student* *achievement*. Unpublished manuscript, UNESCO.

Hambleton, R. K. (1994). Guidelines for adapting educational and psychological tests: A progress report. *European Journal of Psychological Assessment, 19*, 229–244.

Hambleton, R. K., & Kanjee, A. (1994). Enhancing the validity of cross-cultural studies: Improvements in instrument translation methods. In T. Husén & T. N. Postlethwaite (Eds.), *International encyclopaedia of education* (2nd ed., pp. 1020–1024). Oxford, England: Pergamon Press.

Hambleton, R. K., & Kanjee, A. (1995). Increasing the validity of cross-cultural assessments: Use of improved methods for test adaptation. *European Journal of Psychological Assessment, 11*, 147–157.

Holly, S., & Hallett, A. H. (1989). *Optimal control, expectations and uncertainty*. Cambridge, England: Cambridge University Press.

Hulin, C. L., Drasgow, F., & Parsons, C. K. (1982). Applications of item response theory to analysis of attitude scale translations. *Journal of Applied Psychology, 67*, 818–825.

Jensen, A. R. (1969). How much can we boost IQ and scholastic achievement? *Harvard Educational Review, 39*, 162–170.

Jonckheere, A. R. (1954). A distribution-free k-sample test against ordered alternatives. *Biometrika, 41*, 133–138.

Kapinus, B. A., & Atash, N. (1995). Exploring the possibilities of contructed-response items. In M. Binkley, K. Rust, R. Winglee (Eds.), *Methodological issues in comparative educational studies: The case of the IEA Reading Literacy Study* (pp. 105–133). Washington, DC: U.S. Department of Education, National Center for Education Statistics.

Keeves, J. P. (1992). *Methodology and measurement in international educational surveys*. The Hague, The Netherlands: IEA.

Kendall, G. M., & Stuart, A. (1967). *The advanced theory of statistics, Vol. 2* (2nd ed.). London: Griffin.

Kish, L. (1994). Multipopulation survey designs: Five types with seven shared aspects. *International Statistics Review, 62*(2), 167–186.

Knight, G. P., Yun Tein, J., Shell, R. , & Roosa, M. (1992). The cross-ethnic equivalence of parenting and family interaction measures among Hispanic and Anglo-American families. *Child Development, 63*, 1392–1403.

Lambin, R. (1995). What can planners expect from international studies? In W. Bos, & R. H. Lehmann (Eds.), *Reflections on educational achievement: Papers in honour of T. Neville Postlethwaite* (pp. 230–246). Münster, Germany: Waxmann.

Lehmann, E. L. (1955). Ordered families of distributions. *Annals of Mathematical Statistics, 26*, 399–419.

Little, R. J. A. (1982). Models for nonresponse in sample surveys. *Journal of the American Statistical Association, 77*, 237–250.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Melbourne, Australia: Addison Wesley.

Lundberg, I., & Hastedt, D. (1995). Development of international constructs. In R. M. Wolf (Ed.), *IEA Reading Literacy Study: Technical report* (chap. 7). The Hague, The Netherlands: IEA.

Lundberg, I., & Linnakylä, P. (1993). *Teaching reading around the world*. Hamburg, Germany: IEA.

Maris, E. (1995). Psychometric latent response models. *Psychometrika, 60*(4), 523–547.

Marsh, H. W., & Grayson, D. (1994). Longitudinal stability of latent means and individual differences: A unified approach. *Structural Equation Modeling, 1*, 317–359.

Maxwell, S. E., & Delaney, H. D. (1985). Measurement and statistics: An examination of construct validity. *Psychological Bulletin, 97*, 85–93.

McCauley, D. E., & Colberg, M. (1983). Transportability of deductive measurement across cultures. *Journal of Educational Measurement, 20*, 81–92.

McGraw, K. O., & Wong, S. P. (1992). A common language effect size measure. *Psychological Bulletin, 111*, 361–365.

Mehan, H. (1973). Assessing children's language using abilities: Methodological and cross-cultural implications. In M. Armer & A. D. Grimshaw (Eds.), *Comparative social research: Methodological problems and strategies* (pp. 341–350). New York: Wiley.

Messick, S. (1987). Large-scale educational assessment as policy research: Aspirations and limitations. *European Journal of Psychology and Education, 2*(2), 157–165.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed. , pp. 13–103). New York: Macmillan.

Mishra, S. N., Shah, A. K., & Lefante, J. J. (1986). Overlapping coefficients: The generalized t approach. *Communications in Statistics, 15*, 123–128.

Mulaik, S. A. (1972). *The foundations of factor analysis*. New York: McGraw-Hill.

Neale, M. C., Eaves, L. J., Kendler, K. S., Heath, A. C., & Kessler, R. C. (1994). Multiple regression with data collected from relatives: testing assumptions of the model. *Multivariate Behavioral Research, 29*, 33–61.

Novy, D. M., Nelson, D. V., Goodwin, J., & Rowzee, R. D. (1993). Psychometric comparability of the state-trait anxiety inventory for different ethnic subpopulations. *Psychological Assessment: A Journal of Consulting and Clinical Psychology, 5*, 343–349.

OECD. (1996). *Adult literacy in OECD countries: Technical report of the first international adult literacy survey*. Paris: Author.

OECD, & Human Resource Development Canada. (1997). *Literacy skills for the knowledge society*. Paris: OECD.

OECD, & Statistics Canada. (1995). *Literacy, economy and society: Results of the first international adult literacy survey*. Paris: OECD.

Poortinga, Y. H. (Ed.). (1976). *Basic problems in cross-cultural psychology*. Selected papers presented at the Third International Congress of the International Association for Cross-Cultural Psychology, Amsterdam: Swets & Zeitlinger.

Postlethwaite, T. N., & Ross, K. N. (1992). *Effective schools in reading. Implications for educational planners*. Amsterdam, The Netherlands: IEA.

Ragin, C. C. (1984). *The comparative method: Moving beyond qualitative and quantitative strategies*. Berkeley: University of California Press.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen, Denmark: Danish Institute for Educational Research.

Raudenbush, S. W. (1995). Hierarchical models: The case of school effects on literacy. In M. Binkley, K. Rust, R. Winglee (Eds.), *Methodological issues in comparative educational studies: The case of the IEA Reading Literacy Study* (pp. 231–241). Washington, DC: U.S. Department of Education, National Center for Education Statistics.

Robitaille, D. , & Garden, R. (1989). *The IEA Study of Mathematics II: Contexts and outcomes of school mathematics*. Oxford: Pergamon Press.

Rosenbaum, P. R. (1995). *Observational studies*. New York: Springer-Verlag.

Rust, K. (1995). Issues in sampling for international comparative studies in education: The case of the IEA reading literacy study. In M. Binkley, K. Rust, R. Winglee (Eds.), *Methodological issues in comparative educational studies: The case of the IEA Reading Literacy Study*. Washington, DC: U.S. Department of Education, National Center for Education Statistics.

Schleicher, A. (1995). Comparability issues in international education comparisons. In W. Bos & R. H. Lehman (Eds.), *Reflections on educational achievement: Papers in honour of T. Neville Postlethwaite* (pp. 532-540). Münster, Germany: Waxmann.

Sireci, S. G. (1997). Problems and issues in linking tests across languages. *Educational Measurement: Issues and Practices, 16*(1), 12–19.

Steyer, R., & Eid, M. (1993). *Messen und Testen* [Researching and testing]. Berlin: Springer.

Turban, D. B., Francis, D. J., Orburn, H. G., & Sanders, P. (1989). Construct equivalence as an approach to replacing validated cognitive ability selection tests. *Journal of Applied Psychology, 74,* 62–71.

U.S. Department of Education. (1995). *Methodological issues in comparative educational studies: the case of the IEA Reading Literacy Study*. Washington, DC: Author, National Center for Education Statistics.

Wadsworth, S. J., DeFries, J. C., Fulker, D. W., Olson, R. K., & Pennington, B. F. (1995). Reading performance and verbal short-term memory: A twin study of reciprocal causation. *Intelligence, 20*, 145–167.

Wagemaker, H., Taube, K., Munck, I., Kontogiannopoulou-Polydorides, G., & Martin, M. (1996). *Are girls better readers? Gender differences in reading literacy in 32 countries*. Delft, The Netherlands: Eburon.

Wainer, H., & Braun, H. I. (Eds.). (1988). *Test validity*. Hillsdale, NJ: Erlbaum.

Webb, E. J., Campbell, D .T., Schwartz, R. D., Sechrest, L. (1971). *Unobtrusive measures: Nonreactive research in the social sciences* (7th printing). Chicago: Rand McNally.

Wolf, R. M. (Ed.). (1995). *The IEA Reading Literacy Study: Technical report*. Washington, DC: IEA.

Zwick, R. (1993). How do scale properties shape our conclusions about achievement trends? Examples from the National Assessment of Educational Progress. In R. Steyer, K. F. Wender, & K. F. Widaman (Eds.), *Psychometric methodology*. Proceedings of the 7th European Meeting of the Psychometric Society in Trier (pp. 568–572). Stuttgart & New York: Verlag. ■