Pergamon

Studies in Educational Evaluation, Vol. 24, No. 2, pp. 99–125, 1998

Printed in Great Britain
0191-491X/98 $19.00 + 0.00

S0191-491X(98)00006-6

THE VALIDITY OF INTERNATIONAL SURVEYS OF READING LITERACY: THE CASE OF THE IEA READING LITERACY STUDY

Timo M. Bechger*, Erik van Schooten**
C. De Glopper, and Joop J. J. Hox****

**National Institute for Educational Measurement (CITO)*
***Faculty of Education, University of Amsterdam,*
The Netherlands

1. Introduction

In the last twenty years the interest in international comparative surveys of educational achievement has increased as can be seen from the number of studies and participating nations. As the number of participating countries increases, the range of cultural differences becomes larger, and it becomes increasingly difficult to obtain internationally comparable data. Especially when it is no longer possible to use the same questionnaire in each group. This article describes the requirements that measures of reading literacy must meet to be comparable across nations, and the conditions that affect the usefulness of international comparisons. Before discussing the methodological problems of comparative surveys in Section 4 and 5, we discuss the definition of reading literacy in Section 2, and the possible reasons why comparative educational research has become so popular in Section 3. In Section 6 to 8 we consider the conditions that affect the comparability of measurements of reading literacy in more detail. In the final sections we summarize our findings.

As an illustration we consider an international comparative study, which was recently conducted by the *International Association for the Evaluation of Educational Achievement* (IEA). The results of this study were reported in detail by Elley (1992, 1994), Postlethwaite and Ross, (1992), Lundberg and Linnakylä (1993), Wagemaker et al. (1996), the IEA (1995), and the US Department of Education (1995). We will refer in less detail to the *International Adult Literacy Study* (IALS), which was conducted in the autumn of 1994 by the *Organisation for Economic Co-operation and Development* (OECD). The results of this study were published by the OECD in 1995 and 1996.

2. The Definition of Reading Literacy

The IEA used the following definition of reading literacy[1]: "Reading Literacy is the ability to understand and use those written language forms that are required by society and/or valued by the individual" (Elley, 1994, p. 5). The category of language forms "required by society" refers to those kinds of literacy tasks which are needed to live in an organized society: reading notices, directions, maps, graphs, government circulaires, to name a few; the latter part of the definition allows for the inclusion of leisure reading (narrative prose or popular magazine articles) which may be valued by the individual readers, but are less important for survival in a society.

The definition suggests that reading literacy tests must be composed of literacy tasks that are sampled from a "universe" of literacy tasks that subjects in each group may encounter in their daily life. To cope with their limited resources, however, the IEA used short passages followed by multiple-choice questions to assess reading literacy. Thus, although both understanding and use are important according to the definition, the IEA study focused mainly on understanding (Elley, 1994, pp. 5-6).

The IEA 's definition requires a definition of "society". When multiple languages are spoken within the national borders there are persons whose home language is different from the official language. In this situation one has to decide which language is of interest, whether separate translations are made, or whether it is feasible to produce a translation that is suitable to all inhabitants (see Section 6).

3. The Importance of International Studies of Reading Literacy

Especially in the developed nations, governments and international organisations believe that a well educated and literate working population is vital for national social and economic development as well as for the personal well-being of individuals. This is why 1990 was declared "International literacy year" by UNESCO. Albert Tuijnman, the secretary of the OECD states it as follows:

> Literacy has moved to centre stages in the policy agenda because of a new phase of globalization bringing uncertainty and opportunities in terms of the use of labour as growth industries require high skills which changes the relationship between skills and job prospects and implies a growth in demand for literacy. At the same time there is a mismatch between this demand and the present supply of skills, exacerbated by the ageing of the population, with implications for social as well as economic objectives. (OECD, 1995, Chapter 1)

The desire to conduct international comparisons is motivated by the argument that strong educational systems can only be built if countries have timely information about how their students and educational system compare to those in other countries.

Several studies present direct estimates of the economic value of literacy. Bishop (1989), for instance, estimated that the poor reading and mathematical competencies of much of the US workforce will cost the US economy nearly 170 billion dollars each year by the year 2000. These figures led the US government to implement a large-scale plan to

improve the science achievement of American students (US Department of Education, 1991). Similar developments were found in the U.K. (Sarland, 1995, pp. 202-203). Kozol (1985) estimated the costs of illiteracy for the US to be $20 billion in 1985 and $237 billion in lifetime earnings forfeited by men aged 25-30 with less than high school attainment. The OECD (1995) established that lower levels of literacy are associated to unemployment and low income, and that growing sectors of western economies have the most literate workers.

The economic argument appears to prevail in governmental circles, as in the IEA (e.g., Robitaille, 1990).

4. Comparability Across Nations Versus Validity Within Nations

The IEA 's definition of reading literacy appears to allow cultural differences in societal - and individual preference for reading. However, a comparison across nations requires that at least some aspects of reading literacy are common to all societies involved. At the operational level, any two subjects with a score of, say, 10 are assumed to be equally literate. Furthermore, if one scores higher than the other he or she is also assumed to be more literate. Without these assumptions, measurements of literacy are *not comparable.*

In general, the need for comparable measurement conflicts with cultural relativism. First, a comparative survey is impossible unless reading literacy is common to every subject in the survey. Therefore, those who believe that literacy can only be discussed within the context of a particular culture will a priori reject any quantitative comparison (Au, 1995).[2] Second, measurements may not be suitable for comparison when the measurement procedure has been adapted to particular groups of subjects. The reason is that differences in reading literacy may then become confounded with differences in the manner in which examinees are asked to respond to the individual test items and the conditions under which these responses are elicited, i.e., the *measurement procedure* (Lord & Novick, 1968, p. 302).

In general, adaptation of the measurement procedure is necessary when a test that gives valid results in one group, is not valid in another group (Schwarz, 1963). In reading literacy research, further adaptation may be required since there are many aspects of written language that are *conventional*, and different across nations (Smith, 1988). Examples are the names of characters in stories, layout, etc. If conventions differ across nations, the tests should be adapted because understanding the conventions of the "language forms" in one's own society is an essential aspect of reading literacy. Using only conventions that are, say, common in western societies, would give an unfair advantage to these countries and render the measures beyond comparison (see Section 7.3). On the other hand, if researchers in each nation were free to develop their own measures, differences in reading literacy would become confounded with differences in measurement procedure. Hence, validity within nations and comparability across nations may be conflicting aims. The extent to which measurement is both comparable and valid within each group determines the feasibility of cross-national comparative research.

5. Valid and Artefactual Differences Between Groups

Valid differences result from differences in variables that are related to reading literacy. For example, some countries are wealthier than others which, among other things, implies that their schools are better equipped. This is likely to induce changes in the distribution of individual abilities that are related to component processes of reading literacy, such as meta-cognitive skills, and word recognition skills (Daneman, 1991). Indeed, with data from the IEA study, Raudenbush, Cheong, and Fotiu (1995) found a small, facilitating effect of gross national product on reading literacy.

Differences in reading literacy score would be *artefactual* if measurements were not comparable. We believe that the conditions that affect the comparability of measurement can be classified in two broad categories:

1. *Cross-national differences between measurement procedures.* When multiple-choice items are used the most likely sources of cross-national differences in measurement procedure are; (i) translation of the tests (items, instructions, response-format), (ii) the possibility that different groups of respondents respond differently to the measurement procedure, and (iii) differences in the behavior of examiners.

2. *Qualitative differences between subjects from different nations.* Next to causing people to respond differently to the testing, cultural differences are known to be associated with qualitative differences in cognitive abilities (Cole, 1977). By this we mean that the processes that underlie the responses to the reading literacy test, are different for subjects with different cultural backgrounds, or for subjects that speak different languages. Qualitative differences may arise from language differences, or as a consequence of development; when children pass through different stages of cognitive development, as they become more experienced with a subject, or due to instructional and curriculum variation, i.e., when children are taught to do things differently (Juel, 1991). As a consequence, items that are difficult for one group of children may be relatively easy for another group.

In the next sections we shall write in more detail about the conditions that affect the comparability of measurements of reading literacy; translation (Section 6), subjects that respond differently to similar items (Section 7), differences in the behavior of examiners (Section 8), and qualitative differences between subjects in different groups (Section 9).

6. Translation in Comparative Studies

6.1. Introduction

As we noted before, the translation of tests should be such that item responses remain comparable. This may provisionally be stated as a definition :

> In cross-national comparative studies a good translation: (a) does not change whatever the test is measuring in the original instrument, and (b) does not make the test more or less difficult.

This definition is not very precise. It refers to psychometric properties of the test and the definition does not give precise rules that a translator should follow to produce a good translation. The definition suggests that translators should know precisely what the test intends to measure. This is easier to explain when theoretical considerations guided the development of the original test and translators know how the items were written and what skills are required to give the right answer. In the IEA study, for instance, the items were classified according to the "transformations" which, it was hypothesized, the students used in responding (Elley, 1994, pp. 10-15; Wagemaker, Taube, Munck, Kontogiannopoulou, & Martin, 1996). For example, whether subjects would have to go beyond the information given and make inferences in arriving at the correct answer.

Even when translators are well informed about the intended meaning of the items, the translation itself may be difficult for the following reasons:

1. There may not be good dictionaries or grammars.
2. Apparently equivalent words may be altered due to different emotive meaning or cultural differences (e.g., Armer & Grimshaw, 1973; Mehan, 1973, pp. 324-326).
3. It may be difficult to equate levels of concreteness and abstractness in two different languages.
4. There may not be a meaningful common definition for some concepts and circumlocution may be complicated (e.g. Armer, 1973, pp. 54-55; Hofstede, 1980, p. 35).
5. Simple structures in one language may not be simple in another language.
6. When multiple languages are spoken within the national borders there are persons whose home language is different from the official language. One has to decide whether these languages are sufficiently similar.

In view of these difficulties, translators are likely to disagree and since there are usually only a few of them involved they might be an important source of bias.

Whether the translated test does indeed measure the same abilities as the original test can be judged subjectively or investigated by means of psychometric methods. The next paragraph deals with judgmental designs, while psychometric methods are discussed in Section 11. The difficulty of the test is the subject of Paragraph 6.4. Note that we deliberately consider the translation of the test, including questions, text, and the test instructions. The reason is that the items and the test instructions play an important role in the answering process and their translation requires as much attention as the translation of the passages they refer to (Farr, Pritchard, & Smitten, 1990; Kirsch, Jungeblut, & Campbell, 1992).

6.2. Judgmental Designs for Translating Instruments

The two most popular judgmental designs are forward translation and backward translation (Brislin, 1970, 1976, 1986; Campbell, Brislin, Steward, & Werner, 1971).

With *forward translation* a single translator, or preferably a group of translators, translates the test from the source language to a target language. Then, the equivalence of

the two versions of the instrument is judged by another group of translators. 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 judgements are made about their equivalence.

One of the main shortcomings of backtranslation is that the comparison of instruments is carried out in the source language. It is perfectly possible that the translation could be poor while the evidence on the comparability of the original instrument and the back-translated instrument would suggest otherwise. This might happen if the translators used a shared set of translation rules that ensured that the back translated instruments looked like the original instrument (Hambleton & Kanjee, 1994). There is also the danger of "translationalese", i.e., a target language style which is heavily influenced by the source language and consequently looks unfamiliar to native speakers (Jacobsen, 1988).

In general, the weakness of judgmental designs is the high level of inferencing that must be done by the translators. Translators, or expert judges, have often been found incapable of distinguishing biased items from unbiased items. Jensen (1980), and Angoff (1993) report numerous studies in which the items which were judged to be biased against a particular group were actually unbiased according to psychometric criteria and vice versa.

Another shortcoming of the aforementioned and other judgmental designs is that samples of the intended populations for the instruments never actually take the instrument under test-like conditions. This problem is made worse by the fact that the bilinguals used in the translation are likely to be cognitively different from monolinguals (e.g., Ervin, 1964; Hambleton, 1993, pp. 62-64; Jensen, 1980, pp. 605-607; Lambert, Havelka, & Crosby, 1958; Landar, Ervin, & Horowitz, 1960; Ricciardelli, 1992). For example, bilinguals are in general more capable than monolinguals and they do not necessarily interpret test items in the same way that monolinguals do.

6.3. Further Adaptations of the Test

Translators will often find it necessary to adapt the tests when the literally translated tests, including items, test instruction and written texts, do not follow the conventions in the target language, or when literal translation is impossible. Measurement units, street names, national events and names of persons, for instance, are usually adapted (e.g., IEA, 1995, Chapter 5, p. 11). Even societies speaking the "same" language are not identical in their conventions and the test may need to be adapted. African newspapers, for example, are visibly different from French newspapers even when the language of both is French.

In the IEA study, some of the participants were concerned that the preparation of test booklets by researchers in each country could give rise to differences in print size and layout that would be a source of artefactual differences between students in different

countries. Thus, the IEA advised on a standard type size and layout to be used where possible. In addition, the IEA had students complete the language test with different print sizes and found no differences between the test forms indicating that no artefactual differences were due to standardization of print size (IEA, 1995, pp. 16-17). Studies by Tinker (1966), Watts and Nisbett (1974), and Zacharisson (1965) likewise suggest that students can tolerate a wide diversity of print sizes, accommodating quickly to different types without problems. Modu and May (1977) investigated the difference between long and short passages and found very high correlations indicating that the length of the passages did not influence the scores.

6.4. Obtaining Tests of Similar Difficulty

6.4.1. *Introduction*

The difficulty of a test can be defined as the ability needed to achieve a certain score. It is related to the way the texts are constructed, and to the type of questions asked about the texts (Kirsch et al., 1992). The difficulty of texts has been studied in the context of readability studies. The difficulty of multiple-choice items has been investigated by Drum, Calfee and Cook (1981), Embretson and Wetzel (1987), and Kirsch and Mosenthal (1995a). From a practical point of view, these studies provide translators with characteristics of tests that they should try to keep as similar as possible. We now discuss these issues in more detail.

6.4.2. *The Readability of the Text*

The difficulty of a text is known as its "readability". *Readability* can be defined as the (reading) ability necessary to understand a given text (Staphorsius, 1994). The purpose of readability studies is to predict the readability of texts from their semantic and syntactic characteristics.

The predictor variables that have been used in readability studies can be classified as measures of the lexical and syntactic complexity of the text. The most familiar measures of *lexical complexity* are word length, the density of familiar or frequent words, and the proportion of different words. Indicators of *syntactic complexity* are the length of sentences, and the density of specific syntactic structures (i.e., the proportion of simple sentences). Following an extensive review of the literature, Staphorsius (1994) concludes that word length, the frequency of words, and sentence length have proven to be the most effective predictors in almost all investigations conducted so far. We suggest that these properties should be as similar as possible among translated versions of reading literacy tests.

6.4.3. *The Difficulty of Multiple-choice Paragraph Comprehension Items*

Kirsch and Mosenthal (1995a) used data from American students to investigate the effect of three variables on the difficulty of the items of the reading literacy test used by the IEA:

1. Type of information: The nature of information that readers must identify to complete a question or directive.

2. Type of match: The processes used to relate information in the question to information in the text to information in the choices. This refers to the degree of inference required to obtain the answer form the text.
3. Plausibility of distracting information: The degree of difficulty associated with selecting the correct answer from among a list of multiple-choice answers.

Kirsch and Mosenthal (1995a, Figure 6.1) found that, while readability showed no relation to item difficulty, the type of match, as well as the plausibility of distractors had a significant positive effect on item difficulty in the IEA reading study. According to them this suggests that the IEA reading literacy scales tended to be more of a measure of how well students were able to respond to multiple-choice questions than how well they were able to read and understand a wide range of texts (Kirsch & Mosenthal, 1995a, p. 178).

Anderson (1972) proposed a classification of multiple-choice paragraph comprehension items that is very similar to the type-of-match variable used by Kirsch and Mosenthal. In accordance with the findings by Kirsch and Mosenthal, Embretson and Wetzel (1987) found that Anderson's taxonomy predicted item difficulty in multiple-choice tests for reading comprehension. Unfortunately, these studies were limited to subjects in the United States and it is unknown if the results generalize to other languages or cultures.

7. Why Different Groups of People May Respond Differently to a Similar Measurement Procedure

7.1. Introduction

Even when the measurement procedure is physically identical, people from different nations may respond differently to the testing. Some may, for example, refuse to answer direct questions (e.g. Elder, 1973 pp. 126-127), or be differentially motivated to achieve a high score or give a "correct" answer. Another example was given in the previous section, where we noted that apparently equivalent words may have a different meaning to different subjects. As a consequence, test instructions, questions, or texts that contain these words may be understood differently.

In this section, we survey some of the conditions that are relevant to comparisons of reading literacy and discuss how their effect can be diminished. The influence of these conditions is complex and some of them are likely to interact. For example, speededness is found to increase the effect of testwiseness and anxiety (Guida, Ludlow, & Wilson, 1985; Kerstiens, 1990, pp. 7-8). Unfortunately, we known very little about these interactions. To assess the impact of the conditions mentioned in this section, one has to obtain measures of them. Where possible we will therefore discuss how such measures can be obtained.

7.2. Different Familiarity with the Response Format

Differential familiarity with particular item formats presents a source of incomparability. Persons with less experience with the selected response format, usually multiple-choice,

will be at a disadvantage. It is likely that experienced respondents are better at guessing, when the response format allows it, or at any other score-inflating test-taking strategies. That is, some are more "testwise" than others (Kerstiens, 1990).

To avoid non-comparability due to differential familiarity with the test, one may allow subjects the opportunity to practice and get acquainted with the response format before taking the actual test. In support of this recommendation there are several studies in which test performance has been improved by practice, especially of groups that were unfamiliar with formal testing (Feuerstein, 1972, 1979; Silvey, 1963). In some situations, a balance of item formats may be instrumental to ensure equal familiarity with the response format.

7.3. Differential Familiarity with the Conventions used in the Texts

Psychologists and linguists agree that texts are not simple collections of facts and other kinds of content. Instead, the content of different kinds of texts is organized and presented in distinctive and characteristic ways and these conventions differ across nations and across sub-cultures within nations (Smith, 1988, pp. 41-46).

Many aspects of texts are conventional; the internal structure of texts (e.g. story grammar), spelling, punctuation, capitalisation, paragraphing, and book binding are conventional. There are also conventions concerned with register. That is, one must choose and put the words together differently depending on the subject one is talking about, the person one is talking to, as well as the circumstances in which one is talking. It may, for example, be considered rude to ask direct questions and questions on unlikely events ("What would you do if you were in a war situation ?") might be considered senseless (cf. Elder, 1973). Conventions of register are used by the researcher, when he addresses himself to the respondents in the test instructions and in the questions, and also by the characters in the stories respondents are asked to read.

The bottom line is that these conventions are arbitrary and different across nations. If conventions differ, the texts should be adapted since knowing the language conventions in one's own society is an essential part of literacy. Using only conventions that are, say, typical of western societies, would give an unfair advantage to these countries and render the measures past comparison. This statement is upheld by numerous studies, starting with Bartlett (1932), that demonstrate that texts are better understood, remembered and used if readers (of all ages and abilities) are familiar with the relevant conventions. Lipson (1983), for example, found that subjects who read materials compatible with their religious backgrounds performed better on comprehension measures than subjects who read materials incompatible with their religious backgrounds.[3]

Spyridakis and Wenger (1991) discuss methods to measure the familiarity of members of the populations with the topics selected for the reading literacy test. The method they recommend is to ask representative samples from the involved populations to rank the texts on familiarity. Similar rankings are evidence of equal familiarity.

7.4. Different Interpretation of the Text, Items or Instructions

Their cultural background may lead subjects to interpret the reading material, as well as the items, and/or the test instructions in different ways. Examples are given by Mehan (1973) and Anderson et al. (1977). As a consequence, to some subjects the correct answer may not be the answer the test designers had in mind. It is recommended to use unambiguous texts, questions and instructions and to verify whether subjects have interpreted the (translated) text, items and instructions as intended. As we mentioned in Section 6.2, this is often part of the (forward) translation process.

7.5. Different Reactions to Time-limits

Spearman (1927) differentiated experimentally between two types of speed factors that enter into performance on timed tests (or enter into the total amount of time needed to complete an untimed test). This distinction was corroborated in later studies and supported by cognitive psychological theory (Jensen, 1986). One speed factor is intrinsic to all cognitive abilities and involves the speed of mental operations. It is reflected in the speed with which a person recalls relevant information for answering a question or for solving a problem, e.g., finding words in the mental lexicon. The other form of speed was linked by Spearman to a general attitude or preference for speed in performing any task. It might be called "personal tempo" to distinguish it from cognitive - or mental speed (Jensen, 1980 p. 612).

If speed of response affects the scores and high personal tempo is part of the cultural tradition in one group and not in another, the test may fail to span the cultural distance for that reason. Empirical evidence for this proposition is found in studies by Van Leest and Bleichrodt (1990), Schmidt and Crone (1991) and Van de Vijver and Poortinga (1991) who studied *test- or achievement motivation* which is similar to personal speed. Hence, while reading time may be a valid measure of reading proficiency (Smith, 1988, p.78), time-restrictions may render the measures beyond comparison.

There are two additional arguments against time restrictions: First, in each nation subjects should have the same time to finish the test and this may be difficult to organize. The IEA study illustrates this statement since apparently the Danish students were given less time to complete their test. The Danish National Institute for Educational Research decided to conduct its own data analysis and distinguish between a "reading comprehension score" based on the items that students had reached, and a "reading speed score based on the percentage of items reached (Mejding, 1994). Second, in general, we expect time restrictions to increase the effect of the measurement procedure on the response and thereby to increase the likelihood that different groups of respondents will respond differently to the measurement procedure. The effect of differential familiarity with aspects of the test, for example, may increase because subjects must increasingly rely on routine to get all items answered. Empirical evidence comes from a study by Guida et al. (1985) who found that less time on task increases the negative effect of test-anxiety on achievement.

7.6. Test Anxiety

Research examining the effects of test anxiety in situations involving evaluative stress, such as achievement testing, reveals that high test-anxious examinees do not perform as well as their less anxious counterparts (Saronson, 1980). With regard to reading comprehension, a negative relationship between test achievement and anxiety has been found for all kinds of readers (Bennett & Wark, 1980; Berrent, 1975; Everson, Millsap, & Browne, 1987).

Current test anxiety theory is based on a *cognitive interference model* which assumes that anxiety during testing interferes with the ability to retrieve and use previously learned information. Highly anxious students are believed to divide their attention between task demands and concerns of negative self-preoccupation, while low-anxious students are presumed to allocate a greater proportion of their attention to the task demands (Tobias, 1985; Saronson & Saronson, 1987). The cognitive inference model of test anxiety has been challenged with an alternative hypothesis, which assumes that the performance decrements observed in test anxious students are attributable directly to skill deficits. This may be called the *cognitive deficiency model.* (Brozo, 1984; Meijer, 1996).

In support of the cognitive interference model, Everson et al. (1987) found that test anxiety has a large effect on reading comprehension, independent of prior skill. Although experimental studies are needed to decide which model is correct, we feel confident to conclude that test anxiety should be avoided as much as possible.

7.7. Other Variables

Jensen (1980, pp. 615-617), and Anastasi (1964) further mention self-esteem, intrinsic interest in the test content, past habits of solving problems individually or collectively, and reflection-impulsivity as factors that may influence achievement scores and cause artefactual differences. When subjects from different nations are equally familiar with irrelevant aspects of the tests and there are no time restrictions these factors are not expected to have any influence on differences in literacy score. The reason is that past habits of problem solving, as well as interest in the test content are related to familiarity with aspects of the testing while personality factors such as carefulness, persistence and reflection-impulsivity describe the subjects' relative emphasis on speed versus correctness.

8. Differences due to the Behaviour of Examiners

There are many different ways in which examiners can introduce differences in test conditions. For example, in the recent *International Assessment of Educational Progress Study* (IAEP: Lapointe, Mead, & Askwe, 1992), all the randomly selected Korean students were made aware of the great honour of being chosen to represent their school and country, and thus had a responsibility to perform their best. For American students, on the other hand, participation on this international comparative study was just another activity. This may well have introduced a difference in achievement motivation between students

from the two countries. Examiners should be aware of these biases and the need to standardize the measurement conditions across nations.

Communication problems between examiners and examinees can prove to be a serious threat to the validity of results (Hambleton & Kanjee, 1994, p. 3). Examples of studies where problems of this kind occurred are given by Van de Vijver and Poortinga (1991). According to these authors, problems between examiners and examinees can be circumvented by ensuring that; (i) test administrators are familiar with the culture and the language of the examinees, and (ii) that the instructions of the test are clear and self-explanatory, with minimal reliance on verbal behavior.

9. Qualitative Differences in Reading Literacy

In Section 5 we noted that cultural differences may change the component processes that underlie the responses to a reading literacy, or reading comprehension test. We also noted that properties of the language, development or instruction may be responsible for these differences. Unfortunately, research in this area is scarce. As a consequence, it is difficult to make a clear statement with regard to the likelihood, the nature, or the causes of qualitative differences in reading literacy in different languages. We will briefly summarize what we did find.

First, it is well established that beginning readers rely more heavily on phonological recoding than advanced readers (McCusker, Hillinger, & Bias, 1981). The review by Juel (1991), however, shows that word recognition skills are only important in the first stages of reading development. In grades 2-3, most children become fluent readers and there are no theories of reading development beyond the fluent reading stage (Juel, 1991, p. 765).

Second, there is evidence that ideographic script is processed differently from alphabetic or syllabic script (Sasanuma & Fujimura, 1971, 1972). Briefly, phonological recoding plays no part in ideographic script while in alphabetic script it is used along with visually mediated reading. That was the reason why the IEA did not assess recoding skills (IEA, 1995, p. 5). It has also been suggested that the role of phonological recoding can be different from one language to another, depending on the regularity of the grapheme-phoneme correspondence (Staphorsius, 1994, p. 24), but this has not been investigated.

In spite of the uncertainty with regard to the origin and nature of process differences in reading literacy we believe that they can be detected. First of all, verbal protocols collected before and after (forward) translation may reveal that subjects use different strategies. Unfortunately, we did not find empirical evidence for this proposition. Think-aloud protocols have been collected by Farr et al. (1990) but, as far as we know, no "comparative protocol analyses" have ever been published. Second, psychometric methods may detect the problem. When subjects from different groups employ different strategies to answer the items we expect the rank of item difficulties, for instance, to be different across groups (Bechger, 1997).

10. The Validity of Multiple-Choice Paragraph Comprehension Items as Measures of Reading Literacy

In Section 4, we argued that a comparison has no meaning unless "reading literacy" is common to each subject in the populations that are involved in the survey. This means that the same theoretical mechanisms (strategies, information process and knowledge store) underlie task performance in each group (Embretson, 1983). Unfortunately, as we discussed in the previous section, in the case of reading comprehension or literacy the processes underlying the response variables are not well understood, and it is unclear what differences are induced by differences between languages or subjects.

Leaving aside possible qualitative differences between subjects in different groups, the validity of multiple-choice paragraph reading tasks as measures of reading literacy is questionable. This is the subject of this section.

The multiple-choice paragraph comprehension tests used in the IEA study have little face validity as measures of literacy. First, although both the understanding and the use of written language forms are important aspects of reading literacy, the IEA study focused only on understanding. Second, finding the answer to multiple-choice items is, in itself, one minor aspect of literacy. A different approach was taken by the *National Assessment of Educational Progress* (NAEP), conducted by the Educational Testing Service (ETS), and their example was followed by the OECD (OECD, 1995). In their 1985 survey of literacy skills of America's young adults, the ETS used items that actually simulated the diverse literacy demands of daily life (Kirsch & Jungeblut, 1986). For example, respondents were directed to fill in a deposit slip, determine eligibility from a table of employee benefits, fill out an order form from a catalogue, and follow a set of directions to travel from one location to another using a map.[4] The construct validity of these items has been investigated by Sheehan and Mislevy (1990) and found to be satisfactory.

The source of item format in reading comprehension tests was also debated within the IEA. The IEA researchers were particularly worried about the distinction between multiple-choice versus open-ended items. IEA associates reviewed the research in this area and an empirical study was conducted with 9 year old subjects (Elley & Mungubhai, 1992; IEA, 1995, Chapter 3). The conclusions confirmed that multiple-choice items did not rank subjects differently from open-ended questions. Furthermore, multiple choice items required less time, and students preferred the multiple-choice items. Similar conclusions were reached by Kapinus and Atash (1995). It would have been more relevant, however, to compare multiple-choice paragraph comprehension items to the simulated literacy tasks that were used by the ETS and the OECD.

11. Psychometric Methods to Test the Comparability of Measurement

The comparison of measurements in multiple groups has been given considerable attention in psychometrics, where comparability of measurements is referred to as measurement invariance with respect to group membership, Differential Item Functioning (DIF), or bias (Mellenbergh, 1994). That is, the informal operational criteria we discussed in the Section 4 hold when measures are unbiased at group level. Measurements are unbiased at group

level when groups that are matched on reading literacy obtain, on average, the same score. There is an extensive literature on bias, which has recently been reviewed by Millsap and Everon (1993), Holland and Wainer (1993). Various methods have been used to test the quality of a translation, as defined above (e.g. Angoff and Cook, 1988; Björginsson and Thompson, 1994; Ellis, 1989; Hulin, 1987; Thorndike, 1973), and there are numerous applications to reading comprehension tests.

When compared to qualitative arguments, the importance of psychometric methods in providing evidence for comparability of measurement should not be overestimated. First of all, bias is not a property of any statistical model but an interpretation. Second, the absence of psychometric bias can not be regarded as "proof" for comparability of measurement. Statistical models are based on auxiliary assumptions that may be very unrealistic. Third, the results may not generalize to other samples. Fourth, statistical power may be insufficient to reach a correct decision. The best way to use psychometric methods is in combination with systematic translation. Substantive arguments by translators will guide the statistical analysis, facilitate the interpretation of the results, and enhance the statistical power.

12. Investigating the Causes of Differences in Reading Literacy

If the international comparability and validity of measurements can be established, the next step is to investigate how differences in languages, policies and instructional practices (teacher and school characteristics) relate to the students' reading achievement.

It is now widely recognized that educational phenomena are defined at several levels (students, classes, schools), and that the analysis of multilevel phenomena requires a special set of statistical techniques called "multilevel analysis" (Creemers & Scheerens, 1994, Creemers, Stringfield, Scheerens, & Reynolds, 1992; Keeves, 1994; Muthén, 1989), especially when researchers are interested in estimates of the variability at different levels (Kreft, 1996). *Multilevel analysis* includes statistical techniques to model the variability at nation, school, and pupil level simultaneously, and to investigate associations of variables at one level with variables at the other levels (Bryk & Raudenbush, 1993; Goldstein, 1987, 1991; Hox, 1994). Applications of multilevel analysis to cross-national surveys are reported by Mason et al. (1983), Lockheed and Komenan (1989), and Scheerens and Creemers (1989), and by Raudenbush (1995), Raudenbush et al. (1995), Lundberg and Rósen (1995), Gustafsson (1997), and Bechger (1997) who apply multilevel techniques to data from the IEA reading literacy study.

Several authors (e.g. Creemers et al., 1992, p. 7; Bosker & Scheerens, 1994) have noted that effective application of multilevel models requires *multilevel theory*, "specifying which variables belong to which level, and which direct effects and cross-level interaction effects can be expected" (Hox, 1994, 1.2). Scheerens and Creemers (1989) developed general schemes for multilevel school effectiveness studies but detailed multilevel theories of reading literacy appear to be absent at present.

Analysis of associations within and across levels is dependent on the comparability of measurement. There may be differences in the dispersion of variables at different levels

and different sets of variables may be related to reading literacy in each nation. However, whenever we mention a variable in two or more countries we assume that the same construct is measured in each of these countries. Otherwise, the control achieved through the use of covariates may be illusory because these covariates have a different meaning in the two groups. If explanatory variables are not comparable this hinders statistical analysis of the data. In the IEA study, for example, Raudenbush et al. (1995, p. 272) used only four out of many predictors of students literacy due to the supposed incomparability of the other explanatory variables, and several countries were excluded because of "apparent irregularities in test administration " (ibid. p. 254).

13. On the IEA Reading Literacy Project

Frey (1970, pp. 244-245), Hambleton (1993), Goldstein (1993), Hambleton and Kanjee (1994), and many others argue in favour of active participation of all involved countries in the planning of the study because it increases the capacity of the organisation to detect cultural biases and take corrective action. The IEA followed this advice. Representatives from each of the participating nations were involved in each stage of the reading literacy study (IEA, 1995). The IEA reading literacy tests were developed collaboratively by the IEA international steering committee and National Research Coordinators (NRCs). The tests were pilot tested in all countries and the NRCs had the chance to identify problems for their students before the final selection of items took place.

The specific abilities the items were intended to measure were carefully defined and balanced across three "domains" of reading literacy: narrative prose, documents and expository prose (IEA, 1995, pp. 5-9; Wagemaker et al. 1996). This distinction was based on theoretical grounds, but some empirical evidence was also presented. First, differences in the reading literacy score between domains could be explained as the consequence of the special emphasis on certain types of reading material in a school system. Second, time spent on reading was found to differ across domains. Finally, factor analyses provided evidence for a single combined expository-narrative factor and a separate document factor (Gustafsson, 1997). Hence, the difference between documents and narrative or expository prose was well established.

The NRCs were responsible for the translation (Elley et al., 1994, p. 8; IEA, 1995, 3) but the IEA provided them with detailed guidelines in line with the recommendations in Section 6. Unfortunately, the actual translation was not documented in much detail. The number of translators, for instance, was not reported. Following the translation, pilot tests were conducted with samples from the intended populations to check the psychometric properties of the instruments, detect ambiguous and noncomparable items, and try-out the organization of the study.

The *dichotomous Rasch model* (Fischer & Molenaar, 1995) was employed in the pilot testing to detect biased items in the reading literacy test (IEA, 1995, pp. 14-15). If the Rasch model fits the data in all groups, its parameters should be the same across samples (Steyer & Eid, 1993). With this criterion accompanying the subjective judgement of the NRCs, the IEA was able to detect items that behave differently in different groups and construct a reading literacy score that is suitable for comparison (Elley et al., 1994, Table

1.8; IEA, 1995, Table 14). Following the pilot study, a large number of items (IEA, 1995, p. 30), which might have been useful in some countries, were dropped from the international tests. According to the IEA, this is "(..) the price to be paid for an international analysis" (IEA, 1995, 7).

For practical reasons the IEA required subjects to complete test booklets within 49 minutes (Wagemaker et al., 1996, Table 3 and 4). Afterwards, there was some debate whether these time limits were indeed implemented in the same way among the participating nations (Mejding, 1994). Whatever happened, studies by Lundberg and Rosen (1995), and Gustafsson (1997) show that these time-limits are likely to have had an effect on the rank order among nations for the narrative/expository domain.

Along with a cross-national comparison of achievement levels in literacy, the IEA intended to identify differences in policies and instructional practices in reading that cause differences in literacy. Although the comparability of the explanatory measures is just as important as the comparability of the dependent measures, the former received notably less attention. The validity of independent measures within nations was not further investigated.

The analyses published by the IEA appear to have been guided by general ideas about the relation between "input variables" and "output variables" at different levels of educational systems (IEA, 1995, Chapter 5). For example, the availability of books in the school is believed to enhance the reading ability of the students. A detailed theory of reading literacy was not presented. Developing this theory is a topic for future research.

14. Conclusion

Recent decades have seen a rapid increase in the number of multinational comparative studies launched both by official international agencies, such as the OECD, and by academic researchers. International comparisons are motivated by the argument that strong educational systems can only be built if countries have precise information about how their students and educational system compare to those in other countries. However, international comparative studies do not immediately lead to an improvement of the literacy of the population unless the measurements are comparable and the observed differences are interpreted correctly.

If measurements are not comparable the observed differences may not be due to differences in literacy and consequently be difficult to interpret. Generally, measurements become incomparable when measurement conditions are more favourable to some respondents than to others, or when different abilities underlie the response in different nations. To avoid artefactual differences we recommend the following:

1. Subjects should have had sufficient opportunity to get acquainted with the item-format.
2. Texts, questions and instructions should be unambiguous.
3. Passages should be equally familiar to all groups.
4. Time restrictions should be avoided as much as possible.

5. Test anxiety should be avoided as much as possible.
6. Test administrators should be familiar with the culture and the language of the examinees.
7. The instructions of the test should be clear and self-explanatory, with minimal reliance on verbal behavior.
8. Translators should be familiar with the intended use of the test and the information processing demands of the items.
9. The lexical and syntactic complexity of the material should be balanced among translations, as much as possible.
10. The tests should be adapted where necessary to the conventions of each population. Differences in layout, print size and passage length were found to have no influence on the responses.
11. Although cautious translation contributes to the quality of the translation, translators or other judges are not always capable of predicting the comparability of versions of an instrument and we recommend that psychometric methods are used in combination with judgmental methods.

Following the example of the OECD, we recommend that reading literacy be measured with items that actually simulate the diverse literacy demands of daily life. These activities should be sampled from a "universe" of literacy tasks that subjects in each group may encounter in their daily life. How to define this universe, and how to sample from it are topics for future research.

Throughout, the IEA reading literacy study was used to illustrate the text and we concluded our review with an evaluation of this study. We found that the IEA followed most of the recommendations that emerged from our study. The organization took great pains to establish comparable measurements of reading comprehension. Although the IEA study is among the best in the field, we believe that the study must be criticized for three reasons. First, the IEA study lacked theoretical underpinning, and the comparability of the explanatory variables received too little attention. Second, the IEA imposed time-limits on the testing which probably had an effect on the observed differences. Third, the IEA measured reading comprehension rather than reading literacy. This is not a serious point of critique since reading comprehension is also an interesting construct. We conclude that the IEA succeeded neither in assessing the relative level of reading literacy in each nation, nor in establishing why differences in reading literacy (or reading comprehension) occur.

Even with a comprehensive theory of reading literacy, there are good reasons to doubt whether valid inferences can be made as to what makes some school systems more effective than others. In general, cross-sectional studies do not allow "hard" inferences to be made with regard to the direction, or the strength of causal relations (e.g., Cook & Campbell, 1979; Daneman, 1991; Dogan, 1994, pp. 46-47; Rosier, 1994, p. 5855; Sobel, 1995). To improve our ability to make causal statements the first requirement is that the study be supported by educational theory. Secondly, the cross-sectional design may be improved by gathering longitudinal data or data on relatives, which permit stronger conclusions regarding the direction of causality (Neale et al., 1994; Wadsworth, DeFries, Fulker, Olson, & Pennington, 1995).

In the literature several other disadvantages of international surveys are mentioned, which also apply to the IEA study. First, the results of international studies are often useless to policy makers because the independent variables cannot be manipulated (e.g. gender or gross national product: Lambin, 1995). Second, while process variables at the classroom level are believed to be important predictors of differences between students, large scale surveys are not effective in measuring process variables (Creemers et al., 1992, p. 17). The reason is that surveys use paper-and-pencil tests, which are not likely to provide valid measurements of the actual behavior of the teacher. A potential disadvantage, finally, is that international studies might support developments towards the international standardization or globalization of curricula at the expense of the unique properties of national educational systems (Vedder, 1994).

Quantitative international surveys, however, do have a number of advantages. First, these studies generate international data sets, which are suitable for secondary analysis. The availability of cross-national data makes it possible to conduct policy-relevant research that cannot be conducted with data on a single society (Nowak, 1989, pp. 37-38; Schleicher, 1995, p. 217). For example, the study of the relationship between the length of the school year and educational achievement requires cross-national data. As reading literacy theory develops, existing data may be used to investigate new hypotheses. Similarly, the development of more advanced techniques of statistical modelling may give rise to new insights in existing data sets. A good example is the recent development of multilevel analysis. Second, the availability of well-developed questionnaires may facilitate future research (Kotte & Witt, 1995). Third, international comparative studies have proven to be an effective means to acquire government funding for further study (e.g. Bracey, 1996). Finally, cross-national studies may be instrumental in encouraging the development of professional staff in developing countries (Vedder, 1994).

Notes

1. The general problem of defining literacy is discussed in Venezky (1990). Definitions similar to the IEA's are used by the ETS (Kirsch et al. , 1986, 1991), Statistics Canada (1989), and the OECD (1992; 1995, p. 14;1996 Chapter 7).

2. The belief that human properties are universal and cross-cultural comparisons are conceivable is often referred to as the "etic" approach in contrast to the "emic" point of view, which rejects cross-cultural comparison.

3. References of studies demonstrating that texts are better understood, remembered and used if readers (of many ages and abilities) are familiar with the relevant conventions: Daneman, 1991, pp. 530-532; Freebody and Anderson, 1983; Goodman, 1982; Heath, 1983; Kintsch and Greene, 1978; Lipson, 1983; Mandler, 1987; Poissant, 1990; Smith, 1988; Steffenson et al., 1979; Thorndyke, 1977.

4. The tests were published by Prentice Hall: the Educational Testing Service.

References

Anastasi, A. (1964). Culture fair testing. *Educational Horizons, 43,* 26-30.

Anderson, R.C. (1972). How to construct achievement tests to assess comprehension. *Review of Educational Research, 42,* 145-170.

Angoff, W.H. (1993). Differential item functioning methodology. In H. Holland & P.W. Wainer (Eds.). *Differential item functioning.* New Jersey: Erlbaum.

Angoff, W.H., & Cook, L.L. (1988). *Equating the scores of the Prueba de Aptitud Academica and the Scholastic Aptitude Test.* New-York: College Entrance Examination Board.

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

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

Au, K.H. (1995). Multicultural perspectives on literacy research. *Journal of Reading Behavior, 27,* 85-100.

Bartlett, F. (1932). *Remembering: A study in experimental and social psychology.* Great Britain: University Press.

Bast, J. (1995). *The development of individual differences in reading ability.* Unpublished dissertation. Amsterdam: Paedologisch Instituut.

Bechger, T.M. (1997). *Methodological aspects of educational comparison.* Amsterdam: TT-Publications.

Berrent, H.I. (1975). *The effect of anxiety on cloze measures of reading comprehension for third and fifth grade average readers.* Unpublished doctoral dissertation. Hofstra University.

Binkley, M., Rust, K., & Winglee, M. (Eds.) (1995). *Methodological issues in comparative educational studies: The case of the IEA reading literacy study.* Washington, DC: National Centre for Educational Statistics.

Bishop, J.H. (1989). Is the test score decline responsible for the productivity growth decline? *American Economic Review, 79,* 178-197.

Björginsson, T., & Thompson, A.P. (1994). Psychometric properties of the Icelandic translation of the basic personality inventory: Cross-cultural invariance of a three factor solution. *Personality and Individual Differences, 16,* 47-56.

Bosker, R.J., & Scheerens, J. (1994). Alternative models of school effectiveness put to the test. *International Journal of Educational Research, 21,* 159-180.

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

Brislin, R.W. (1970). Back-translation for cross-cultural research. *Journal of Cross-cultural Psychology, 1,* 185-216.

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.

Brozo, W.G. (1984). *Pre-questions for prose learning with reading-anxious college students.* ERIC Document ED257633.

Bryk, A.S., & Raudenbush, S.W. (1993). *Hierarchical linear models: Applications and data analysis methods.* Newbury Park: Sage.

Campbell, D.T., Brislin, R.W., Steward, V.M., & Werner, O. (1971). Back-translation and other translation techniques for cross-cultural research. *International Journal of Psychology.*

Cole, N. (1977). An ethnographic psychology of cognition. Chapter 28. In P.N. Johnson-Laird & P. C. Wason (Eds.), *Thinking.* Cambridge: Cambridge University Press.

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

Creemers, B.P.M., Stringfield, S., Scheerens, J., & Reynolds, D. (1992). *National and international school-effectiveness research in retrospect and prospect.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco April 1992.

Creemers, B.P.M., & Scheerens, J. (1994). Developments in the educational effectiveness research programme. *International Journal of Educational Research, 21,* 125-140.

Daneman, M. (1991) Individual differences in reading skill. In R. Barr, M.L. Kamil, P. Mosenthal, & P.D. Pearson, (Eds.), *Handbook of reading research.* London: Longman.

Dogan, M. (1994). Use and misuse of statistics in comparative research. In M. Dogan & A. Kazancigil, A. (Eds.), *Comparing nations.* Oxford: Blackwell.

Drum, P., Calfee, R., & Cook, L. (1981). The effect of surface structure on performance in reading comprehension tests. *Reading Research Quarterly, 16,* 486-514.

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.* New York: Wiley.

Ellis, B.B. (1989). Differential item function: Implications for test translation. *Journal of Applied Psychology, 74,* 912-921.

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

Elley, W.B. (1994). *The IEA study of reading literacy: Achievement and instruction in thirty-two school systems.* The Hague: Pergamon.

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.

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

Ervin, S.M. (1964). Language and TAT content in bilinguals. *Journal of Abnormal and Social Psychology, 68,* 500-507.

Everson, H., Millsap, R.E., & Browne, J. (1987). *Test anxiety and skill deficits: A test of two competing hypothesis.* ERIC document ED301593.

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.

Feuerstein, R. (1972). Cognitive assessment of the socioculturally deprived adolescent. In L.J. Cronbach & P.J.D. Drenth (Eds.), *Mental tests and cultural adaptation.* The Hague : Mouton.

Feuerstein, R. (1979).*The dynamic assessment of retarded performers.* Baltimore: University Park Press.

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

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

Flier, van der H. (1980).*Vergelijkbaarheid van individuale testprestaties* [Comparability of individual test results]. Lisse: Swets en Zeitlinger.

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

Foster, P., & Purves, A. (1991). Literacy and society with particular reference to the non-western world. In Barr, M.L. Kamil, P. Mosenthal, & P.D. Pearson (Eds.), *Handbook of reading research,* Vol. II. (pp. 46-67). White Plains, NY: Longman.

Freebody, P., & Anderson, R.C. (1983) Effects of vocabulary difficulty, text cohesion and schema availability on reading comprehension. *Reading Research Quarterly, 18,* 277-294.

Frey, F.W. (1970). Cross-cultural survey research in political science. In Robert T. Holt & J.E. Turner (Eds.), *The methodology of comparative research.* New York: Free Press.

Goldstein, H. (1987). *Multilevel models in educational and social research.* London: Oxford University Press.

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

Goody, J., & Watt, T. (1962). The consequences of literacy. *Comparative Studies in Sociology and History, 5,* 105-110. Graesser, A., Golding, J.M., & Long, D. (1991). Narrative representation and Comprehension. In R. Barr (Ed.), *Handbook of reading research,* Volume II. New York and London: Longman.

Guida, F.V., Ludlow, L.H., & Wilson, M. (1985). The mediating effect of time-on-task on the academic anxiety/achievement interaction: A structural model. *Journal of Research and Development in Education, 19,* 21-26.

Gustafsson, J-E. (1997). Measurement characteristics of the IEA reading literacy scales for 9-10 years-old at country and individual levels. *Journal of Educational Measurement, 34,* 34-38.

Hambleton, R.K. (1993). Translating achievement tests for use in cross-national studies. *European Journal of Psychological Assessment, 9,* 57-68.

Hambleton, R.K., & Kanjee, A. (1994). Enhancing the validity of cross-cultural Studies: Improvements in instrument translation methods. In T. Husen & T.N. Postlethwaite (Eds.), *International encyclopedia of education* (2nd ed.) Oxford: Pergamon.

Heath, S.B. (1983). *Ways with words: Language, life and work in communities and classrooms.* Cambridge: Cambridge University Press.

Hofstede, G. (1980). *Culture's consequences.* London: Sage.

Holland, H., & Wainer, P.W. (1993). *Differential item functioning.* London: Erlbaum.

Hox, J.J. (1994). *Applied multilevel analysis.* Amsterdam: TT Publications.

Hulin, C.L. (1987). A psychometric theory of evaluations of items and scale translations. *Journal of Cross Cultural Psychology, 18,* 115-142.

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.

IEA (1995). *The IEA reading literacy study: Technical report.* Edited by R. M. Wolf. The Hague: IEA.

Jacobsen, A.L. (1988). *Lexical selection and creation in translation.* Stockholm: NAES Proceedings, 3.

Jensen, A.R. (1980). *Bias in mental testing.* London: Methuen.

Just, M.A., & Carpenter, P.A. (1987). *The psychology of reading and language comprehension.* Boston: Allyn & Bacon.

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: The case of the IEA reading literacy study.* Washington, DC: National Centre for Educational Statistics.

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

Keeves, J.P. (1995).The contribution of IEA research to Australian education. In W. Bos & R. H. Lehmann (Eds.), *Reflections on educational achievement: Papers in honour of T. Neville Postlethwaite.* Munster: Waxmann.

Keeves, J.P., & Adams, D. (1994). Comparative methodology in education. In T. Husen & N.T. Postlethwaite (Eds.), *The international encyclopedia of education* (2nd ed.). Amsterdam: Elsevier.

Kerstiens, G. (1990). A slow look at speeded reading comprehension tests. *Review of Research in Developmental Education, 7,* 3-14.

Kintsch, W., & Greene, E. (1978) The role of culture-specific schemata in the comprehension and recall of stories. *Discourse Processes, 1,* 1-13.

Kirsch, I., & Jungeblut, A. (1986). Literacy: profiles of America's young adults, National Assessment of Educational Progress (NAEP). Princeton: Educational Testing Service.

Kirsch, I.S., & Mosenthal, P.B. (1995a). Interpreting the IEA reading literacy scales. In M. Binkley, K. Rust & M. Winglee (Eds.), *Methodological issues in comparative educational studies: The case of the IEA reading literacy study.* Washington, DC: National Centre for Educational Statistics.

Kirsch, I.S., & Mosenthal, P.B. (1995b). Literacy performance on three scales: Defintions and results. In: *Literacy, Economy and society: Results of the first international adult literacy survey.* Paris: OECD.

Kirsch, I., Jungeblut, A., & Campbell, A. (1992). *Beyond the school doors. The literacy needs of job seekers served by the US Department of Labor.* Princeton, NJ: Educational Testing Service.

Kotte, D., & Witt, R. (1995) Change and challenge: Assessing economic literacy. In W. Bos & R.H. Lehmann (Eds.), *Reflections on educational achievement: Papers in honour of T. Neville Postlethwaite.* Munster: Waxmann.

Kozol, J. (1985). *Illiterate America.* New Jersey: Anchor Press.

Lambert, W.E., Havelka, J., & Crosby, G. (1958). The influence of language-acquisition contexts on bilingualism. *Journal of Abnormal and Social Psychology, 56,* 239-244.

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.* Munster: Waxmann.

Landar, J.J., Ervin, S.M., & Horowitz, A.E. (1960). Navaho color categories. *Language, 36,* 368-382.

Lapointe, A.E., Mead, N.A., & Phillips, G.W. (1989). *A world of differences.* Princeton: Educational Testing Service.

Lipson, M.Y. (1983). The influence of religious affiliation on children's memory for text information. *Reading Research Quarterly,* summer 1983.

Lockheed, M.E., & Komenan, A. (1989). *Teaching quality and student achievement in Africa: The case of Nigeria and Swaziland.* Oxford: Pergamon.

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

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

Lundberg, I., & Rósen, M. (1995). *Two-level structural modeling of reading achievement as a basis for evaluating teachers effects.* Paper presented at the AERA conference in San Francisco, April 1995.

Hambleton, R.K. (1992). *Translation achievement tests for use in cross-national studies.* Vancouver: IEA International Coordinating Centre.

Mandler, J.M. (1987). On the psychological reality of story structure. *Discourse Processes, 10,* 1-29.

McClure, E., Mason, J., & Gordon, C. (1979). Sociocultural variables in children's sequencing of stories. *Discourse Processes, 6,* 131-143.

McCusker, L.X., Hillinger, M., & Bias, R.G. (1981). Phonological recoding and reading. *Psychological Bulletin, 89,* 217-245.

Mehan, H. (1973). Assessing children's language using abilities: Methodological and cross-cultural implications, Chapter 11. In M. Armer & A.D. Grimshaw (Eds.), *Comparative social research: methodological problems and strategies.* New York: Wiley.

Mejding, J. (1994). *Den grimme aelling og svanerne? - om danske elevers laesefaerdigheder.* Copenhagen: Danmarks Paedagogiske Institut.

Meijer, J. (1996). *Learning potential and fear of failure.* Amsterdam: Guust Bauer.

Millsap, R.E., & Everon, H. (1991). Methodology review: Statistical approaches for assessing measurement bias. *Applied Psychological Measurement, 17,* 297-334.

Modu, C.C., & May, G. (1977). *Exploring the use of shorter reading comprehension passages in the SAT-V.* Princeton, NJ: Educational Testing Service, Report SR-77- 33.

OECD (1992). *Adult literacy and economic performance.* Paris: OECD.

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

OECD (1996). Adult literacy in OECD countries: Technical report of the first international adult literacy study, edited by Irwin Kirsch and Scott Murray. Paris: OECD.

Olmedo, E.L. (1981). Testing linguistic minorities. *American Psychologist, 36,* 1078-1085.

Poissant, H. (1990). Linguistic and cultural factors in text comprehension. Paper presented at the world congress on reading, Stockholm, Sweden, July 3-6; available as ERIC document 340010.

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

Raudenbush, S.W. (1995). Hierarchical models: The case of school effects on literacy, Chapter 8. In M. Binkley, K. Rust & M. Winglee (Eds.), *Methodological issues in comparative educational studies: The case of the IEA reading literacy study.* Washington, DC: National Centre for Educational Statistics.

Raudenbush, S.W., Cheong, Y.F., & Fotiu, R.P. (1995). Synthesizing cross-national classroom effects data: Alternative models and methods. In M. Binkley, K. Rust & M. Winglee (Eds.), *Methodological issues in comparative educational studies: The case of the IEA reading literacy study.* Washington, DC: National Centre for Educational Statistics.

Riciardelli, L.A. (1993). An investigation of the cognitive development of Italian-English bilinguals and Italian monolinguals frome Rome. *Journal of Multilingual and Multicultural Development, 14,* 345-346.

Robitaille, D.F. (1990). *The third international mathematics and science study.* The Hague: IEA.

Rosier, I. (1994). Survey research methods. In T. Husen & N.T. Postlethwaite (Eds.), *The international encyclopaedia of education* (2nd ed.) Amsterdam: Elsevier.

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 & M. Winglee (Eds.), *Methodological issues in comparative educational studies: The case of the IEA reading literacy study.* Washington, DC: National Centre for Educational Statistics.

Sarland, C. (1995). Versions of literacy ? Re-thinking reading research. *Cambridge Journal of Education, 25,* 201-212.

Saronson, I.G., & Saronson, B.R. (1987). Cognitive interference as a component of anxiety: Measurement of its state and trait aspects. In H van der Ploeg, R. Schwarzer, & C.D. Speilberger (Eds.), *Advances in test anxiety research,* Vol. 5. Lisse, The Netherlands: Swets & Zeitlinger.

Sasanuma, S., & Fujimura, O. (1971). Selective impairment of phonetic and non-phonetic transcription of words in Japanese aphasic patients: Kana vs. Kanji in visual recognition and writing. *Cortex, 7,* 1-8.

Sasanuma, S., & Fujimura, O. (1972). An analysis of writing errors in Japanese aphasic patients: Kanji vs. Kana words. *Cortex, 8,* 265-282.

Scheerens, J., & Creemers, B.P.M. (1989). Conceptualising school effectiveness. *International Journal of Educational Research, 7* (13), 691-707.

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.* Munster: Waxmann.

Schmitt, A.P., & Crone, C.R. (1991). *Alternative mathematical aptitude item types: DIF issues.* Research report 91-42. Princeton, NJ: Educational Testing Service.

Schwarz, P.A. (1963). Adapting tests to the cultural setting. *Educational and Psychological Measurement, 23,* 672-686.

Silvey, J. (1963). Aptitude testing and educational selection in Africa. *Rhodes Livingstone Journal, 34,* 9-22.

Smith, F. (1988). *Understanding reading: A psycholinguistic analysis of reading and learning to read.* Hillsdale, NJ: Erlbaum.

Spyridakis, J.H., & Wenger, M.J. (1991). An empirical method of assessing topic familiarity in reading comprehension research. *British Educational Research Journal, 17,* 353-360.

Steffenson, M.S., Joag-Dev, C., & Anderson, R.C. (1979). A cross-cultural perspective on reading comprehension. *Reading Research Quarterly, 15,* 10-29.

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

Thorndike, R.L. (1973). *Reading comprehension education in 15 countries.* Uppsala: Almquist & Wiksell.

Thorndyke, P.W. (1977). Cognitive structures in comprehension and memory of narrative discourse. *Cognitive Psychology, 9,* 77-110.

Tinker, M.A. (1966). Experimental studies on the legibility of print: An annotated biography. *Reading Research Quarterly, 1,* 67-118.

Tobias, S. (1985). Test anxiety: cognitive inference, defective skills, and cognitive capacity. *Educational Psychologist, 2,* 135-142.

UNESCO (1970). Mass-media in society - The need for research. *Reports and papers in mass-communication,* No. 59. Paris: UNESCO.

UNESCO (May 1990). *International Literacy Year: Year of opportunity.* Paris: UNESCO.

US Department of Education (1991). *America 2000: An education strategy.* Washington, DC.

Van de Vijver, F.J.R., & Poortinga, Y.H. (1991). Testing across cultures. In R.K. Hambleton & J. Zaal (Eds.), *Advances in educational and psychological testing* (pp. 277-308) Boston: Kluwer.

Van Leest, P.F., & Bleichrodt, N. (1990). Testing of college graduates from ethnic minority groups. In N. Bleichrodt & P.J.D. Drenth (Eds.), *Contemporary issues in cross-cultural psychology.* Amsterdam: Swets & Zeitlinger.

Vedder, P. (1994). Global measurement of the quality of education: A help to developing countries? *International Review of Education, 40,* 5-17.

Venezky, R.L. (1990). Definitions of literacy. In R.L Venezky, D.A Wagner, & B.S. Ciliberti (Eds.), *Towards defining literacy* (pp. 2-16). Newark, DE: International Reading Association.

Venezky, R.L. (1991). The development of literacy in the industrialized nations of the west. In M. Barr, L.Kamil, P. Mosenthal, & P.D. Pearson (Eds.), *Handbook of reading research, vol II,* (pp. 46-67). White Plains, NY: Longman.

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: Eburon.

Watts, L., & Nisbett, J. (1974). *Legibility of children's books.* Sussex: NFER.

Wolfe, R., & Wiley, D. (1992). *Third International Mathematics and Science Study: Sampling plan.* The Hague: IEA.

Zacharisson, B. (1965). *Studies in the legibility of printed text.* Stockholm: Almquist & Wiksell.

The Authors

TIMO BECHGER is currently employed as a post doc at the Dutch National Institute for Educational Testing (CITO). The present article is based on his doctoral thesis, which concerned the methodology of international educational comparisons. His research interests centre around applied statistics.

ERIK VAN SCHOOTEN is an Educational Researcher at the SCO-Kohnstamm institute forEducational Research of the University of Amsterdam. His major research interests are mothertongue and second and foreign language acquisition, reader response, reading attitudes and literary education.

JOOP HOX is Associate Professor of methods and statistics at the University of Amsterdam. His major research interests are multilevel modeling, structural equation models and metaanalysis.

KEES DE GLOPPER is Professor of Education at the Faculty of Educational Sciences of the University of Amsterdam. He was involved in the IEA reading literacy study as a nationalcoordinator. His major research interests are language learning and teaching and international comparative research.