Assessment in Education, Vol. 8, No. 2, 2001

International Adult Literacy Survey (IALS): an analysis of international comparisons of adult literacy

ALAIN BLUM[1], HARVEY GOLDSTEIN[2] & FRANCE GUÉRIN-PACE[1]
[1]*Institut National d'Études Demographiques, 125/133 bd. Davout, 75980 Paris, France, and* [2]*Institute of Education, University of London, 20 Bedford Way, London WC1H 0AL, UK*

ABSTRACT *The International Adult Literacy Survey (IALS) raises a number of important issues that are inherent in all attempts to make comparisons of cognitive and behavioural attributes across countries. This article discusses both the statistical and interpretational problems. A detailed analysis of the survey instruments is carried out to demonstrate the cultural specificity involved. The data modelling techniques used in IALS are critiqued and alternative analyses performed. The article argues for extreme caution in interpreting results in the light of the weaknesses of the survey.*

Introduction

The International Adult Literacy Survey (IALS) represents the collaboration of a number of countries that agreed to investigate co-operatively adult literacy on an international basis. The main findings are published in a report (OECD, 1997) and there is also a technical report (Murray *et al.*, 1998)

Five EU member countries (France, Germany, Ireland, the Netherlands and Sweden) took part in the first round of the IALS in 1994, as part of a larger programme of surveys, which included the USA, Canada, Poland and Switzerland. The UK and (Flemish) Belgium took part later in Spring 1996, together with Australia and New Zealand. Several other EU member countries joined in a second round in 1998.

A draft report of the results of the IALS in December 1995 revealed concerns about the comparability and reliability of the data, and the methodological and operational differences between the various countries. In particular, France withdrew from the reporting stage of the study and the European Commission instigated a study of the EU dimension of IALS. The present article uses results from that investigation that was managed by the Social Survey Division of the Office for National Statistics, London (Carey, 2000).

ISSN 0969-594X print/ISSN 1465-329X online/01/020225-22 ©2001 Taylor & Francis Ltd
DOI: 10.1080/0969594012006266

The ostensible aim of IALS was to provide a comparison of levels of 'prose', 'document' and 'quantitative' literacy among the countries involved using the same measuring instrument that would yield equivalent interpretations in the different cultures and different languages. Respondents, about 3000 in each country, were tested in their homes. Each participant responded to one booklet, which contained items of each literacy type, and there were seven different booklet versions that were rotated. Background information was collected on the respondents and features in some of the analyses. The results of the survey received wide publicity.

There have been several commentaries and critiques of IALS. Most of these (e.g. Street, 1996; Hamilton & Barton, 1999) are concerned with how literacy is measured and are critical of the relative lack of involvement of literacy specialists. These critiques take particular issue with the notion that there can be a valid common definition of literacy across cultures and maintain that it is only meaningful to contextualise measures of literacy within a culture. In the present article, we seek to complement these views by criticising the technical procedures and assumptions used in IALS and by presenting evidence from IALS itself that there are serious weaknesses due to translation problems, cultural specificity and inherent measurement problems. There are further weaknesses that have been identified in IALS, which are not the subject of this article, including sampling problems, scoring variability and response rates; these are discussed in the report of the European Commission-funded study (Carey, 2000).

We begin by looking at the procedures used in IALS to define literacy, including specifically the way in which test items were selected, and how 'scales' were constructed and reported on. We also consider some alternatives to the analyses actually used. We illustrate some of the technical issues raised through a detailed re-analysis of IALS data and then offer a more general discussion of translation problems with respect to measurement issues. There is a re-analysis of IALS data at the item level and an analysis of respondent motivation. Finally, we attempt to draw some conclusions about international comparative studies in general.

Defining the Domains of Literacy

From the outset IALS considered literacy measurement in three 'domains'—prose literacy, document literacy and quantitative literacy—the domains being based upon earlier US work. Scales were constructed and results are reported for each of these three 'measures'.

Three major US studies in the 1980s and 1990s (Kirsch & Murray, 1998) were used to produce the three domains. This was done in each case by Educational Testing Service (ETS) using 'item response models' (IRMs), which are referred to in the IALS reports as 'item response theory'.

For each domain different tasks are used. The analysis carried out by Rock in the Technical report (Murray *et al.*, 1998, Chapter 8) shows that there are high correlations (around 0.9) between the domain scores—each domain score being effectively the number of correct responses on the constituent items. The

justification for the use of three scales, rather than just one therefore seems rather weak. Section 8.3 of the report states that 'a strong general literacy factor was found in all 10 populations, (but) there was sufficient separation among the three literacy scales to justify reporting these scales separately'.

No attempt is made in IALS properly to explore the dimensionality of the complete set of tasks. (In the Statistical Appendix we give a brief formal description of what is meant by 'dimensionality' of a set of items.) There is a reliance on the original US studies, with little discussion of whether it is possible to assume that any results will apply to other populations. The three scales are treated quite separately, yet Chapter 7 discusses some of the reasons for expecting high correlations.

The implication of this is that underneath the chosen domains there may well be further dimensions along which people differ. It may be the case, for example, that such dimensions exist and are common to all three domains, and that these are responsible for the observed high intercorrelations. In future work, this is one area for research, using multi-dimensional item response models of sufficient complexity. The IRMs used in IALS are all uni-dimensional, i.e. allow no serious possibility for discovering an underlying dimensionality structure, other than by using global and non-specific 'goodness of fit' statistics.

Dimensionality

The upshot of the initial decision to use three separate domains is that these constrain the outcomes of the study. We can see this as follows.

Suppose that for a collection of tests or test items, a two-dimensional (factor) model was really underlying the observed responses [model (3) in the Appendix]. If a one-dimensional (uni-dimensional) model [for example model (1) in the Appendix] is fitted then, given a large enough sample, it will be found to be discrepant with the data. Typically, this will be detected by some tests or items 'not fitting'. This is what actually occurs in IALS and such 'discrepant' items tend to be removed. This then results in a model that better satisfies the model assumptions, in particular the assumption that there is only a single dimension. The problem is that the 'discrepant' items will often be just the ones that are expressing the existence of a second dimension. If, initially, only a minority of items are of this kind, then the remainder will dominate the model and determine what is finally left. We see therefore that, when a uni-dimensional model is assumed, initial decisions about which items to include and in what proportions, will determine the final scale. We shall return to this issue in more detail later.

The real problem here comes not just from the decisions by test constructors about what items to include in what tests or domains, but also in the subsequent fitting of possibly over-simplified models, which then lead to further selections and removals of items to conform to a particular set of model assumptions.

There are two consistent attitudes one can take towards scale construction. One is to decide what to include on largely substantive grounds, modified by piloting to ensure that the components of a test are properly understood and that items possess

a reasonable measure of discriminatory power. The final decision about how to combine items together in order to report proficiencies, or whatever, will then be taken with reference to the substance of what is being tested. The other is to allow the final decision to be made following an exploration of the dimensionality structure of data obtained from a large sample of respondents. In practice, of course, a mixture of these might be used. The problem with the IALS procedure is that it neither allows a proper exploration of the dimensionality of the data nor allows substantive decisions to be decisive. It should also be pointed out that procedures for exploring dimensionality have existed for some time (see, for example Bock *et al.*, 1988), yet the existence of these is ignored in the technical report.

Item Exclusion

According to Chapter 10 of the technical report, 12 of the 114 items originally trialled for IALS were dropped because they did not fit very well (see Statistical Appendix), involving a large discrepancy value in three or more countries. A further 46 items (Chapter 9.3) also did not fit equally well in all countries and for 14 of these (available in French and English versions) a detailed investigation was made to try to ascertain why. When the final scale was constructed, however, these 46 remained.

The conclusion of Chapter 9 is that the IALS framework is 'consistent across two languages and five cultures'. This is a curious statement since the detailed analysis of these 14 items reveals a number of reasons why they would be harder (that is have different parameter values associated with them) in some countries than others. It would seem sensible to carry out a detailed analysis of all items in this kind of way in order to ascertain where 'biases' may exist, rather than just the ones that do not fit the model

An item that does not 'fit' a particular uni-dimensional model is providing information that the model itself is inadequate to describe the item's responses. There may be several reasons for this. One reason, in an international study such as IALS, may be that translation has altered the characteristics of the item relative to other items for certain countries; a different translation process might allow the item to fit the model better. Of itself, however, this does not imply that the latter translation is better; a judgement of translation accuracy has to be made on other grounds. Another reason for a poor fit, as noted earlier, is that there are in reality two or more dimensions of literacy that the items are reflecting, and the lack of fit is simply indicating this. In particular there may be different dimensions and different numbers of dimensions in each country.

If, in fact, these discrepancies are indicating extra dimensions in the data, then removing some 'non-fitting' items and forcing all the remaining items to have the same parameter values for each country in a uni-dimensional model will tend to create 'biases' against those countries where discrepancies are largest.

The problem with scale construction techniques that rely upon strong dimensionality assumptions is that the composition of the resulting test instruments will be influenced by the population in which the piloting has been carried out. Thus, for

cultural, social or other reasons the intercorrelations among items, and hence the factor and dimensionality structure, may vary from population to population. IALS assumes that there is a common structure in all populations and this drives the construction of the scale and decisions as to which items to exclude. Furthermore, since it appears that the previous US studies were included in the scaling it seems that the US data may have dominated the scaling and weighted the scale to represent the US pattern more closely than that in any other country. In this way the use of existing instruments developed within a single country can be seen to lead to the possible introduction of subtle biases when applied to other cultures.

We are arguing, therefore, that a broader approach is needed towards the exploration of dimensionality. While we accept that for some purposes it may be necessary to summarise results in terms of a single score scale (for each proficiency) we believe that this should be done only on the basis of a detailed understanding of any underlying more complex dimensionality structure. Techniques are available for the full exploration of dimensionality and there seems to be no convincing case for omitting such analyses.

Scale Interpretations

In order to provide an indication of the 'meaning' to be attached to particular scores on each scale, the scale for each proficiency is divided in IALS into 5 levels (with 1 the lowest and 5 the highest). Within each level, tasks are identified such that there is an (approximately) 80% probability of a correct response from those individuals with proficiency scores at that level. A verbal description of these tasks, based upon a prior cognitive analysis of items, is used to typify that level.

This approach derives from the uni-dimensionality assumption. Since only a single attribute is (supposedly) being measured the resulting scale score summarises all the information about the attribute. It is therefore sufficient to characterise an individual. It follows that any verbal label attached to a scale score need only indicate the attributes that an individual with that score can be expected to exhibit. Thus, for all individuals with the same (one-dimensional) proficiency score, the relative difficulties of all the items is assumed to be the same.

However, if in fact some such individuals find item A more difficult than item B, and vice versa for other individuals, then there is no possibility of describing literacy levels consistently in the manner of IALS: individuals with very different patterns of responses could achieve the same score. Thus, the issue of dimensionality is crucial to the way in which scale scores can be interpreted. If there really are several underlying dimensions the existing descriptions provided by IALS will fail to capture the full diversity of performance by forcibly ranking everyone along a single scale.

The attempt to give 'meaning' to the IALS scale thus seems difficult to justify. Any score or level can be achieved by correct responses to a large number of different combinations of items and the choice of those items that individually have a high probability of success at each scale position is an over-simplification and may be very misleading. What is really required for interpretations of a scale, however it

may have been produced, is a description of the different combinations or patterns of tasks that can lead to any given scale position.

Alternatives

We now look at some of the alternative approaches to scaling and analysis that were ignored by IALS, but which, nevertheless, could produce useful insights and correct some of the restrictions of the IALS approach. Chapter 11.4 of the technical report presents a comparison of the scaled average proficiencies for each country compared to a simple scoring system consisting of the proportion of correct responses for each of the three proficiency sets of items. The country level correlations lie between 0.95 and 0.97 and essentially no inference is changed if one uses the simpler measure. This result is to be expected on theoretical grounds and, if one wishes to restrict attention to 1-dimensional models, there seems to be a strong case for using the proportion correct as a basis for country comparisons. The model underlying the use of the (possibly weighted) proportion correct, is in fact model (1) of the Statistical Appendix as opposed to model (2), and the whole IRM analysis could in principle be carried out based upon model (1), rather than model (2) (see Goldstein & Wood, 1989, for a further discussion). In fact, one might wish to argue for reporting the proportion correct simply on the grounds of this being a useful summary measure without any particular modelling justification.

Secondly, it would also be advantageous for a separate scaling to be done for each country. In this way differences can be seen and investigated directly. This will make the scaling procedure more 'transparent' and allow more substantively informed judgements to be made about country differences.

A third important approach is to see whether item groupings could be established for small groups of items that, on substantive grounds, were felt to constitute domains of interest. Experts in literacy with a wide variety of viewpoints and experiences could be used to suggest and discuss these and a mechanism developed for reaching consensus. These groupings would then describe 'literacy' at a more detailed level than the three proficiencies used in IALS, and for that reason have the potential for greater descriptive insights. If this were done, then for each such group or 'elementary item cluster' a (possibly weighted) proportion correct score could be obtained for each individual, and it would be these scores which would then represent the basic components of the study design. Each booklet would contain a subset of these clusters, using a similar allocation procedure to that in IALS. The analysis would then seek to estimate country means for each cluster, the variances and the correlations between them. Differences due to gender, education etc could readily be built into the multivariate response models used so that fully efficient estimates could be provided. Goldstein (1995, Chapter 4) describes the analysis of such a model. In addition, multilevel analysis could be performed so that variations between geographical areas can be estimated.

In addition to reporting at the cluster level, combinations of clusters could be formed to provide summary measures, but the main emphasis would be upon the detailed cluster level information. No scaling would need to be involved in this, save

perhaps to allow for different numbers of constituent items in each cluster if inter-cluster comparisons are required. This procedure would also have the considerable advantage of being relatively easy to understand for the non-technical reader. A serious disadvantage of the current IALS model-based procedures is their opaqueness and difficulty for those without a strong technical understanding.

In the main IALS report (OECD, 1997) and the technical report there is some attempt to carry out analyses of proficiency scores that introduce other individual measurements as covariates or predictors. There is little systematic attempt, however, to see the extent to which country differences can be explained by such factors. There appears to be a reluctance in the published IALS analyses to fit models which adjust for more than one, or at most two, factors at a time: this is a fourth approach for any future analyses.

For example, in Chapter 3 of the main report literacy scores are plotted against age with and without adjusting for level of education and separately by parents' years of education, but not in a combined analysis. Yet, the report (OECD, 1997, p. 71) warns that because of the marked relationship with age, comparisons should take account of the age distribution. (This remark is made in the context of comparisons between regions within countries, but applies equally to comparisons between countries). Indeed, since countries differ in their age distributions it could be argued that all comparisons should adjust for age. It would also appear that there are interactions with age, such that there seem to be fewer differences between countries for the older age groups.

It will be important, if in future multi-dimensional item response models are fitted, to incorporate factors such as age and education, into these models directly. Such a model, of the kind exemplified by (3) in the Statistical Appendix, could include such covariates. As Goldstein & Wood (1989) point out, it is quite possible that dimensions, which emerge from an analysis of a heterogeneous population, could be explained by such factors.

Fifthly, as we shall show later, IALS tasks can be classified according to their contextual characteristics, such as familiarity, repetitiveness, precision, etc. Such characteristics, at least in principle, can be applied to all tasks and therefore can be used in the analysis of task responses. Thus, for example, in comparing countries a measure of average familiarity could be used to adjust differences. More usefully, comparisons could be carried out at the task level to see how far country differences can be explained by such characteristics, also allowing for age etc as suggested above.

Finally there is no attempt in IALS to carry out multi-level analyses, which take account of differences between geographical areas, etc. These techniques are now in common use and it is well known that a failure to take proper account of multilevel structures can lead to misleading inferences, especially when carrying out analyses of relationships between scores and other factors.

We now look at a detailed re-analysis of IALS data to illustrate some of the technical points we have made. Having carried out the analysis described below and established a large number of problems in the data we did not consider it was worthwhile to invest further efforts exploring dimensionality on this dataset.

Comparing Literacy Between Countries: the case of France

The results of the IALS survey (Murray *et al*, 1998) suggest that three quarters of the French population have an ability level in terms of 'literacy' which prevents them from handling the normal matters of everyday life: reading a newspaper, writing a letter, understanding a short text, payslip, etc. Based on the scales proposed by the originators of the IALS survey, 75% of French adults have a low literacy level, estimated at 1 or 2, for comprehension of prose texts, whereas 52% of British, 49% of Dutch, 47% of Americans and 28% of Swedes are at this level. For comprehension of schematic texts the percentages at levels 1 and 2 are 63, 50, 42, 50 and 25%, respectively, and finally for comprehension of texts with a quantitative content the percentages are 57, 51, 34, 46 and 25%, respectively.

The percentages of people having a level 1 or 2 are high in France, but also surprisingly high in other countries. Being at level 1 supposedly means that you may just 'locate one piece of information in the text that is identical to or synonymous with the directive given in the instruction' (OECD, 1997, p. 16) and for level 2 that you may 'locate one or more pieces of information in the text, but several distractors may be present' (OECD, 1997). At the other end of the scale of ability, the percentage of people anywhere who are at level 5 is extremely low, so low in fact that it was not published separately: levels 4 and 5 were grouped in the same class in all the publications issued by IALS. A level 5 task 'requires the reader to search for information in a dense text that contains a number of plausible distracting elements'. In France, out of a sample of nearly 3000 people there are 11 people at level 5 for the prose texts, eight for the schematic texts and 16 for the questions with quantitative content, although 648 of interviewees were educated to a level higher than the *baccalauréat* (the upper secondary leaving certificate which also provides for entry to higher education). In Great Britain, 51 people are at level 5 for the prose texts out of a sample of over 6000 people. Sweden, which has the best results, has 121 people at this level for prose texts out of a sample of just over 2500 respondents.

The extent of the differences between countries on one hand, and the discrepancy between this and other data available for France, have led to considerable doubt about the validity of this survey and the international comparisons resulting from it. Because the methodologies and forms of definition of illiteracy are extremely diverse, it is difficult to compare the various assessments that exist. We merely note that according to the French national statistical office (INSEE), 5.4% of the adult population 'has at least one of the manifestations of illiteracy' and the definition given by INSEE corresponds in large part to the concept of literacy (Bodier & Chambas, 1996). According to a survey of conscripts, 8% of young people aged from 16 to 24 years have reading difficulties (Bentolila & Fort, 1994).

To understand these results, we put forward two hypotheses. The first one is that there is a lack of equivalence of the tasks in the different countries. More precisely, it suggests a change in the difficulty of items once they have passed through the translation filter. The second one is the possible effect on the measure of literacy of unequal motivation of interviewees faced with a survey of this type. The following section addresses translation issues in greater depth.

Translation Effects

The IALS survey had two main objectives. The first was:

> to develop scales that would permit comparisons of the literacy performance of adults with a wide range of abilities. Then, if such an assessment could be created, the second goal was to describe and compare the demonstrated literacy skills of adults in different countries. This second objective presented the challenge of comparing literacy across cultures and across languages. (Kirsch & Murray, 1998, p. 16)

Thus, the validity of the survey is based on a strong hypothesis of an identical difficulty scale of tasks among cultures and languages.

Translations of the questionnaires and documents from their original English-language version were done in each participant country and checked by Statistics-Canada. For instance, three versions of the questionnaire exist in French: Canadian, Swiss and French. Similarly, the British questionnaire differs from the English-speaking Canada one. The translation had to be both high quality and faithful to the original text. However, no precise accuracy criterion was defined and, as some authors have commented (Kalton *et al.*, 1998) the usual, and we believe essential, rule of back translation (new translation into English of the translated text and comparison with the original) was not followed.

If the main hypothesis of the survey is verified, i.e. that the questions are 'psychometrically equivalent' among social groups and linguistic groups, then the item success profile must be independent of the language of the questionnaire. We have examined whether the difficulty of the questions could actually be considered equivalent in each of the languages. A necessary (but not sufficient) condition of equivalence of the difficulty levels is compliance with the difficulty hierarchies. A question that is more difficult than another one in the original questionnaire must remain so in all the versions of the questionnaire. A simple way to test this hypothesis is to compare the *a priori* difficulty of items and their success in different countries. Each item is allocated a score *a priori* (on a scale defined from 0 to 500) and its difficulty can be then classified [1]. This score is calculated using a series of criteria, taking into account the complexity of the document to which the question refers and the complexity of the link between question and document (Kirsch *et al.*, 1998).

On the basis of the individual data and general results provided by the survey in 13 countries, [2] there are large differences between the actual and theoretical hierarchies. It suggests a discrepancy between the theoretical difficulty of the questions and the actual difficulty in a given country (Guérin-Pace & Blum, 1999).

Another procedure is to compare, for each item, the observed success rates in different countries. If the questions are presumed to be of equivalent difficulty for two countries, the graphic representation of the proportion of correct answers for each item will then approximate to a straight line with a gradient that is a function of the country literacy level. However, examination of the graphs shows clusters with high dispersion in most cases. The dispersion is especially important when

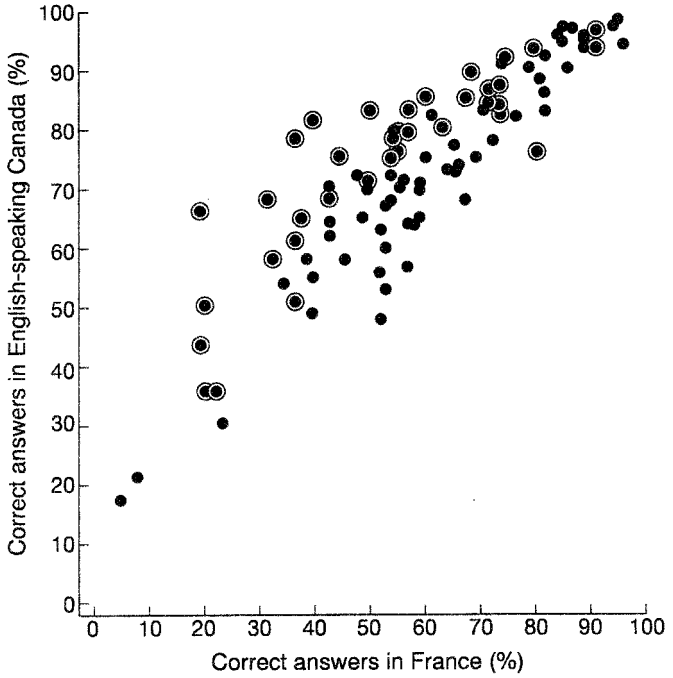

FIG. 1. 1994 IALS survey—Proportion of correct answers, French- and English-speaking Canada.

comparing France and Great Britain, for example. More generally, the dispersion is bigger when comparing two countries with different languages than two countries with the same languages [3].

From a comparison between the items' success rates in France and English-speaking Canada (Fig. 1) we have identified a set of questions with high variation, for which we have looked deeper and attempted to understand the reasons for large differences. We have compared the way these questions were formulated in French and English and found that there were significant differences in the translations. On a broader scale, we examined the whole questionnaire and the set of responses to establish the questions where the wording was not equivalent in French and in the Canadian one. We found 35 questions (circled in figure 1) where the translations differed and analysed these divergences under three headings.

Omission of Repetition of Terms

Repetition of a term in the question and in the accompanying text adjoining the answer is more frequent in the original English-language documents and forms the first source of bias. For example, in a question referring to the use of '*couches jetables*', the phrase containing the answer uses the term '*changes complets*'. In the Anglo-Canadian questionnaire the term 'disposable diapers' is repeated, as is the term 'disposable nappies', used in the British one. The respondent is drawn in English more easily towards the phrase containing this term and therefore to the

correct answer, whereas in French the reader has to understand that these terms are equivalent before being able to answer. The resultant bias considerably increases the difficulty of the questions in French.

Greater Precision of English Terms

As a general rule, the questions are drafted in a more precise form in English. For example, one question rendered in English by 'What is the most important thing to keep in mind?' is translated into French as *'Que doit on avoir à l'esprit?'* [What must be kept in mind?]. However the sentence containing the answer reads in English 'the most important thing' and in French *'la chose la plus importante'* [the most important thing]. The link is therefore easier to establish in English. Another task is defined in English as 'List all the rates'. It is translated into French as *'Quels taux'* [What rates], omitting to state 'all the rates'. This omission frequently led French interviewees to state only one rate instead of the list required for the answer to be considered correct.

Translation Errors

Some translation errors may be relatively unimportant from a strictly linguistic point of view but become important in relation to comprehension. The following example is particularly characteristic. A question rendered in French as *'soulignez la phrase indiquant ce que les Australiens ont fait pour …'* [underline the sentence indicating what the Australians did to …] is linked to a text worded: '*Une commission fut réunie en Australie'* [A commission was set up in Australia]. In English the question is 'What the Australians did to help decide …', the corresponding wording being 'The Australians set up a commission'. The answer is ambiguous in French because the place is given instead of the people.

These examples illustrate the problems associated with a test that originates in one language and then has to be translated into another.

Other Sources of Error

Sources of error that are less widespread, nevertheless, reveal the complexity of a definition of difficulty equivalence between items expressed in different languages. The example below is typical and can be interpreted in terms of 'cultural bias'. The task required is to work out, which are the comedies in a review covering four films. In two of these reviews, in both English and French, the term 'comedy' appears, which makes the question easy. In France, however, we find that many interviewees gave as their answer a third film, which from the description is obviously not a comedy. The only possible explanation is the presence in that film of the actor Michel Blanc, who is well known in France for his roles in many comedies, but is little known abroad. Here, association predominated in the answering process to the detriment of careful reading of the reviews.

The Importance of Translation Effects

The importance of the translation effect has been confirmed by a retest survey, conducted in 1998 as part of the EU-funded project (Carey, 2000). A sample of French individuals who were interviewed for IALS in 1994 were interviewed again in 1998. About 40% of the sample (300 respondents) were questioned with the original French questionnaire, while 60% (422 respondents) were interviewed using the French-Swiss questionnaire. Indeed, some of the problems found in the French questionnaire are not present in the Swiss version.

We have plotted the proportion of correct answers in 1998 from the Swiss questionnaire, against the proportion of correct answers in 1998 from the French questionnaire, for all items (Fig. 2). We have circled items where in the French questionnaire, but not the Swiss, the translation has created potential problems. Triangle items are those having both problems in the French and the Swiss questionnaire (same or different problems). Finally, squared items are those that we

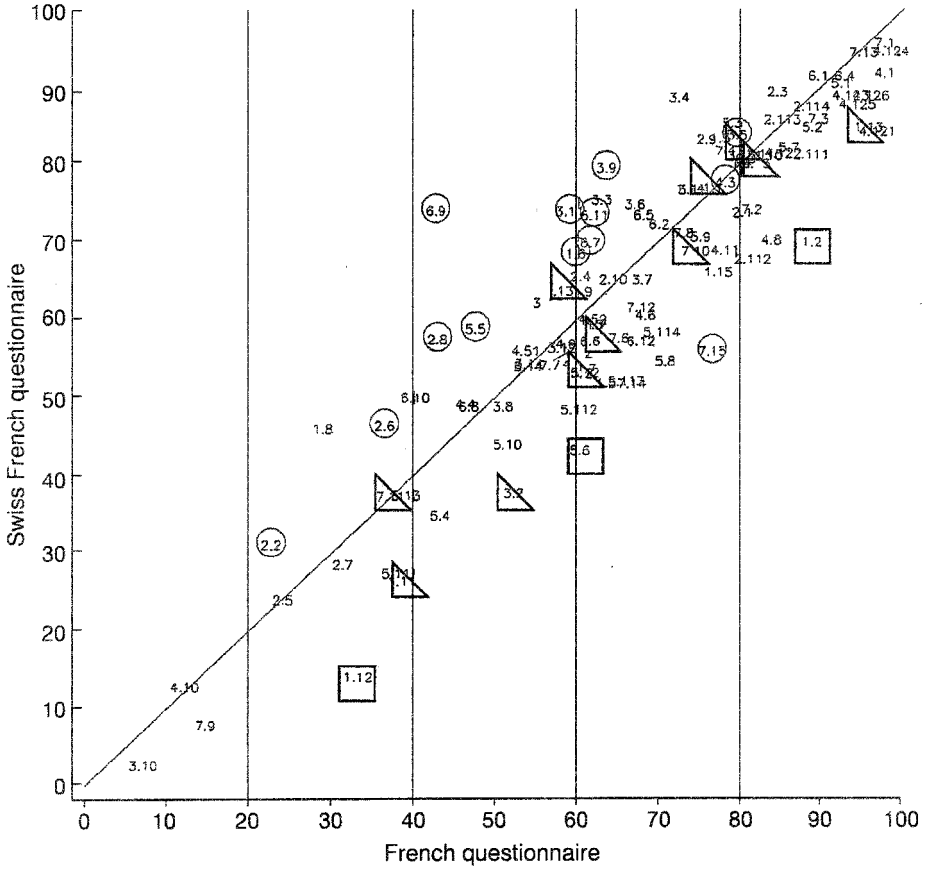

FIG. 2. 1998 retest in France, comparing proportions of correct answers for each item between the two samples (Swiss questionnaire or French questionnaire).

have identified as involving no translation problem in the French questionnaire, but which do in the Swiss questionnaire.

This figure clearly shows that, for almost all items, the proportion of correct answers will be expected to differ between the Swiss and French samples. We notice that almost all circled questions are above the diagonal (11 items among 12 circled items), with a difference in proportion that reaches 34% (item 6.9). On the other hand, all the 10 triangle items having problems in both questionnaires are on the diagonal or below it. These results clearly confirm the important effect of translation biases.

Problems Associated with Translation of Items: general case

Various measurements of the degree of similarity between the hierarchies of items, ranked according to the proportion of correct answers, show that the success rate of each item differs quite significantly from one country to another. We have systematically calculated correlations between the country success hierarchies, given by the percentage of correct answers. Although all are significant, the correlation values are stronger between hierarchies relating to questionnaires in the same language (American and Canadian English; Guérin-Pace & Blum, 1999). For example, the correlation is 0.86 between English-speaking Canada and the United States, and between English-speaking Canada and the UK, but is only 0.67 between English-speaking Canada and Sweden.

To study more systematically the relationship between the different questionnaires, we classified the countries into different groups according to their items' success hierarchy [4]. Two countries are grouped in the same class if the difficulty hierarchy is similar in both, irrespective of the general success rate. France is thus grouped with French-speaking Switzerland, as we shall see below, despite a very high variation in the population distribution among the five literacy levels in the two countries. The classification is established on the basis of 97 IALS items [5].

The clusters are characterised by a combination of geographic proximity and linguistic proximity (Fig. 3). All the English-speaking countries are grouped in one class. Thus, the USA is first grouped with English-speaking Canada and then New Zealand; Great Britain, Northern Ireland and Southern Ireland form a class. The two groups are combined and then form a single class with French-speaking Canada. Another class includes the non English-speaking European countries, divided according to the language of the questionnaire. France and French-speaking Switzerland form one class, Germany and German-speaking Switzerland another, and Flemish-speaking Belgium and the Netherlands another. Sweden remains isolated before being added to the class formed by Germany and German-speaking Switzerland.

We verified that this classification is not the result of any other artefact. For example, classifications made separately according to the type of document (prose, document or quantitative) give similar classes.

These results indicate that the item success rate is associated with geographic and linguistic factors, which contradicts the hypothesis of comparability that underpins

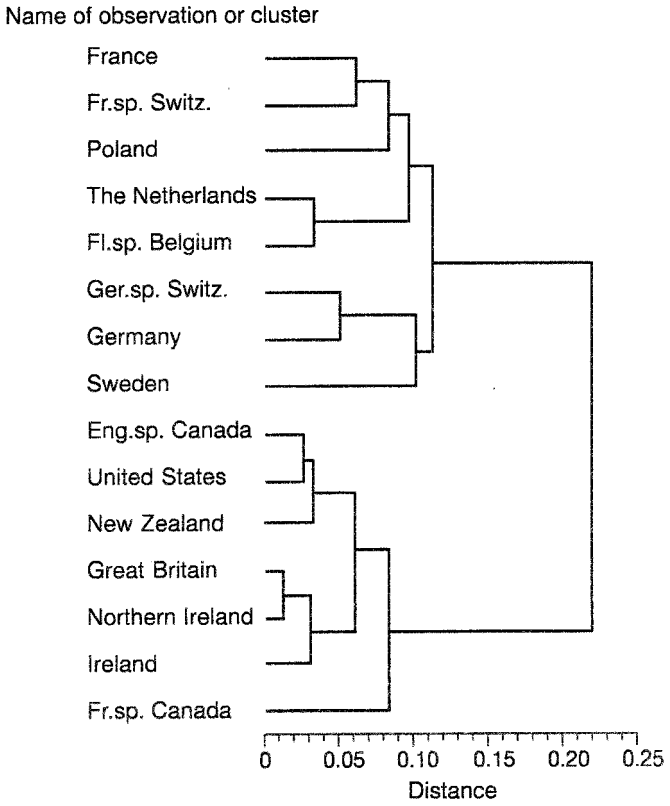

FIG. 3. Classification of countries according to their profile of success in IALS. Single linkage cluster analysis using standardised Euclidean distance metric.

this survey, based on the assumption that performance is independent of the language of questioning.

What Does IALS Actually Measure?

We pointed out earlier that the interpretation of scaled scores and the assignment of levels assumed a uni-dimensional scale; otherwise, it does not make sense to provide the verbal descriptions used in reports of the IALS analyses. In fact, in France, 750 individuals among the 3000 interviewed in 1994 successfully answered at least one question classified as level 5, but have been estimated to be functioning at level 1 or 2! In Great Britain, among 6718 interviewed, 271 were likewise estimated.

The assumption underlying the item response scaling in IALS is not the only one that can be used. In IALS an individual level is based on a weighted average of their responses to all items in a domain. There are, however, many different ways of assigning levels based on the pattern of responses. One very simple alternative is to define the level in terms of the most difficult item that has been answered correctly. This is related to a simple Guttman scale where anyone answering successfully a

question at a given level is then assumed to be able to answer any other question at a lower level.

The individual scores have been re-estimated in this way for interviewees with non imputed IALS scores [6]. We thus define a 'literacy profile' as follows: it is a set of five digits d_1 to d_5, d_i is equal to 1 if at least one question of level i (as defined by IALS) has been answered correctly; otherwise it is equal to 0. Then an interviewee is at level i if d_i is equal to 1, d_{i+1} to d_5 is equal to 0, and d_1 to d_{i-1} equal to 0 or to 1. The sets $(d_i)_{I=1,5}$ are named literacy profiles. Such profiles are said to be *coherent* if d_1 to d_{i-1} are all equal to 1, incoherent if at least one of the $i-1$ digits d_1 to d_{i-1} is equal to 0. In this last case, it would mean that an individual is considered to have level i, but failed in a task of lower level. Among these incoherent profiles, a profile is called 'weakly incoherent' if the only level of failure is just below the literacy level. Tables I and II give the distribution of interviewees, in France and Great Britain, according to the different profiles. Coherence is high: 91% of the profiles are coherent in France, 94% in Great Britain. Moreover, when the profiles are incoherent, it is often a weak incoherence.

Distributions of literacy level, using this measure, are completely different from IALS distributions (see Tables III and IV). Using the IALS measure, 65% of French interviewees with non imputed scores have a prose literacy level of 1 or 2; with a measure based on 'upper level' of success, the proportion falls to 5%. For the UK, the proportions are, respectively, 48% at level 1 or 2 using the IALS measure, and 3% at the same level using the 'upper-level' measure.

Tables III and IV give the transition matrix between the two measures, that is the change in the distribution of individuals among the five literacy levels. It shows a greater concentration on levels 3 and 4. We observe in France that for people at level 1 (IALS), 8% stay at this level (upper measure), 9% move to level 2, 56% to level 3, and about 18% to level 4 or 5. These transfers demonstrate the completely different conclusions that emerge from using the different definitions.

Our goal is not to provide another measure of literacy, but to demonstrate how, using the actual item correct percentages as grouped into levels by IALS, an alternative scaling of individuals produces different results. In addition, the problems already raised of translation of items, attention of individuals, etc., still remain.

This shows that a very simple measure, which can be easily interpreted, may be at least as informative as a very complex measure.

Interviewee Motivation

The consequences of inattention and lack of interest on the part of interviewees towards a long questionnaire requiring real concentration were not discussed in the reports published by IALS.

Nevertheless, this question is crucial. Is it realistic to make the assumption of a uniform behaviour of populations at regional and national level? A bias related to people's attitudes towards the survey can be established with the help of the retest that has been made on a part of the original sample in 1998, in France, Great Britain

TABLE I. Distribution of interviewees (unweighted sample) according to their response profile. Upper level measurement—non-imputed scores

Level	5	4	3	2	1	%
France (2556 interviewees)						
Profiles: coherent						
	0	0	0	0	0	1.5
	0	0	0	0	1	0.7
	0	0	0	1	1	2.2
	0	0	1	1	1	21.2
	0	1	1	1	1	45.9
	1	1	1	1	1	19.7
Coherent					91.2	
Profiles: weakly incoherent						
	0	0	0	1	0	0.9
	0	0	1	0	1	0.8
	0	1	0	1	1	0.2
	1	0	1	1	1	3.6
Weakly incoherent					5.5	
Profiles: incoherent						
	0	0	1	0	0	0.5
	0	0	1	1	0	1.7
	0	1	1	0	0	0.0
	0	1	1	0	1	0.0
	0	1	1	1	0	0.7
	1	0	0	1	1	0.0
	1	0	1	1	0	0.2
Incoherent					3.1	
Great Britain (3306 interviewees)						
Profiles; coherent						
	0	0	0	0	0	0.6
	0	0	0	0	1	0.2
	0	0	0	1	1	1.5
	0	0	1	1	1	18.6
	0	1	1	1	1	45.4
	1	1	1	1	1	27.8
Coherent					94.1	
Profiles: weakly incoherent						
	0	0	0	1	0	0.5
	0	1	0	1	1	0.1
	0	0	1	0	1	0.2
	0	1	0	1	1	0.1
	1	0	1	1	1	2.8
Weakly incoherent					3.7	
Profiles: incoherent						
	0	0	1	0	0	0.1
	0	0	1	1	0	1.2
	0	1	0	1	0	0.0
	0	1	1	0	1	0.1
	0	1	1	1	0	0.4
	1	0	1	1	0	0.2
	1	1	1	1	0	0.1
Incoherent					2.1	

TABLE II. Comparing distributions of literacy level (France and England): non-imputed scores

	Literacy level				
	1	2	3	4	5
France					
IALS level	27	38	31	4	0
Upper-level	2	3	24	47	24
Great Britain					
IALS level	17	31	35	16	1
Upper-level	1	2	20	46	31

TABLE III. Redistribution of population, from IALS-level of prose literacy to upper-level of literacy (unweighted sample), France, non-imputed scores

	Upper-level					
IALS level	Level 1	Level 2	Level 3	Level 4	Level 5	Total
Level 1	7.8	9.1	55.8	13.7	4.1	100
Level 2	0.3	0.6	21.1	58.1	19.8	100
Level 3	0.0	1.3	3.8	55.2	39.7	100
Level 4	0.0	0.0	0.9	33.0	66.1	100
Level 5	0.0	0.0	0.0	0.0	100.0	100
Total	2.2	3.1	24.3	46.8	23.6	100

TABLE IV. Redistribution of population, from IALS-level of prose literacy to Up-level of literacy (unweighted sample): Great Britain, non-imputed scores

	Upper-level					
IALS level	Level 1	Level 2	Level 3	Level 4	Level 5	Total
Level 1	5.3	11.4	58.4	22.5	2.4	100
Level 2	0.0	0.4	29.7	54.8	15.1	100
Level 3	0.0	0.0	3.9	54.3	41.8	100
Level 4	0.0	0.0	0.0	37.7	62.3	100
Level 5	0.0	0.0	0.0	16.3	83.4	100
Total	0.9	2.0	20.1	46.0	31.0	100

and Sweden. For this retest, each interviewee answered a booklet in which one-third of the questions were the same as in the first IALS round.

The first finding is that the proportion of correct answers for each individual is very unstable from one test to the other, both for France (comparing 1994 results with 1998 results: Fig. 4) and Great Britain (comparing 1996 with 1998

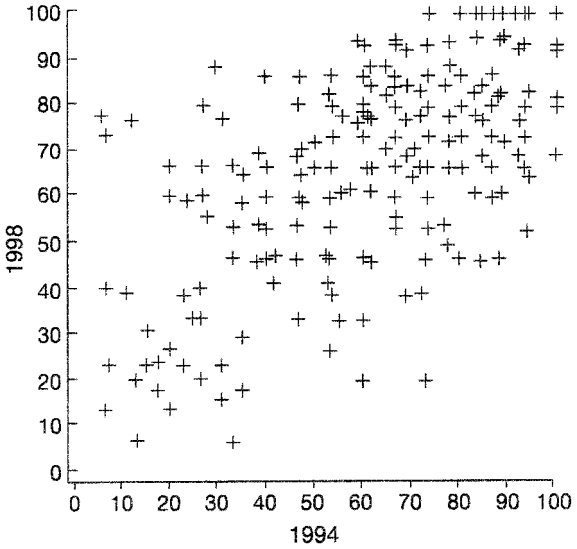

FIG. 4. Comparing proportions of correct answers on French individuals, between IALS 1994 and retest 1998 (300 respondents to French questionnaire).

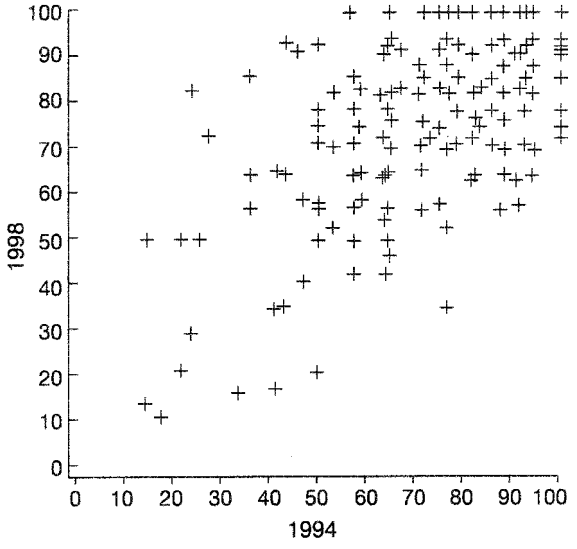

FIG. 5. Comparing proportions of correct answers for UK respondents, between IALS 1996 and retest 1998, same procedure as in 1996 (300 respondents).

results: Fig. 5). This result reinforces the assumption of a strong relation between literacy level and the behaviour of each interviewee at the time of the survey, behaviour that appears to be variable.

The method of processing missing answers also has considerable effects. It is difficult to deal with missing answers in a conventional survey, but the bias induced

is not usually very high. In a survey about literacy, a missing answer may signify a refusal, as well as ignorance. If it is treated as a refusal then the overall result will be over-estimated. If it is taken as a mark of ignorance, the measure will be under-estimated.

An analysis carried out on the French data for the questions scored as 'wrong' has enabled us to demonstrate that many 'wrong answers' should in reality be considered as omissions. In some regions, respondents wishing to ignore a set of questions crossed them out. In other regions the interviewees merely omitted them without putting any marks on the questionnaire. These answers were judged to be wrong in the first case (when crossed out) and omissions in the second. The results therefore are likely to be unreliable and biased. Our analysis has demonstrated that there was actually a geographical bias, which had a major influence on the assessment of ability levels (Guérin-Pace & Blum, 1999). Some of the disparities found, in terms of success in the survey, reflect different attitudes towards the survey that are not allowed for in the scoring process, although they affect the calculation of the individual score.

Another indirect consequence of differential motivation is related to the difficulty of establishing a satisfactory scoring of answers. It is not easy, indeed, to indicate clearly what a correct answer is. This fact has been demonstrated during the 1999 rescoring of a part of 1994 French questionnaires, independently of the first scoring. For some questions, a considerable change in the proportion of correct answers occurred (which could reach 60–80% of changes).

These changes can be interpreted, again, as related to motivation and behaviour of interviewees. People who tried to respond to items very seriously often answered very precisely and their answers are easy to score. Conversely, people who answered very quickly, often gave short sentences and parts of answers, which are more difficult to score, and thus to decide whether they really understood the question.

Conclusions

In the light of our critique we believe that there are important lessons to be learnt from the IALS survey. To begin with we offer the following recommendations for future surveys that might be conducted:

1. The psychometric criteria used by IALS do not provide a satisfactory basis for country comparisons. The one-dimensional models used fail properly to explore the complexity of the data with the result that the conclusions of IALS may well be over-simplifications about the state of literacy in the member countries. These criteria need modification.
2. There is a need to carry out sensitivity analyses of the assumptions made in any Item response modelling. In particular, multi-dimensional models should be explored and rankings of item difficulties compared between countries.
3. Attention should be directed at providing greater validity and recognising that absolute comparability may not be achievable. The survey data should be viewed

as potentially casting light on factors that are locally specific and not amenable to simple scale comparisons between countries.
4. Country comparisons should be carried out at task or 'small task set' level with particular attention paid to translation issues and cultural differences.
5. Multilevel modelling needs to be considered in all analyses of the data in order fully to explore within-country variability.
6. A variety of alternative procedures need to be explored for combining and reporting items with clearly set out assumptions that are used.

The IALS survey, as it stands, should be treated with caution at national level and more so at an international level. The instability of the item success hierarchies due to a combination of linguistic and cultural differences shows that the survey cannot be used on a comparative basis. The operation of translation leads to important biases in the estimated levels of the tasks. The scoring and the processing of omissions in the IALS survey also resulted in a biased assessment of the ability levels due to unequal motivation on the part of interviewees which was not taken into account.

On the basis of our analyses, it is not possible to assume that IALS measures only literacy. It seems to measure a combination of different factors: motivation (reflected in the different ways of filling in the questionnaire), understandings of what items mean, and differences in test taking behaviour more generally. We are not arguing against any kind of international comparative study. Indeed, we think they can be useful. However, we do want to make both the constructors and the users of such surveys more aware of the complexities of design and interpretation, and the caveats that need to be entered about their use.

Acknowledgements

This article has benefited greatly from discussions with Siobhan Carey, Lars Lyberg, Patrick Heady and Kentaro Yamamoto.

NOTES

[1] The questions with the highest scores are the most difficult.
[2] The eight countries in the first wave of the survey, i.e. Canada, the United States, France, Germany, the Netherlands, Poland, Sweden and Switzerland, and five countries in the second wave, Great Britain, New Zealand, Flemish-speaking Belgium, Ireland and Northern Ireland. These countries were chosen from the ones in the second wave of the survey only on the basis of availability of data.
[3] See, for an example of such figures, Guérin-Pace & Blum (1999).
[4] We used an agglomerative hierarchical clustering procedure (Ward's minimum-variance method).
[5] The questions dropped in the IALS analysis do not appear in the analysis.
[6] In France, more than 13% of interviewees didn't fill in the main booklet; in England, less than 1%. We don't include them in this analysis.
[7] Although the logistic 'link function' is commonly used, others are possible. Goldstein (1980) shows that the choice of link function can substantially affect proficiency estimates and argues that this exposes an undesirable arbitrariness of these models.

REFERENCES

BARTHOLOMEW, D. J. (1998) Scaling unobservable constructs in social science, *Applied Statistics*, 47, pp. 1–14.

BENTOLILA, A. & FORT, P. (1994) *Contribution à l'analyse de l'illettrisme en France* (Paris, GPLI).

BOCK, R. D., GIBBONS, R. & MURAKI, E. (1988) Full information item factor analysis, *Applied Psychological Measurement*, 12, pp. 261–280.

BODIER, M. & CHAMBAS, C. (1996) Lire, écrire: les difficultés des adults, *Données Social*, 1996, pp. 490–497.

CAREY, S. (Ed) (2000) *Measuring Adult Literacy—the International Adult Literacy Survey in the European Context* (London, Office for National Statistics).

GOLDSTEIN, H. (1980) Dimensionality, bias, independence and measurement scale problems in latent trait test score models, *British Journal of Mathematical and Statistical Psychology*, 33, pp. 234–246.

GOLDSTEIN, H. (1995) *Multilevel Statistical Models* (London, Arnold).

GOLDSTEIN, H. & WOOD, R. (1989) Five decades of item response modelling, *British Journal of Mathematical and Statistical Psychology*, 42, pp. 139–167.

GUÉRIN-PACE, F. & BLUM, A. (1999) L'illusion comparative. Les logiques d'élaboration et d'utilisation d'une enquête internationale sur l'illettrisme, *Population,* 54, pp. 271–302.

HAMILTON, M. & BARTON, D. (1999) *The International Adult Literacy Survey: what does it measure?* (Lancaster, University of Lancaster, Literacy Research Group).

KALTON G., LYBERG, L. & REMPP, J-M. (1998) Review of methodology in Adult Literacy in OECD Countries, in National Center For Education Statistics, *Adult Literacy in OECD Countries, Technical Report on the First International Adult Literacy Survey*, Appendix A (Washington DC, National Center for Education Statistics).

KIRSCH, I. & MURRAY, S. (1998) Introduction, in T. S. MURRAY, I. S. KIRSCH & L. B. JENKINS (Eds) *Adult Literacy in OECD Countries* (Washington DC, National Center for Education Statistics).

KIRSCH, I. S., JUGENBLUT, A. & MOSENTHAL, B. (1998) *The Measurement of Adult Literacy. Adult Literacy in OECD countries* (Washington DC, US Department of Education, Office of Educational Research and Improvement).

MURRAY, T. S., KIRSCH, I. S. & JENKINS, L. B. (1998) *Adult Literacy in OECD Countries* (Washington DC, National Center for Education Statistics).

OECD (1997) *Literacy Skills for the Knowledge Society* (Paris, OECD).

STREET, B. (1996) Literacy, economy and society, *Literacy Across the Curriculum*, 12, pp. 8–15.

Statistical Appendix: defining dimensionality

Dimensionality refers either to a set of items, or alternatively to a set of test scores. While the detailed procedures for investigating dimensionality will differ in each case, the essential underlying models are the same. The essence is captured in the following simple uni-dimensional factor model for a set of test scores:

$$y_{ij} = a_i + b_i f_j + e_{ij}$$
$$f \sim N(0,\ \sigma_f^2),\ e_{ij} \sim N(0,\ \sigma_{ei}^2) \qquad (1)$$

where y_{ij} is the score for the i-th test for the j-th individual, f_j is the underlying factor value for the j-th individual and the $e_{ij}$are mutually independent 'residual' terms. The intercept term, a_i, is often omitted if all the measured variables are standardised to have zero means. If the responses y_{ij} are replaced by a set of item binary responses then with minor modifications we can write:

$$\begin{aligned} &\text{logit}(\pi_{ij}) = a_i + b_i f_j \\ &y_{ij} \sim Binomial(\pi_{ij}, 1) \\ &f \sim N(0, \sigma_f^2) \end{aligned} \tag{2}$$

The basic similarity resides in the fact that a single underlying variable f determines the response through a simple regression type relationship, apart from random variation. Both models (1) and (2) are a special kind of 2-level model in which individuals are at level 2 and tests (or test items) at level 1. In addition to the uni-dimensionality assumption, the independence of the e_{ij} in (1) and the independence of the y_{ij} given f; i.e. the item coefficient values and the individual's proficiency, is a further assumption that underlies the use of significance testing of the model, construction of confidence intervals and as a basis for testing for the degree of dimensionality which may exist. It is worth noting that in section 10.4 this assumption is incorrectly described. Model (2) is precisely the model used by IALS and is often known as a 'binary factor model' and is referred by IALS as the 'two-parameter logistic model', and the notation used by IALS is also slightly different (Chapter 10). A useful discussion of these models is given by Bartholomew (1998) [7].

The aim of the statistical analysis of these models is to estimate the parameters, and, in particular, to provide estimates of the values of f_j, one for each individual. These are known as factor or trait scores or 'proficiencies'. They are, in effect, weighted averages of the responses—in the case of test items the (0,1) responses, where the weights depend on the values of the b_i estimates. Here, we shall explore a little further what the use of such a model implies substantively.

For simplicity we shall use the traditional factor model (1), but everything we say will apply in general terms to (2) also. Suppose that individuals' responses were, in fact, determined by two underlying responses according to the following model:

$$\begin{aligned} &y_{ij} = a_i + b_i f_j + c_i g_j + e_{ij} \\ &f \sim N(0, \sigma_f^2),\ g \sim N(0, \sigma_g^2),\ e_{ij} \sim N(0, \sigma_{ei}^2) \\ &\text{cov}(f, g) = \sigma_{fg} \end{aligned} \tag{3}$$

In the IALS case, such a model would be fitted for a collection of items that are assumed to reflect two domains, say prose and document literacy. IALS makes the strong assumption that for each domain the items used reflect that domain and only that domain. Yet the high intercorrelations observed among the proficiency scores suggests that this is very unlikely. The advantage of a full multi-dimensional analysis is that it would provide some insight into how any underlying domains that can be identified from the analysis predict the responses to the test items.

Section 10.3 of the Technical report describes a (MH) test for detecting individual items that have different parameter values in some countries. While one would expect the existence of more than one dimension to lead to such a situation the non-existence of such items does not imply uni-dimensionality. In any case, as this section points out, the test is very approximate.