December 1999

International comparisons of adult literacy

by

Harvey Goldstein

Institute of Education, University of London

h.goldstein@ioe.ac.uk

Introduction

The International Adult Literacy Survey (IALS) represents the collaboration of a number of countries who agreed to co-operatively investigate 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 US, 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 paper is based upon a fuller version that will appear in a report to be published early in 2000 by the UK Office for National statistics ONS).

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. Respondents, about 3,000 in each country, were tested in their homes using three booklets, one for each scale. Background information was collected on the respondents and features in some of the analyses. The results of the survey received wide publicity and a new survey on 'life skills' has been set up by OECD with a greater number of countries and using similar procedures.

There have been several commentaries and critiques of IALS. Most of these (e.g. Street, 1996), 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 paper 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 paper, including sampling problems, scoring variability and response rates; these are discussed in the ONS report.

We look at the procedures used in IALS to define literacy by the way in which test items are selected, how 'scales' were constructed and reported on and the ways in which the data have been analysed. 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 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 report as 'item response theory' techniques.

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 3 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. 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; such as the presence of exemplars, literality etc. 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 there are such dimensions which are common to all three domains and which are responsible for the observed high intercorrelations. In future work this is one area for research, using multidimensional item response models of sufficient complexity. The IRMs used in IALS are all unidimensional, 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. In the Appendix we give a brief formal description of what is meant by 'dimensionality' of a set of items.

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 (unidimensional) 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 which better satisfies the model assumptions, in particular 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 initial decisions about which items to include and in what proportions, will determine the final scale when a unidimensional model is assumed. We shall return in more detail to this issue 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 oversimplified models which 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 posses 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 on substantive grounds. The other is to allow the final reporting decision to be made following an exploration of the dimensionality structure of the 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 falls between these, neither allowing a proper exploration of the dimensionality of the data nor allowing 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 relevance of these is ignored in the technical report.

Item exclusion

According to Chapter 10 of the technical report, twelve of the original 114 items were dropped because they did not fit very well (model (4) given in the appendix), involving a large discrepancy value in 3 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 in 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 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 which do not fit the model.

An item which does not 'fit' a particular unidimensional 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 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 is that there are in reality two or more dimensions which 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 'non-fitting' items and forcing all the remaining items to have the same parameter values for each country in a unidimensional 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 more closely the US pattern 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. 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. Such an attempt to give 'meaning' to the scale 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 oversimplification 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.

The logic of the unidimensionality assumption, however, is that since only a single attribute is being measured the resulting scale score summarises all the information about the attribute and 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 (1-dimensional) proficiency score, the relative difficulties of all the items is assumed to be the same. 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.

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 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 and Wood, 1989 for a further discussion). In fact one might wish to argue for a summary such as the proportion correct simply on the grounds of this being a useful summary measure without any particular modelling justification.

It would be advantageous for a separate scaling to be done for each country. In this way differences can be seen directly (and tested) rather than concentrating on fitting a common scale. This will make the scaling procedure more 'transparent' and allow more substantively informed judgements to be made about country differences.

Another important approach is to see whether item groupings could be established for small groups of items which, 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 which 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.

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, (P71) the report 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. In particular it would 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 multidimensonal 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 appendix, could include such covariates. As Goldstein and 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.

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 multilevel analyses which take account of differences between schools, 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.

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 oversimplifications 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, multidimensional 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 which 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 used 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.

For all the reasons given the IALS survey, as it stands, should be treated with caution at national level and more at so at an international level. We are not arguing against any kind of international comparative study, indeed we think they can be useful. Rather we want to make both the constructors and the users of such surveys more aware of the complexities of design and interpretation and the caveats which need to be entered about their use.

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 unidimensional factor model for a set of test scores

$$\begin{aligned} &y_{ij} = a_i + b_i f_j + e_{ij} \\ &f \sim N(0, \sigma_f^2), \quad e_{ij} \sim N(0, \sigma_{ei}^2) \end{aligned} \tag{1}$$

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

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

The basic similarity resides in the fact that a single underlying variable f_j 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 unidimensionality assumption, the independence of the e_{ij} in (1) and the independence of the y_{ij} given π_{ij}, i.e. the item coefficient values and the individual's proficiency, is a further assumption which 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).[1]

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

[1] 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.

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

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

In the IALS case, such a model would be fitted for a collection of items which 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 multidimensional analysis is that it would provide some insight into how any underlying domains which 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 which 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 unidimensionality. In any case, as this section points out, the test is very approximate.

References

Bartholomew, D. J. (1998). Scaling unobservable constructs in social science. *Applied Statistics* **47**: 1-14.

Bock, R. D., Gibbons, R. and Muraki, E. (1988). Full information item factor analysis. *Applied Psychological Measurement* **12**: 261-280.

Goldstein, H. (1980). Dimensionality, bias, independence and measurement scale problems in latent trait test score models. *British Journal of mathematical and statistical psychology* **33**: 234-246.

Goldstein, H. (1994). Recontextualising mental measurement. *Educational Measurement: Issues and Practice* **13**: 16-19.

Goldstein, H. (1995). *Multilevel Statistical Models*. London, Arnold; New York, Wiley.

Goldstein, H. and Wood, R. (1989). Five decades of item response modelling. *British Journal of mathematical and statistical psychology* **42**: 139-167.

Mislevy, R. J. (1991). Randomisation based inference about latent variables from complex samples. *Psychometrika* **56**: 177-196.

Murray, T. S., Kirsch, I. S. and 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**: 8-15.