Proceedings of Statistics Canada Symposium 2001
Achieving Data Quality in a Statistical Agency : A Methodological Perspective

METHODOLOGICAL PROBLEMS RAISED BY AN INTERNATIONAL SURVEY – *THE INTERNATIONAL ADULT LITERACY SURVEY*

Alain Blum, France Guérin-Pace[1]

ABSTRACT:

The International Adult Literacy Survey (IALS), coordinated by Statistics Canada, was conducted in some 20 countries between 1994 and 2000. Based on the survey's findings, a wrap-up report containing a comparative analysis of reading skills in participating countries was published in 2000 under the auspices of the OECD. Unfortunately, there are very serious methodological problems with the survey that make it unusable for comparative purposes.

This presentation discusses the survey's weaknesses and, more generally, the pitfalls of applying a universal measure of skills in countries with different cultures and different languages. Analyses of the survey results reveal the extent to which translation and wording altered the difficulty of the test questions. The range of scores on the various items among the participating countries confirms this linguistic bias. Moreover, respondents' attitudes (motivation, attentiveness, refusal, etc.) are not only culturally marked and tightly bound to the survey tradition in their own countries, but a determining factor in a survey intended to construct a measure of skills. Only highly detailed coding could separate out respondents' attitudes toward this long, difficult survey and differentiate between a deficiency of skills and a lack of motivation or attention.

This paper is based on an assessment of the survey within the European Community. Its findings have led to the publication of a series of articles and a book on the subject.

KEY WORDS: Literacy, error of measures, comparative surveys.

1. THE IALS SURVEY. PRESENTATION AND GENESIS

The IALS (International Adult Literacy Survey) launched in 1994 illustrates comparative approach which are now often initiated by the European Union or international organizations. The goal of the IALS survey was to measure the literacy skills required to read and understand documents encountered in everyday life (instruction manuals for household appliances, directions for use on medicines, press articles on current affairs, various diagrammatic representations of descriptive statistics, transport timetables, and so forth).

Coordination of the IALS has been done by Statistics Canada which was responsible for encouraging its implementation at the international level in 1992, in collaboration with the private American institute ETS (Education Testing Service) and the OECD. Eight countries took part in the first round of surveys: Canada, United States, France, Germany, The Netherlands, Poland, Sweden and Switzerland. A further round of surveys was conducted in 1996 in the following countries: Great Britain, Northern Ireland, New Zealand, and Ireland. The survey has then been conducted in other OECD countries (OECD, 2000). In all 21 countries have been surveyed, including 19 OECD countries.

The results of the international survey were presented in the form of comparative tables giving the distribution of the population among five rising literacy levels. Level 1 is the lowest level and comprises

[1] INED, 133 Bd Davout, 75020 Paris, France. blum@ined.fr; guerin@ined.fr. This paper is a synthesis of articles refered in the bibliography.

the people who are almost illiterate, able perhaps to decipher sentences or perform simple calculations. The highest level (level 5) comprises the people who have no difficulty understanding non-specialist articles and texts, and can write letters and perform calculations, and so on.

Using this procedure the population of each country in the survey is distributed over the literacy scale, thus encouraging comparisons between proportions of people among the literacy's levels. These data were then widely circulated and received extensive coverage in the press and from politicians. In the case of France, this media and political attention was accentuated by the fact that three-quarters of French people had levels of literacy that were too low for them to be able to perform normal everyday tasks such as reading a newspaper, writing a letter, making sense of a short text or a pay slip. The scales used by the originators of the IALS survey, suggested that 75% of French adults had a literacy level estimated at 1 or 2 for the understanding of prose texts, compared with 46% of Americans, 40% of Netherlanders and 28% of Swedes. For the understanding of schematic texts, the proportions at levels 1 and 2 were respectively 63%, 50%, 36% and 25%, while for the comprehension of texts with a quantitative content the proportions for the same countries were 57%, 46%, 35% and 25%. Also, for the highest skills level (level 5), only 11 of the 3000 people who took part in the survey in France possessed the skills associated with this level, far fewer than the number of respondents who had received higher education.

2. THE LIMITS OF THE COMPARATIVE APPROACH

Several studies have examined the value of the IALS survey for comparative purposes and, more generally, the validity and significance of its findings. In this presentation we focus on the validity of the comparisons based on an examination of the individual-level data and general results from the survey in the 13 countries listed above.

In particular three important questions are examined:
- is the equivalence of difficulty's questions preserved when they undergo the translation process? In other terms, does a good translation of items necessarily ensure an equivalent test between languages?
- Is it plausible to assume uniform behaviour of the populations at regional and national levels? A crucial question concerns respondent motivation in relation to a questionnaire that is both long to answer and intended to measure individual ability.
- is the synthetic measure given by the scores robust? More precisely, does another measure based on similar materials provide equivalent results ?

2.1 Comparison, level of difficulty and translations

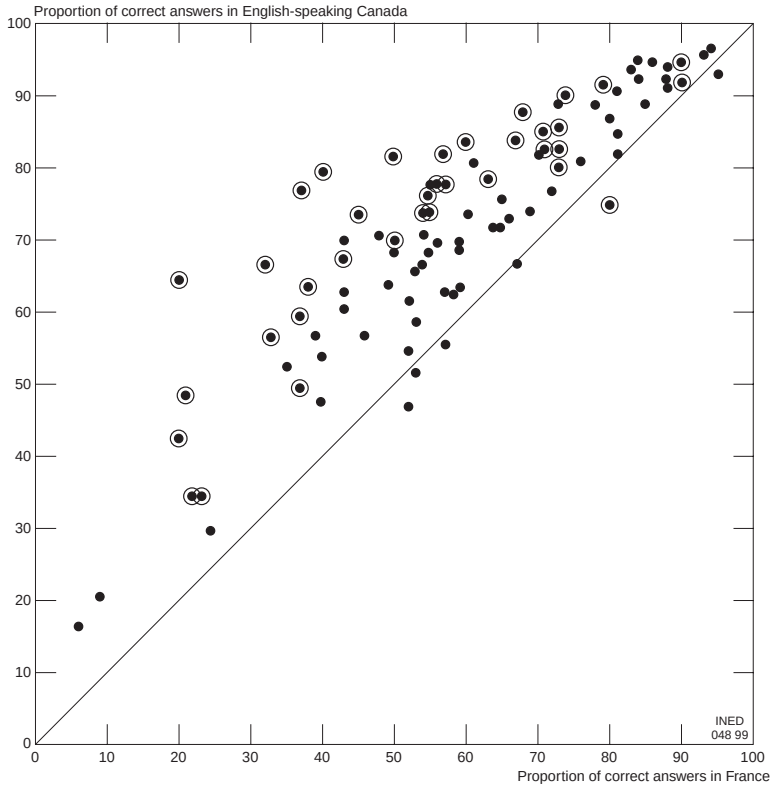
The validity of the IALS survey is based on the strong hypothesis of an identical difficulty scale existing for assessment across cultures and languages. In order to check this hypothesis we compared the success rates observed, question by question, in the countries taken two at a time. If the questions are presumed to be of equivalent difficulty in the two countries being considered, the graphic representation of the percentage of correct answers to each question should approximate to a straight line whose slope is a function of the overall literacy level. But examination of the graphs shows the clusters to be extremely diverse in form. From a comparison of the percentages of success in France and English-speaking Canada - whose questionnaire provided the basis for the French translation - we identified a set of questions with high variations in success rates and attempted to explain these disparities (figure 1).

To begin with, a comparison of the way the questions were formulated in the two languages revealed significant differences in the translations. On a broader scale we examined the questionnaire as a whole to identify the questions whose wording was not 'equivalent' in French and English. We identified 35 questions

(circled in figure 1) for which the translations differed and have analysed these divergences under three headings[2].

Omission of repetition of terms: Repetition of terms from the questions in the text that supplies the answer is more frequent in the original documents (English) and represents the first source of bias. The effect is draw the respondant in English more easily to the sentence containing the answer**.**

Figure 1: France – English-speaking Canada comparison; Percentage of right answers



Greater precision of English terms: The questions are in general more precisely worded in English. For example, one question which in English reads '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 between question and sentence is therefore easier to establish in English.

Translation errors: Some translation errors though often minor from a strictly linguistic point of view can have important implications for understanding. 'If you wanted to more than double your principal within five years, what rate..” is translated into French as '.. doubler votre capital *en moins* de cinq ans ..”. The translation of “within 5 years” by “”en moins de 5 ans” has the effect of excluding a lenght of 5 years in

[2] The formulation of all these questions and the precise problems which arose for each of them can be found in Guérin-Pace and Blum, 2000

French, and there is no solution to the question in the corresponding table. People who indicated that there was no response to this item were considered as beeing wrong.

Other sources of error: Sources of error can be identified which, though less widespread, reveal the complexity of a definition of difficulty equivalence between items expressed in different languages. The example that follows is typical and can be interpreted in terms of 'cultural bias'. The task involved is to decide which are the comedies in a review covering four films. In two of these reviews, in both English and French, the word comedy appears, which thus simplifies the question. In France, however, it was found that many interviewees gave as their answer a third film even though from the description it is not a comedy. The explanation is the presence in that film of the actor Michel Blanc, who is well known to the French public for his roles in many comedies but little known abroad. In this case, the answering process was dominated by the association of ideas, to the detriment of a careful reading of the reviews.

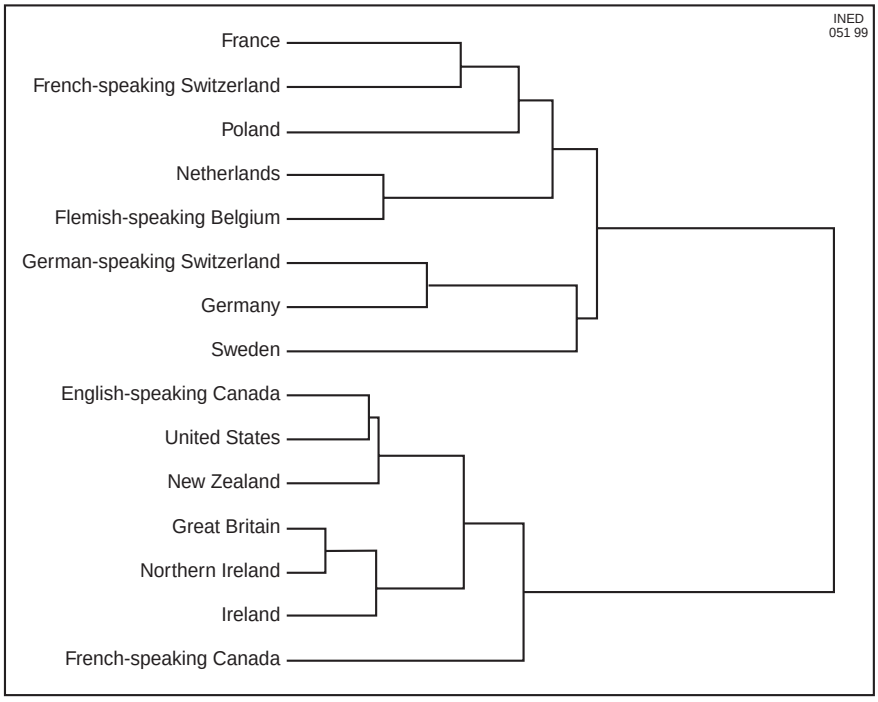
2.2 Classification of countries by the item success profile

Divergences between the level of difficulty of the questions defined in advance and their success can also be seen in countries other than France. Various measurements of the degree of similarity of the hierarchies show that even with a high degree of association they differ quite significantly from one country to another. To place the emphasis on the similarities in terms of relative success in the questions regardless of the percentage of right answers to each of the items, we set up a partition of countries by the rank of each of the questions in the success hierarchy using a cluster analysis. Irrespective of the population distribution by levels of literary, among the various countries a correlation can be seen between the ranking of the questions, the language of the interview and the country in which the survey was carried out[3].

The contents of clusters show that the parallels established are based on a combination of geographic proximity and linguistic proximity (figure 2). 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. These latter countries have the same version of the questionnaire. 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.
This overall result goes strongly against the foundations of the protocol of this survey, which are based on the fact that performance is independent of the language of questioning.

[3] That is $r_{i,j}$ the rank of item *i* (from the most to the least successful) in country *j ;* the cluster analyses are made on the table R(i,j) = (r_{ij})

Figure 2: Classification of countries by success hierarchy, eliminating dropped questions



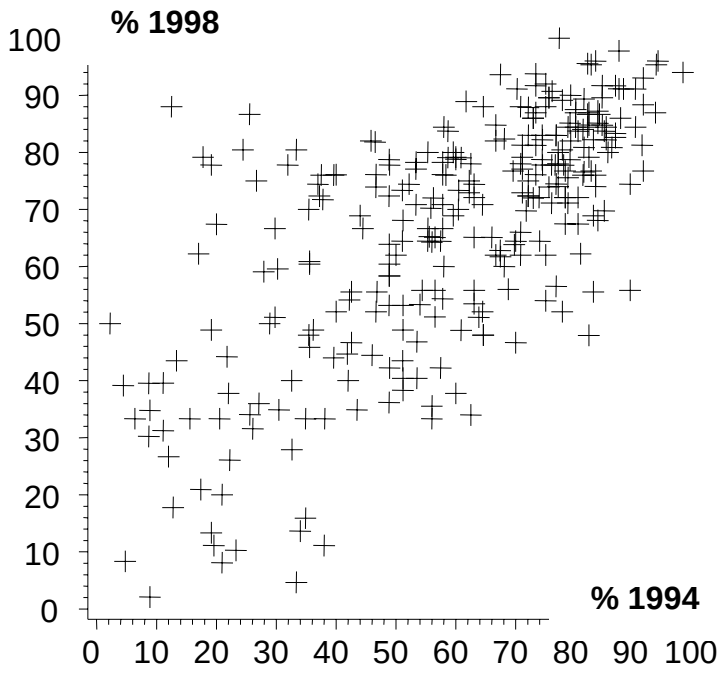
2.3 Retest and translation effects

A retest has been done (See Carey, 2000) in which a sample of French individuals interviewed in 1994 have been interviewed again in 1998. About half of the sample (300 respondents) was questioned with the original French questionnaire while the other half (422 respondents) has been interviewed with the Swiss questionnaire[4]. The main purpose of doing this is to estimate the impact of using a different translation, knowing that some of the problems found in the French questionnaire are not present in the Swiss version.

First of all, differences in response profile between 1994 and 1998 show a strong instability of the results. (Figure 3). This important variability raises the question of the reliability of the literacy measure.

[4] A first analysis shows that the two samples are almost similar, in terms of their socio-demographic structure. Using weights computed on the basis of each sample does not affect the response profile of items.

Figure 3: Proportion of correct answers for each individual in 1994 and 1998 (individuals with French questionnaire in 1998)



Secondly the result demonstrates a strong effect of translation, especially its impact on the change of "incorrect" answers between 1994 and 1998. For almost all items, the proportion of correct answers is much more important using the Swiss questionnaire than using the French one (Blum and Guérin, 2000). It demonstrate a significant positive effect of the translation on the proportion of correct answer per individual. In other words, using the Swiss questionnaire in 1998 instead of the French one, increases significantly the probability to give a correct answer for an item that had a translation problem in the 1994.

2.4 Analysis of missing answers

The consequences of inattention and lack of interest on the part of interviewees towards a long questionnaire requiring real concentration is a very important one. Nevertheless, the results in the various European and North American countries suggest that this question is important. The percentage of the population in literacy levels 1 and 2 varies between 28 % (in Sweden) and 77 % (in Poland) for the tasks associated with prose texts. These percentages are considerable, not only in France but in the other countries as well: being in level 1 means that one can 'pinpoint an element of information contained in the text which is identical to or synonymous with the information given in the instruction' and in level 2: ' pinpoint one or more elements of information in the text, which may contain a number of distracting elements, or the reader may have to make low-level deductions' (Statistics Canada, OECD, 1995). At the other end of the scale of ability, the percentage of people in level 5 is particularly low, so low in fact that it was not published: levels 4 and 5 were grouped in the same class in all the publications issued by IALS and in the database distributed. A level 5 task 'requires the reader to search for information in a dense text which contains a number of plausible distracting elements'.

These surprising results point us towards a more detailed analysis of how seriously the people interviewed responded to the survey. We carried out this study using the French data, by examining the consequences of this lack of interest on the regional differentiations. The subject studied is as follows: Can different attitudes to the survey be detected in the different regions of France?

We explored the theory of unequal interviewee motivation by region, reflected in different behaviour in the way the questions were answered. For example, when faced with a question which is not understood or perceived as difficult, some interviewees try to give an answer, even if wrong, while others ignore the question and move on to the next one. If an interviewee omits a question because of its difficulty, it is desirable for the omission to be considered as a wrong answer, as the reason for the omission is doubtless the expectation of failure. However, not every omission can be interpreted in this way because some of the people interviewed freely choose not to answer because of either lack of interest or laziness, but not due to lack of understanding.

Two approaches were followed by the authors of the IALS survey to try to overcome this pitfall. The instructions supplied to the interviewers state that the people questioned should be told as soon as they read a question that if they do not know how to answer they should put some kind of mark on the question (put a cross, cross through the lines for the answer etc.). The answer is then considered to be wrong. There should be no such mark if the question is actually omitted, i.e. if interviewees do not look at the document because they have given up.

We then tried to identify what could be interpreted as omissions included as wrong answers. As we have seen, a single document is used as the basis for several questions. The document and the questions referring to it are located opposite each other. An examination of the questionnaires completed by interviewees in France often reveals whole pages crossed out, even in the questionnaires filled in successfully elsewhere. It is likely that some of the answers coded as 'wrong', are in fact expressions of lack of interest and motivation rather than lack of ability.

We have shown elsewhere (Guérin-Pace and Blum, 2000) for France that the regional map of the omissions (partial omissions or quitting before finishing the test) is complementary with the map of 'refusals', which nothing could have led us to assume. The regions with a large number of answers 'crossed out' have a small number of no answers and reversely. This finding suggests that the answers 'crossed out' actually correspond to a form of no answer.

The determinants of the various types of answer, computed using different logistic regression models, confirm this result. Each of the three variables - number of 'wrong answers' per interviewee, number of 'refusals' and number of omissions - was explained as a function of the following individual characteristics: region of residence, age, qualifications, score for the question and time taken to complete the questionnaire. The determinants of the wrong answers have a more demographic and social origin, with qualifications having a significant influence, along with age. The 'refusals' variable is closer to the 'omission' variable than to the 'wrong answer' variable (Guérin-Pace, Blum, 2000).

This analysis demonstrated then that there is no reason to assume homogeneous attitudes within the same country and therefore even less between countries. We did not make these assessments for the other countries but that differentials are introduced for the same reasons, again leading to doubts about the comparative nature of the survey. The variations of motivation from one individual to another, become a strong element of bias in itself that changes the estimated literacy level for each individual. Many of the failures are probably the result of lack of seriousness or lack of interest when faced with a long survey. Moreover, not any comparative analysis can be made, taking into account that this lack of interest could be different from one country to another one, and, moreover, could be expressed differently on the questionnaires.

3. ALTERNATIVE SCALINGS

The assumption underlying the item response scaling in IALS is not the only one that can be used. To be robust, a measure has to be confirmed by other alternative measures, based on the same principles and on the same materials. 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. 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; else 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. 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. 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 Great-Britain, 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.

In France for people at level 1 (IALS), 8% stay at this level with the 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 which emerge from using the different scaling.

In presenting this very simple scaling, we don't want 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.

4. CONCLUSION

Serious doubts thus exist about the value of the IALS survey for cross-national comparison. The study of these problems lead to important conclusions concerning more general comparative surveys, of an interest broader than interest concerning only this survey. The translation and adaptation into the different languages, and the attitude of different national populations towards this type of survey, are responsible for disparities in the measurements that cannot be interpreted in terms of literacy. The production of several synthetic measurements obscures the complexity of the phenomena under examination, and causes the process behind construction of the measurement to be overlooked.

In the case of the IALS survey the American experience appears to have been over-hastily extended to an international context. The collaboration that occurred between the international bodies did not in itself justify the international dimension given to the survey. In its conception and content the questionnaire was almost exclusively derived from what was originally a North American initiative. Consequently it was not the product of a comparison between the differing national experiences on these questions. More seriously, the statistical methods employed, when tested in a national context, are found to be inadequate to validate the comparative value of the survey.

In addition, the survey's organizers under-estimated the degree of cultural bias associated with the differing attitudes of respondents in the different countries.

Finally, the contrasts we have identified are revealing of some of the underlying mechanisms which account for success or failure on an item and can again be linked with actual survey procedures. Interviews conducted with a number of respondents show that answering quickly is often preferred at the expense of understanding the document fully. In these conditions it is reasonable to ask exactly what is being

measured? A visual perception skill or a real ability to understand and interpret information for the task in hand? In other words, doubts must exist about the suitability of this assessment instrument for measuring the skills needed to 'get by' in everyday life.

We have shown that the data produced by the IALS survey are far from satisfying the original ambitions. The bias observed occurs to varying degrees across the participating countries and its respective effects are hard to measure. For France, for example, we have shown that the translation of the questionnaire caused an increase in difficulty for over a third of the questions asked.

This analysis shows the need for great care in each stage of the elaboration of a survey instrument. This is even more of an imperative when, as was the case here, once the information has been collected and the estimations established, the survey emerges from the field of research and enters the public domain, acquiring an irrefutable status, regardless of the reservations that specialists may have expressed.

REFERENCES

Blum, A., Guérin-Pace, F. and Goldstein, H. (2001), "An analysis of international comparisons of adult literacy", *Educational Assessment*, July.

Blum, A. and Guérin-Pace, F.(2000), *Des lettres et des chiffres – Des tests d'intelligence à l'évaluation du "savoir lire", un siècle de polémiques*, Paris, Fayard, 191 p.

Carey, S. (ed) (2000). Measuring Adult Literacy - The International Adult Literacy Survey in the European Context. London, Office for National Statistics.

Darcovich, N., M. Binkley, J. Cohen, M. Myrberg and S. Persson (1998), « Non-response Bias » *in* National Center for Education Statistics, op. Cit..

Développement des ressources humaines Canada, OCDE (1997), *Literacy Skills for the knowledge society, Further Results from the International Adult Literacy Survey*, Paris, 195 p.

Goldstein H. et R. Wood (1989) : « Five decades of item response modelling », *British Journal of Mathematical and Statistical Psychology*, vol. 42, pp. 139-167.

Guérin-Pace, F. and Blum, A. (1999), L'illusion comparative. Les logiques d'élaboration et d'utilisation d'une enquête internationale sur l'illettrisme. *Population* 54:271-302.

Guérin-Pace, F. and Blum, A. (2000), The comparative illusion : The International Adult Literacy Survey ; *Population/ An English Selection*, 12:215-246.

Kalton G., L. Lyberg et J.-M. Rempp (1998), Review of methodology in Adult Literacy in OECD Countries in NCES, op. Cit. Kirsch I, A. Jungeblut et B. Mosenthal (1998), « The Measurement of Adult Literacy », in National Center for Education Statistics, op. cit., ch.7.

Kirsch, I., Murray, S. (1998). Introduction. In: Murray, T. S., Kirsch, I. S. and Jenkins, L. B. (1998). *Adult literacy in OECD countries*. Washington, DC, National Center for Education Statistics

Murray, T. S., Kirsch, I. S. and Jenkins, L. B. (1998). *Adult literacy in OECD countries*. Washington, DC, National Center for Education Statistics.

National Center for Education Statistics (1998), *Adult Literacy in OECD Countries, Technical Report on the First International Adult Literacy Survey*, U.S. Department of Education, Office of Educational Research and Improvement, Washington, p. 215 and appendices.

OECD (1997). *Literacy skills for the knowledge society.* Paris, OECD.

Statistique Canada, OCDE (1995)- *Littératie, Economie et Société, Résultats sur la première Enquête internationale sur l'alphabétisation des Adultes*, Paris, 218 p.

OECD (2000), Literacy at the Information Age – Final report on the International Adult Literacy Survey, Paris.