Validity in adult assessment surveys in France

Dr Jean-Pierre Jeantheau

19/09/2008 11:05

Sommaire

Validity in adult assessment surveys in France

Dr Jean-Pierre Jeantheau

Natioanl Agency for fighting illiteracy

I Background:

A) Assessment of adult competencies in large surveys

Before examining the validity issue, a reminder of the context may be necessary. The evaluation of reading competencies or, more broadly, of basic skills is not new. It has been practised for a very long time by teachers in the form of exercises, exams, "tests", and so on. Schools or national institutions have also used evaluation for a very long time in the process of certification. Recently it has even been focused on "formative assessment", which places the emphasis on evaluation as the driving force of the learning process. All the evaluations are not similar and have neither the same goals nor the same methods.

Assessment of an adult population is more complex, especially if this population is selected according to geographical criteria or age groups, or both. In those cases the population will be very heterogeneous in terms of school background as well as formal, practical, and professional competencies.

Despite this structural heterogeneity, the evaluation of adult competencies has long been completed according to a school pattern or according to a "vocational" pattern. The first one is essentially based on the hypothesis that certain formal

positive performances reflect the possibility of executing tasks which are not, themselves, directly evaluated. The second one is essentially based on the hypothesis that a successful task can be effectively repeated in contexts which are more or less similar. For example, if somebody is able to build one wall as requested, they will also be able to build other walls. These two evaluation types take place in the framework of the learning process or vocational training and lead to certifications. Another method, which is entirely different, the portfolio, enables to list the evidence of the skills to be listed, plus the benefits which can be granted to the bearer of the certificate.

Before using direct assessment, the evaluation of adult (population) competencies was based on the interviewees' self-reports and the inferences made by statisticians and researchers in charge of the data processing and results. This methodology is still used by UNESCO surveys to collect data on very low levels of literacy (illiteracy), but only because direct assessment results are not available.

For years, in various countries throughout the world, direct assessment surveys have been conducted on large adult populations. Probably, the first surveys have been carried out by the army for recruitement. The tested domains were directly linked with the needs of the army . In a recent past we saw the development of adult surveys on a large group of adult population. The USA conducted YALS (a household survey focusing on young adults) in 1985, and its extensions to a wider adult population (16+) in 1992 (NALS) and 2003 (NAAL). Great Britain also carried out many adult literacy surveys, among them those entitled "Older and

Younger" in1992-93, and "Skills for Life Survey"in 2002-03. At international level (introducing the comparison dimension) we may recall IALS (International Adult Literacy Survey), from 1994 to 1998, ALLS (Adult Life Skills and Literacy Survey) in 2003, IVQ (Information and daily life) in France from 2002 to 2008, LAMP (Literacy Assessment and Monitoring Programme) in progress since 2003, and PIAAC (Project for International Assessment of Adult Competencies), which is currently being developed and is scheduled for 2010 or 2011.

B) The concept of validity

When reading the literature we notice that the concept of validity has evolved particularly since the middle of the 20th century, together with the monitoring techniques which were used. Traditionally, the validity of an evaluation tool reflects its capacity to measure whatever such a tool is designed to assess. The essential nature of adequacy is obviously perceived, but immediately afterwards rises the difficulty of monitoring it particularly in the field of arts and science.

The literature focuses on instrument validity. But maybe it would be more reasonable to speak about the validity of surveys, because as the *Sage Encyclopaedia* explains, "validity not only depends on the instrument itself, but [on] how you use the instrument". This point highlights the fact that instrument validity issues are embedded in a more general setoff problema concerning surveys, including the motivation issue, and the method and the conditions of the test administration.

The literature generally distinguishes at least three kinds of validity: "Content Validity" which refers to the extent to which the test items represent the domain or universe of the trait or property being measured, "Construct Validity" which refers to the theory underlying the construct to be measured and to the adequacy of the test in measuring the construct, and "Criterion-Related Validity". This approach is concerned with detecting the presence or the absence of one or more criteria considered to represent traits or constructs of interest.

Usually content validity, construct validity and criterion related validity are addressed by an expert panel, which is set up and consulted specifically for this purpose.

Sometimes internal consistency, which in fact is related to test reliability and is usually tested with Cronbach's alpha, is considered as an estimate of the test validity- but it is a minority view.

C) Methods for testing validity

We can consider that there are 3 main categories of methods designed to test the validity of an instrument (of a survey).

The first category gathers methods, which can be described as "Internal Methods" for examining validity and consist in examining each assessment tool and expressing a judgment on its quality using statistical tests on the coherence of the results and/or on experts' opinions. Though this approach cannot be the only test of validity, especially in the Humanities, it should rather be considered as a validity

presumption. The main risk is that this agreed validity might imply that it itself defines the object that it has to evaluate. An absolute theoretical validity may be reached, even though it is not very reliable in the field of social sciences or cognitive sciences as it restricts reality too much.

The second category assembles methods which can be referred to as "Comparative", and which proceed to compare the examined tool to other tools or indicators which are supposed to measure an identical dimension or a dimension with a similar nature, both of those being considered to be valid.

Finally, the last category, which can be referred to as "External Examination" , sets the results provided by the tool against objective external facts which are considered to be indicators representative of the command of the assessed dimension. In fact, this last category tries to measure the predictive nature of the assessment tool.

In literature and also in practice, especially in the field of psychology, the last category is considered as the better solution to test validity. One of the easiest ways to test for criterion-related validity is to administer the instrument to a group that is known to exhibit the trait to be measured. Such a methodology has been used for the development of tests in psychology. The tested group may be identified by a panel of experts.

The study of an external validity in the Humanities is a long process. With time running and changes in the environment, social expectations, technical improvements and theoretical research findings, the tests definitely become out of date and therefore invalid, which means they must be changed eventually. Whatever the relevance of a test is at time, given that it is the result of the surrounding context analysis at that time, at time t+x, the surrounding context will have changed and the test will not be in agreement with this context anymore. This phenomenon can be observed even when the test focuses on objects as difficult to evaluate as intelligence that would be thought as invariant. The history of intelligence tests shows their evolution throughout time, not only due to the various new conceptions of Intelligence, but also due to the fact that tests (for example texts on which items are developed) which are supposed to measure it are growing old and tend to become obsolete.

II Validity in the French adult surveys

A) Background: Recent adult surveys in France

We are going to examine how those three categories of methods are used to evaluate validity in adult competency assessment surveys carried out in France in recent years, bearing in mind that the validity of an evaluation tool reflects its capacity to measure whatever such a tool is designed to assess.

The three main surveys carried out in France on adult populations in recent years are IALS in 1994, the JAPD (The defence preparation call-up day) every year since

1998, and IVQ (Information and Everyday Life) in 2002 and 2005 in mainland France, extended to French oversea territories in 2006, 2007 and 2008.

Despite differences, all those surveys have a common goal: producing figures in order to estimate the extent of the illiteracy problem within the adult population. In the 90's, what could be called « the dimensioning of the illiteracy phenomenon » became obvious to decision makers as being crucial in order to justify the construction and implementation of their public policy, and to estimate the efforts required to confront the problem.

B) IALS (International Adult Literacy Survey)

In 1994, the OECD survey seemed to be a way of producing reliable figures. However the results - 40% of French adults of working age were below the study's lowest level, considered as being one of functional illiteracy – led France to withdraw from the survey before the results were officially made public. To justify their withdrawal French authorities claimed that the percentage of illiterate people determined by the Survey did not correspond to the reality. In other words they criticize the validity of the survey. In particular, they criticized the instrument and it translation, the sampling, the way the test was administered, etc …

The same year another study, albeit declarative, published by the INSEE claimed that 5.4% of French people over the age of 18 had difficulties in speaking, reading or writing everyday language. Only 0.8% of interviewees admitted to not being able to read.

However, this unfortunate episode did mark the end of the declarative surveys and

led on to studies using standardised tests.

C) Individual based survey. The 'Defence Preparation call-up day' (JAPD)

The Army was the first to react by giving reading tests to conscripts. First of all in 1995 (334,721 conscripts tested) and then from 1998 with the 'Journées d'appel de préparation à la défense' (JAPD) which enabled measures of reading competencies to be applied to a whole age group (young people between 17 and 19 years old) with no distinction of schooling and using tests developed by the national education Ministry.

The JAPD offered the young people identified as having literacy difficulties the opportunity to meet an adult and talk about their future, receive information and careers guidance, and enter a support process led either by the national education Ministry or by other bodies, depending on whether or not the person had finished schooling. The tests taken during the JAPD also supply national and local statistics which are used as a reference for public authority action in relation to young people. In 2006 11% of the nearly 800,000 youths who took the tests showed some difficulty in reading (4.8% very serious – functional illiteracy – and 6.2% quite serious). But in recent years, a dispute has arisen within the Ministry of education between the directorate in charge of assessment and the directorate in charge of teaching. The latter criticizes the validity of the test results, arguing that many testimonies showed that a lot of students tested as illiterates have in reality no real problem with reading and understanding printed texts.

D) The 'Information and everyday life' survey (IVQ)

In 1998, after refusing to participate in the new OECD survey, the French administration decided to launch its own national quantitative survey designed to measure the writing performance of the adult population living in France. To build this survey of 'Information et Vie Quotidienne' (IVQ), the authorities relied upon the know-how of their statistics services, the qualitative studies carried out in the GPLI period and the skills of the ANLCI. The IVQ survey supplied statistical outline data (in 2002 and 2005) on illiteracy and the competencies of individuals (including immigrants) faced with written or oral communication tasks in standard French, as well as some indications of skills in calculation and everyday life.

The national survey « Information et vie quotidienne » (Information and daily life) has been a success in France. Its results are referred to and used by all the political or administrative decision makers at national or local level, as well as by most of the academics and by the vast majority of non governmental organisations. Currently, there is no real debate about the validity of IVQ. Despite such unanimous recognition; the questions of whether the tools and statistical thresholds used in the IVQ reflect the phenomenon of illiteracy validly, or whether the test used actually enables a saound understanding of the command of writing skills in standard French are far from being elucidated. In other terms, one must certainly make a significant point out of the question of assessing the validity of the tools used within the IVQ, and particularly assessing such validity as regards the tools meant to identify people facing difficulties in writing.

III) Testing validity in the French adult literacy surveys

A) Internal methods

As in all surveys the lower level of validity control is to verify the statistical consistency of the survey results. That has been done for all the surveys. In addition IALS offers an IRT analysis of the item pool.

IALS, IVQ and the JAPD have set up expert groups to estimate content validity, construct validity and criterion related validity. If the process is similar for the 3 surveys, we can point out some differencies. The composition of IALS expert group was large and international. JAPD expert group was built around a/one scholar and was composed mainly by representatives of the Ministry of education (direction for evaluation and forcast) and the Ministry of Defence (department of military service). The IVQ expert group is bigger and its members come from broader horizons, with representatives of ministries such as Social Affairs (department of statistical studies), Education (direction for evaluation and forcast), Culture (department in charge of French language), but also scholars from 4 universities (Lyon, Rennes, Paris, Nanterre), researchers from the national Institute of Demographic studies, specialists from the National Agency for Fighting Illiteracy, and of course from the National Institute of Statistical and Economic Studies which is head of the group. At least in IVQ and IALS experts have been helped in their decisions by the results of field tests which produced figures about items and tests. Those field tests where particulary numerous in IVQ.

B) Comparison methods

IALS has been compared to YALA in the USA, but in other countries it was not possible to follow the same process because IALS was the first survey of this type. In France IALS has been repeated in 2002 in a subsample of the IVQ methodological survey (547 respondents). The study showed significantly different results from the orginal ones (1998). In Prose (the only literacy dimension tested in 2002 with 13 items) the French population classified at level one was divided by 3 (2.73), dropping from 41 percent down to 15 percent!

Thanks to the 2002 methodological survey, comparison between IALS and IVQ is possible and has been done by INSEE (Murat, 2004). Nevertheless, when you want to check the validity of a survey by comparing it to another survey you have to suppose the one chosen as reference is itself valid. In the 2002 context, and because of longitudinal comparison of IALS between 1994 and 2002, it is not possible (for French scholars at least) to consider that IALS is in the French context a reference for the assessment of literacy levels, especially for low levels of literacy. OECD itself aknowledged this situation by recommending in the new adult literacy survey (PIAAC) the development of a specific measure for low levels of literacy.

Another remark could be that the comparison is difficult because the principles underlying each survey were different. IALS tried to measure interviewees' literacy (in fact reading proficiency and text comprehension) and, by establishing means, to measure the average literacy level of a population. The second survey tried to

indentify people facing difficulties regarding reading tasks, and to measure the text comprehension competencies of those who could be considered as mastering the minimal level of proficiency needed for life in a developed country like France.

The difference in approach is very important to the validity testing through comparison. For the measurement approach the first question is: do the tests (surveys) provide the same evaluation of literacy level (in fact same figures) for a common target population? If the results are significantly different there is a problem. But if the figures are the same or if they are not significantly different, maybe there is still a problem, because the apparent similarity of results can be an effect of the mean masking huge differences in the measurement of the competencies for each individual (member of the sample). In the second situation, obviously the first question is: do the two tests identify the same person as corresponding to the common target criteria (predefined profile). In other words in the first situation the question is: are the figures related to the target population identical with the two instruments, the two surveys? In the second situation the question is: if somebody is identified as being at a particular level with test A is s/he identified at the same level with test B? Technically, the question is: how often do the two tests produce the similar identifications? This last approach reproduces exactly the original methodology used in psychology for testing Validity. It is far more demanding than the first one.

Of course, things are not so simple, because in surveys like IALS, PISA, ALL, (based on the measurement approach) the target population is split into

competence groups. Each person, according to his/her own score is identified as belonging to a particular group which is determined by a score range statistically established. Nevertheless in practice the comparison is conducted on measurement and not on group repartition, because correlations are technically higher in the first case (continuous variable).

In 2002 the comparison between IALS and IVQ was limited to measurement comparison between two samples because interviewees had to answer only one of the two questionnaires.

In 2008 the ANLCI carried out a comparison survey between JAPD and IVQ. The principle was to ask young adults selected by Defense literacy tests to go through the IVQ process. In this case, on the same day at the same place, each member of the sample has been identified by the two instruments, and has ftaken both tests. The two tests took about 30 minutes each, and were administred one after the other, with a two to three hour gap between them. The first results showed a flagrant lack of correlation, or very low correlation on some specific points.

This acknowledgement of failure shows that the goals of the tests have to be revisited, maybe modified. The results of this comparison survey don't give any information about the validity of each test, because the testing can not decide on which test is the reference, in fact which test is valid.

C) External validity:

To compare test results and observations is probably the most reliable way to check the validity of a survey or a test, but also the most demanding for test developers.

IALS has been used in specific studies and tested on specific populations mainly in the USA, but never in France.

1) JAPD: Between 2004 and 2007, the department in charge of teaching content and teachers (DGESCO) within the French ministry of education has carried out a validity survey concerning JAPD tests in the whole of mainland France. According to the results, about 12% of French young adults (17-19) tested in JAPD are identified as presenting severe or moderate difficulties in reading tasks (Rocher, . Because practically all young adults are tested, the number of "detected" young adults is very high (about 90 000 a year). For all of those still attending school, a letter is sent to the school they attend. In 2006 just over 17 000 letters were sent. Following the mailing 25% of the young adults identified have benefited from a complete competence diagnosis carried out by "orientation advisors" (guidance) professionnally trained in to psychometric techniques and testing. Following the diagnosis 91% (in 2006) of the retested young adults were presenting no problems with reading and understanding simple texts[1]!

The situation looks very bad following this presentation, but in the field it is considered as very bad; teachers think that the ministry tests question their professional competences, especially those concerning student evaluation; local stakeholders, especially those working directly in the field, cannot decide how to conceptualise the action to remedy the problematic situation described by the test and contested by teachers. Of course, for statisticians the situation is not good

[1] *in 2004, letters sent,: 8741, in 2005: 11526, in 2006: 17181, young adults retested, 2558, 2935 and 4299, young adults without difficulties: 1623, 2257, 3898. from Ministry of Education, DGESCO, 2007 JAPD report*

either, even though if they can argue that the group of retested adults is not a representative sample. Nevertheless there are still doubts on the validity of the test.

2) IVQ

IVQ is probably not perfect. After the IVQ 2005 survey 200 respondents were revisited by a new interviewer team, including sociologists, to collect more information about their life (personal story) or about language competencies (using parts of IVQ tests translated in other languages such as Arabic, Turkish, Spanish, Portuguese, Polish, English, German), or about numeracy competencies, or about their vocational training path and attitude towards professional training). A few interviewers reported that some interviewees identified as facing difficulties with IVQ did not seem illiterate people. It is only just testimony not based on test results, and every specialist knows that in each survey a proportion of interviewees do not want to do their best or do not want to fully participate in the assessment process, but it encourages us to address more completely the external validity question.

In most cases, more than properly checking the validity of an instrument or a tool, we gather validity hints. These might either be statistical hints linked to the collected data, or external hints linked to the trait or the dimension to be evaluated or measured.

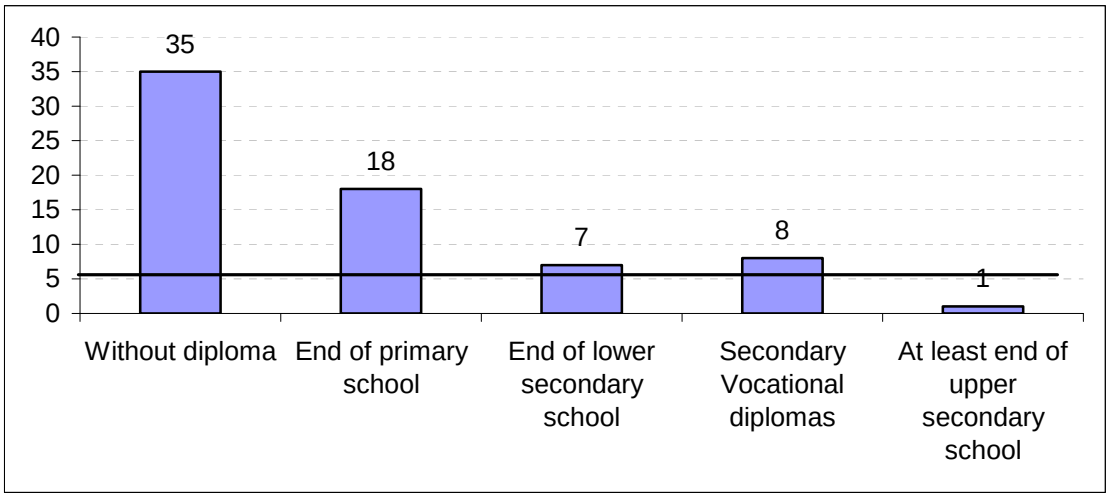
As far as education is concerned, available information often tends to be either on school results of the ongoing year or results at the end of the year (achieving a diploma, exam results, orientation, moving up to the next class or to a specific

course). A statistical "predictivity" index is then set, which would show to what extent the test predicted the obtained result.

Concerning tests for adult populations, a way of comparing survey (test) validity might also be comparing the survey results to available information. However the previously described education situation seldom happens. The only available information of that kind usually concerns either people at school level, or those with a professional qualification or who have passed diplomas. Because of the changing nature (improvement or loss) of written competences (by adults and even pupils), it is difficult, if not impossible, to check on the predictive nature of a test intended for adults, except for specific situations. Indeed, the diplomas to which test results can be compared are anterior to the test, and not the other way round as in the usual school test validation. The diplomas passed by someone a long time ago, might not have anything in common with the same diploma today. On the other hand, one can easily check on that inescapable evolution by trying to do exercises of the level one had reached in one's youth though the diplomas adults got might show that being good at writing (often) depends on meeting to job demands and generating good literacy skills. Yet, it is justifiable to compare test success percentages to the tested person's achieved diploma, even though, for the previously mentioned reasons, it is out of question to reach the statistical precision of the links that it is possible to build within the school context between school achievement and test results.

In IVQ 2004/5, the results of the interviewees have been compared with the diplomas declared by interviewees. As the following chart shows, we found that the results are logical. This could be considered as a first hint encouraging us think the test may be valid.

Chart : Proportion of interviewees considered to be illiterate according to their level of education (diploma)
In percentage %



Source : INSEE data file processed by author Scope: people aged 18-65 living in Metropolitan France and having begun their schooling in France Interpretation: 35% of the people having declared they have no diploma appear to be illiterate.
Note: The persons who are still studying have been included in the "At least end of upper secondary school" group.

Nevertheless, how can we explain that people that followed secondary school in France are still illiterate? Is the reason for this the failing of the school system or the inadequacy of the assessment instrument, or the way the test was administered to interviewees? Did interviewees lack motivation and implication? Is it the result of a lack of professionalism from interviewers?

As a validity study, IVQ could also benefit from a survey carried out by the French association for adult professional training (AFPA) involving 1035 trainees between September 2005 and March 2006.

The interest of that survey consists in the fact that it concerns people who had already been selected by psychologists working with the AFPA. Consequently,

information was available on these people through the results of the acknowledged standardised psychotechnical tests. The survey carried out by the AFPA consisted in giving a so-called orientation test first (actually a preselection test) to the members of the chosen trainee sample. Then all those who failed in this test; together with a few people who did not fail took the ANLCI tests in circumstances identical to those of the IVQ survey. These trainees have been monitored until the end of their training and therefore one owes information is also available on the outcome of their training. Thus it is possible to compare the whole data to the test results to collect validity hints.

The initial tests of the AFPA and the initial interview which goes along with them aim to establish the profile of the future trainee and to anticipate in which courses they would be most likely to succeed, taking into account both the candidate's skills and the inherent difficulty of the proposed training. The selection of the test method has been tried out for many years and it turned out to be very reliable since almost 80% of the trainees on average obtained a full degree or a certificate of professional competence. As a general rule, the trainees are recruited among job-seekers or among people who seek to obtain a professional qualification. The average age of the sample is over 32 years. First encouraging statement, not a single person who was training for administrative tasks has been detected by IVQ tests as facing difficulties. Most people detected by the IVQ as illiterate were concentrated, in builder's mate training courses (as far as men were concerned), and in electricity assembling training courses (as far as women were concerned). This is another

positive hint, because usually, these orientations are recommended by psychologists to adults with low level of literacy. In the same field, until the end of 2008, ANLCI will continue to gather validity hints by setting up a study in builder's mate training centres with the help of local authorities and European funds. In another field, a study designed to compare IVQ tests to the literacy tests used in Prisons is planned for early November 2008. In addition to comparison data this study will bring information about validity because tested prisoners will be followed for months by teachers from the Ministry of Education who will be able to collect information about their competencies by other means.

Conclusion:

If the « Information et vie quotidienne » (IVQ, Information and daily life) national survey has been successful in France and if its results are used as a reference by all political or administrative decision makers at national or local level, it is probably because IVQ results seem valid to them and to most of French experts.

Nevertheless, in France, questions on the reliability of the figures and about what figures represent, still exist concerning international or national studies. The increasing number of studies in the literacy field or in the education domain will boost the publication of diagnosis which can be very different from one another. These concerns are very sensitive in France in the case of low level of literacy because according to 1998 law, the fight against illiteracy is a national priority which involves a lot of partners who are in capacity to carry out their own studies, and to produce their own figures. The development of international studies can increase

the confusion if the results are very far from national studies or national reality. For all these reasons we have chosen to develop a network of literacy assessment, linking all studies with one another. We have chosen the IVQ national study as reference for the system. To substantiate this choice we have to challenge the question of IVQ validity.

Even though we presented a few hints which can back up the IVQ validity, we have to carry on collecting this kind of hints and organising validity studies as often as possible. This could be difficult and costly because testing people with low levels of literacy implies for us a face to face interview carried out according to psychological methods, if possible, but it is the only way to guaranty a good validity to the IVQ test. The efforts to establish validity using external methods could allow other French test makers to use a comparative method which is easier to implement and which will counterbalance the power of internal validity established by statistical methods or by expert reports.

Bibliography and references:

Brooks, G, Heath, K, and Pollard, A. 2005. Assessing adult literacy and numeracy: a review of research instruments (research report). London; National Research and Development Centre for adult literacy and numeracy.

Danish Technological Institute. 2004. Defining a Strategy for the Direct Assessment Skills, European Commission, Leonardo da Vinci Programme.

Dany Laveault and Jacques Grégoire, 2002, Introduction aux théories des tests, [Introduction to the tests theories], De Boeck, Bruxelles

Given, Lisa M., 2008. The Sage Encyclopedia of Qualitative Research Methods, University of Alberta, Canada, Sage inc.

Jeantheau, Jean-Pierre. 2005, Assessing low levels of literacy: IVQ survey 2004-2005, *Literacy and Numeracy Studies*, Vol 14, number 2 2005; 75-92

Jeantheau, Jean-Pierre. 2007. Low levels of literacy in France. *Alphabetisierung und Grundbildung* Band 1: 41-58. Münster(New York): Waxmann.

Manson, Emmanuel J and Bramble William, 1989, Understanding and Conducting Research, McGraw-Hill Inc, USA

Murat, Fabrice. 2005, Application de la méthodologie IALS aux données de l'enquête IVQ, [IALS methodology applied to IVQ data], Actes des journées de méthodologie statistiques 2005, Paris, INSEE
National Research Council. 2005, Measuring Literacy, Performance Levels for Adults, national academies press, Washington D.C.
NCES. 1998, Adult Literacy in OECD Countries, Technical Report on the First International Adult Literacy Survey, Washington
Rocher, Thierry. 2007. Les évaluations en lecture dans le cadre de la JAPD en 2006. [Reading assessment in JAPD]. *Note d'information* 07-25, Paris: MENESR
Trochim William M.K. 2006. Research Merthods Knowledge Base, http:// www.socialresearchmethods.net