

Final Report

# Defining a Strategy for the Direct Assessment of Skills

Funded by the Leonardo da Vinci Programme  
European Commission



Contract No. 2003-0541/001-001 LE2 52STAT

# **Defining a Strategy for the Direct Assessment of Skills**

Funded by the Leonardo da Vinci Programme  
European Commission

Jens Henrik Haahr, Hanne Shapiro, Signe Sørensen, Danish Technological Institute  
(managing partner)

Cathleen Stasz, Erik Frinking, Christian van't Hof, RAND Europe

Francis Green, Ken Mayhew, Rosa Fernandez, SKOPE

November 2004

## **Table of Contents**

<b>1. INTRODUCTION.....</b>	<b>4</b>
<b>2. SUMMARY, CONCLUSIONS AND RECOMMENDATIONS .....</b>	<b>8</b>
2.1. Summary of Findings.....	8
2.2. Practical Choices and Options .....	13
<b>3. THE POTENTIAL AND LIMITS OF A EUROPEAN INITIATIVE .....</b>	<b>20</b>
3.1. The Need for a European Initiative for Assessing Adult Skills.....	20
3.2. The Limits of Adult Skills Assessment.....	23
<b>4. THE FOCUS OF A EUROPEAN SKILLS ASSESSMENT INITIATIVE.....</b>	<b>25</b>
4.1. Political Priorities: Which Skills Are in Focus? .....	25
4.2. Informing Policy Making: Lessons from Previous Efforts.....	28
4.3. Political Priorities and Possibilities .....	30
<b>5. METHODOLOGICAL ISSUES .....</b>	<b>32</b>
5.1. Ways of Measuring Skills in an Adult Population.....	32
5.2. The Data-Collection Process.....	38
5.3. Methodological Issues Concerning the Skills Domains .....	40
5.4. Instrument Design Process.....	44
5.5. Conclusion .....	45
<b>6. REFERENCES.....</b>	<b>46</b>
<b>ANNEX A: CASE STUDIES OF SELECTED ADULT SKILLS ASSESSMENT INITIATIVES .....</b>	<b>48</b>
<b>NAAL AND NALS.....</b>	<b>49</b>
<b>ADULT LIFE SKILLS AND LITERACY SURVEY (ALLS).....</b>	<b>55</b>
<b>INTERNATIONAL ADULT LITERACY SURVEY (IALS).....</b>	<b>62</b>
<b>IVQ – INFORMATION ET VIE QUOTIDIENNE.....</b>	<b>73</b>
<b>DIALANG .....</b>	<b>80</b>
<b>LAMP: LITERACY ASSESSMENT AND MONITORING PROGRAM.....</b>	<b>89</b>
<b>THE 1997 AND 2001 UK SKILLS SURVEYS .....</b>	<b>98</b>
<b>THE SKILLS FOR LIFE SURVEY .....</b>	<b>105</b>
<b>THE “ASSESSING LEARNING TO LEARN”-INITIATIVE.....</b>	<b>110</b>
<b>THE DANISH NATIONAL COMPETENCE ACCOUNT .....</b>	<b>119</b>
<b>TEST OF WORKPLACE ESSENTIAL SKILLS (TOWES).....</b>	<b>127</b>

## 1. INTRODUCTION

In November 1995, the European Commission adopted the white paper on “Education and Training: Teaching and Learning, Towards the Learning Economy” (Commission 1995). This white paper, which took forward the earlier white paper on “Growth, competitiveness, employment” and its emphasis for Europe to invest in knowledge (Commission 1993), defined a new set of European ambitions in the field of education. It emphasised, for example, how investment in training and capital investment should be treated equally, that schools and the business sector should be brought closer together, and that the acquisition of new knowledge should be encouraged.

The Lisbon Strategy, adopted by the European Council in 2000, placed new emphasis on knowledge, education and training. The European Council set itself a new strategic goal for the upcoming decade: to become “the most competitive and dynamic knowledge-based economy in the world capable of sustainable economic growth with more and better jobs and greater social cohesion” (European Council 2000). This, in turn, requires a strategy which, among other things, supports the transition to a knowledge-based economy through

- adapting Europe’s education and training systems to the demands of the knowledge society, making them “a world quality reference by 2010” (European Council 2002)
- ensuring “a substantial annual increase in per capita investment in human resources”
- defining in a European framework those new basic skills that are to be provided through lifelong learning: IT skills, foreign languages, technological culture, entrepreneurship and social skills.

The Lisbon strategy has two main goals. The first is to enhance economic competitiveness through improvements in human capital. Skills, knowledge and competencies are increasingly seen as crucial prerequisites for productivity and competitiveness. The European economies are increasingly confronted with a dual challenge. They face global competition not only from developing countries with cheap and plentiful labour, but also from high-productive, high-tech economies of North America and the Far East.

The second is to promote social inclusion. In the view of Lisbon, competitiveness should be achieved “with more and better jobs and greater social cohesion”, and not at the cost of greater inequality or social marginalisation. A dynamic and competitive economy should benefit all, and the entire European population must be involved in and benefit from reform and development. “The knowledge society” is a society of not only full employment but “all-employment” (Threlfall 2002). It is “an information society for all”--one in which “every citizen must be equipped with the skills needed to live and work”, where “info-exclusion” and illiteracy must be prevented, and where special attention is given to the disabled (European Council 2000).

Within this context, the assessment of adult skills has come to be seen as both politically and economically relevant. Adult skills assessments may help decision makers in government and business, at European and Member State levels, to take stock of their human resources, and to adjust their “human capital investment” accordingly. Assessments may also enable policy makers to gauge the returns to the Lisbon strategy’s hoped for increases in human resource investment and to take any necessary corrective steps. This stock-taking is especially important with respect to the new skill requirements that are believed to be in greater demand in the “knowledge society”. In

addition to basic literacy and numeracy, the economy may require ICT-related skills as well as entrepreneurial, self-management, and learning skills, to name a few.

Adult skills assessments also potentially open up possibilities for strengthening the accountability of the education and training sector by providing one indicator of whether education and training institutions deliver on their promises. This would include not only compulsory schooling, but also opportunities for learning over the lifetime.

### **Promise of International Comparison**

In relation to both of these ambitions, international adult assessment initiatives hold particular promise. As we shall return to shortly, systematic international comparisons have proved themselves efficient in stirring interest, stimulating debate, and affecting political decisions and priorities. They have also been effective in promoting mutual learning and the exchange of good practices. Not least, international comparisons appear to have been rather powerful tools for initiating reforms within the education sector.

### **‘Education and Training 2010’**

‘Education and Training 2010’ is a set of EU activities which have been set in motion to help achieve the strategic objectives of education and training system reform. In 2001, the Council adopted a set of three overall and thirteen associated concrete objectives to support the Lisbon goal. A number of these objectives are relevant in an adult skills assessment context: Increasing numeracy and literacy, maintaining the ability to learn, improving ICT skills, developing the spirit of enterprise, and improving foreign language learning. The objective of ‘making the best use of resources’ is also highly relevant when considering adult skills assessment measures.

In 2002, a work programme was developed to realising these objectives. Subsequently, twelve different working groups, comprised of stakeholders and experts, have been working on one or more objectives of the work programme, for example by supporting the implementation of the objectives for education and training systems at national level through exchanges of good practices, study visits, and peer reviews.

One of the twelve groups, the Standing Group on Indicators and Benchmarks, has focussed on developing indicators to monitor progress on the work programme’s specific objectives. In July 2003, the Standing Group presented a list of indicators to support the implementation of the work programme, and suggested the development of several new indicators, including indicators for language competencies and learning to learn skills, ICT skills and indicators on social cohesion and active citizenship (Standing Group on Indicators 2003). Another standing group has focussed on basic skills. In November 2003, the basic skills working group presented a report which contained proposals for definitions of essential basic skills in eight domains (Working Group 2003).

The so-called ‘Copenhagen process’ is a third set of activities that are relevant in the context of adult skills assessment. With the Copenhagen-declaration, the EU Ministers for Vocational Education and Training (VET) formulated a set of objectives for cooperation in VET, within the broader framework of the Lisbon Strategy and the ‘Education and Training 2010’ Work Programme. Among other things the Copenhagen Declaration calls for common principles for the validation of non-formal and informal learning to help ensure greater compatibility between approaches in different countries and at different levels.

The Commission presented an interim evaluation of the implementation of the Education and Training 2010 programme, and with the Council, developed a joint report for the Spring 2004 European Council (Commission 2003a). The interim evaluation highlighted that too many young people fail to acquire key competencies, that too few adults participate in further learning, and that a language proficiency indicator had not yet been developed.

### **Our Approach: A Broad Perspective**

Within this policy context, this project on ‘defining a strategy for the direct assessment of adult skills’ addresses the possibilities and challenges that must be confronted in pursuing a European adult skills assessment initiative. We began our project with a review of a large number of recent initiatives for the direct measurement of adult skills, including recent efforts that are still in progress. This review comprised the report for the first phase of the study.

In phase 2, the study team more closely inspected and evaluated the most relevant of these initiatives (11 case studies are enclosed in Annex A). We have conducted a series of interviews with relevant experts and users of initiatives at different levels (in government, international agencies and in the training and education sectors), both in relation to the analysis of specific skills assessment initiatives and in connection with the more general professional and political debate about skills assessment in contemporary society. We also discussed our findings and preliminary conclusions with the European Commission.

From an initial exclusive focus on methodological and practical issues, we adopted a broader study perspective to include questions such as the ethical limits of government, the political feasibility of different types of adult skills assessments, the question of acceptance on a European scale, and various economic and social risks involved in carrying out a comprehensive European adult skills assessment initiative.

While methodological and practical implementation issues remain at the forefront of our study, and necessarily must do so if the result is to be relevant and useful for developing a coherent strategy, the stakes are too high to limit the study to methodological exercise. An international adult skills assessment initiative, instigated by the European Union and its Member States, holds certain promises, but its implementation also presents a number of problems, challenges and potential risks. These challenges and risks should be openly presented, discussed and addressed by the Commission and the Member States before any initiative is launched. This report represents a step in that direction.

### **Structure of the Report**

Following this introduction, Chapter 2 presents a summary of findings and a set of conclusions and recommendations. The more detailed analyses that support these these conclusions and recommendations are presented in successive chapters.

Chapter 3 discusses the need for a European initiative for assessing adult skills and addresses the limits of an assessment in broad terms. A European adult skills assessment initiative must inform policy. Chapter 4 addresses issues of political goals and political feasibility, asking how a direct skills assessment initiative can be designed to inform policy making in the best possible way.

Chapter 5 turns to methodological issues and addresses the overall question of how a European adult assessment initiative can be best designed to produce valid and reliable results. Dealing with such issues is important as the Member States can only be expected to support direct adult skills assessment on a European scale if they feel confident that results will be credible and can stand up to scrutiny and criticism.

Annex A presents a number of case studies of the most relevant skills assessment initiatives.

## 2. SUMMARY, CONCLUSIONS AND RECOMMENDATIONS

This chapter presents recommendations and options for a European adult skills assessment. The chapter first summarizes the findings, organised as three issues: purpose, content, and methods. It then presents a number of options that should be addressed in developing an operational strategy for the assessment initiative.

Overall, the study concludes that a European assessment of adult skills is both feasible and policy relevant. If the European Commission accepts this conclusion, we further recommend that before any technical discussions begin it must first initiate a dialog to reach broad policy agreement as to the purposes of the initiative and to establish the collaborations needed to carry it out.

### 2.1. *Summary of Findings*

#### **What is the purpose of a European adult skills assessment?**

There are a number of arguments that support the need for a European initiative for adult skills assessment, especially in the context of the Lisbon strategy.

- It is commonly agreed that skills, knowledge and competencies are becoming more important in modern societies: inadequate levels of skill can impact negatively on the European Union's economic development and social cohesion.
- Lifelong learning—learning outside of formal education or training—is an important policy objective for meeting the demands of the “learning economy”. However, little is known about what learning is needed, who needs it, or what is learned through participation in non-formal learning activities.
- Systematic information on the outcomes of formal education systems in equipping citizens with the necessary skills, given different levels of investment is incomplete.
- Educational attainment is only a rough estimate of skill and therefore has limited uses for understanding the relationship between skill and various economic outcomes or for developing policy interventions. For example, it does not help us to understand issues of skill utilisation in different sectors and contexts and how and why rates of utilisation might differ.
- International comparisons of direct measures of adult skills have proved useful for influencing public debate and political agendas and indeed in initiating policy reform. The EU's voluntary system of open coordination in connection with the Lisbon strategy sets the stage for using comparative data more broadly to review and assess training and education policies of the Member States vis-a-vis European objectives.

While these and other arguments stand in support of a European initiative, there are also some risks and challenges to consider. Skills assessment is not value neutral, and any EU-wide effort must directly face a number of political, ethical and methodological issues. Some broad issues include the following:

- The concept of “skill” can be contested in the definition of jobs (e.g., between employees and employers or between genders) and has been evolving in public discourse. A term that once referred primarily to job-specific high-level analytic capabilities, technical abilities, or

vocational competences is now used for a wide range of generic competencies (e.g. problem solving, teamwork, communications), attitudes, and personal characteristics. Many of these recent additions to the “skill” label are difficult to define and measure.

- The need to quantify outcomes to assess progress toward the Lisbon Strategy or other policy initiatives tends to focus attention on what is measurable. Skills and competences that may be essential to reach economic and social goals (e.g., entrepreneurship), yet difficult to measure, may be overlooked.
- Skills associated with success along various economic and social criteria may be context dependent, yet context is often lost in standardised measurements. Any assessment will provide a somewhat distorted picture of the character and distribution of skills.
- Assessments are typically conducted on individuals, and data are aggregated to estimate skills in particular populations or groups. Some skills of interest to policy makers, however, are instantiated by groups, teams, or organisations and simple aggregation of individual scores is not sufficient to understand the composition of skills at higher levels.
- The broadening of “skill” to new domains risks taking policy into the realm of the personal. To what extent should governments be permitted to assess skills like problem solving, critical thinking, physical/health competencies, or aesthetic competencies?

### **What should be measured?**

The brief of this project was to consider output-based measures of skill rather than input-based measures of learning, such as formal qualifications or training episodes. The skills-related strategic objectives debated by various European Councils help to determine the content of an adult skills assessment, and also give an indication as to how useful the results will be for policy making purposes. Three relevant summits—Lisbon, Barcelona, and Stockholm—defined a number of skills as important. First, we can identify a set of skills for which standardised, reliable assessments are available or feasible:

- Literacy skills (prose, document)
- Numeracy skills (quantitative literacy)
- Problem solving/analytical skills
- Foreign language competency
- Job- or work-related “generic skills”
- Information and communication technology (ICT) skills

Of these, there are some limitations in direct assessment of the latter two categories—generic skills and ICT skills—but indirect assessment of these skills linked to work context is possible. These methodological considerations are discussed later in this chapter.

In the context of a European-wide adult assessment initiative, it is not presently feasible to directly assess some skills of interest to policy makers, for definitional, methodological, or perhaps political reasons. These include:

- Entrepreneurial skills
- Some social skills
- Learning-to-learn skills or the ability to learn

The results of this study indicate that any European initiative should *not* attempt to assess these skill types.

### **What methods and approaches should be considered?**

The study reviewed three broad approaches to skills measurement in an adult population - occupational titles, tests, and survey-based.

#### *Occupational titles*

The idea behind this approach is that if the skills required to be used by any “job” defined by its occupational title could be measured reliably and precisely, and if such measures were available for all occupations, then the skills of the employed part of the population could be assessed through a simple count of occupational employment, as obtained through a census or labour force survey. Unfortunately, neither of these two conditions are satisfied for various reasons: assessments are often unreliable, there is considerable variation in skill requirements within jobs, skill requirements change regularly, and there is as yet no common basis for comparing occupational composition across countries. This method is not part of a recommended strategy.

#### *Survey-based self assessments*

Survey-based self assessments are frequently used to measure skills. Self-assessment methods simply ask individuals to say “how good” or “how effective” they are at certain activities. The advantage of self assessment is that one can ask about a broad range of skills, including those that are difficult to measure directly (e.g. entrepreneurship, learning-to-learn). The disadvantages, however, outweigh the advantages. Self assessments are subject to self-esteem bias, may be unreliable, and are difficult to validate. Self-assessment would not command acceptance in any international assessment. Thus, this method is not recommended as part of a European strategy.

Individual self-reports linked to jobs, however, can provide valid and reliable estimates of skill within the context of the job. Respondents are questioned directly about the work activities and requirements of the job, and asked, for example, “how important” a particular skill is in the job. This method is less susceptible to social-esteem bias, and meets accepted validity and reliability standards.

There are two disadvantages to job-linked self reports. First, this method delivers measures of the skills actually used in jobs, rather than direct measures of the skills that individuals possess. As an indication of the skills of individuals, this method gives only approximate measures. For example, job-holders may have too many or too few skills for the job—if too many, the statistics will underestimate the jobholders skills. Further, the reliability of this method for similar job-holders in different countries would need to be evaluated for results to be comparable. A second disadvantage is that the job analysis method cannot be applied to the economically inactive. However, the use of this method for the employed part of the population (or those recently out of work) is a viable means of measuring some skills of interest to policy makers (e.g. ICT skills, work-related “generic” skills).

#### *Testing*

Direct testing is generally presumed to be the best method for assessing individuals’ skills, because tests are said to be objective. It is assumed that they do not suffer from potential biases that can come from dependent parties, especially the individual being assessed. They therefore start with a

presumption of superior reliability over other methods of assessment. This claim underpins the history of testing in the field of international skills assessments from the International Adult Literacy Survey (IALS) onwards, and helps account for the considerable impact of such assessments for policy.

However, testing presents both methodological limits and possibilities:

- tests are restricted to a narrow range of skills, and
- developing valid, reliable and comparable tests for additional skills will be costly.

In that respect, existing tests of literacy and numeracy (e.g., IALS, Adult Life Skills and Literacy Survey (ALLS)) are useful, and there would be little point in developing new assessments for a European assessment initiative. New methods are being developed that permit differentiation of low levels of literacy or numeracy to be assessed, and these can be adopted if there is particular policy interest in this area. Addition of any other skills is dependent on feasibility—length of testing time, context dependency, and ability to measure validly or reliably. Irrespective of content, an important lesson from the international studies concerns data collection processes. It is essential for participating countries to adopt strict guidelines for data collection, as compromises will threaten the usefulness of the results.

#### *An integrated approach*

There are several arguments for integrating different methods into a single assessment, rather than dividing tests and self-report portions into multiple sessions. Methodologically speaking, in order to properly investigate associations between skills that may be dependent on one another it is important to obtain individual scores in multiple skill domains. An integrated approach is also more cost effective than one requiring two separate data collections. While an integrated approach reduces costs, it also limits the testing time—an average of 90 minutes for the entire assessment is a good rule of thumb. Even within a single-visit assessment, however, it is possible to use matrix sampling techniques in which different groups, for example, receive a core test domain (e.g., literacy tests) but different combinations of tests or surveys to assess additional skills. It is still conceivable to contemplate a separate data collection process for skills of interest to policymakers that are not expected to be highly correlated with other skills. These types of tradeoffs need to be considered as any strategy develops.

Our review indicates that it is feasible to combine direct measures and self reports based on job context within a single assessment battery, carried out by interviewers in the respondent's home.

#### **Ethical and Political Issues**

Some risks in developing a European strategy for adult skills assessment are more political in nature. The study identified a number of issues, some of which arose as a result of experience with other national and international efforts.

#### *Accepting the Limits of Skills Measurement*

The review indicates that many skills of interest to policy makers cannot be defined or assessed in standardised ways and therefore are not good candidates for an adult skills assessment. These limitations will need to be openly discussed and understood so that the methodological constraints are clear. It may be judged that more effort needs to be put into developing suitable assessments, and this decision could affect the overall planning and timing of the strategy.

Even when skills can be measured to an acceptable standard, other limits have been noted and may need consideration at the outset. The notion of “skill” has evolved to include personal characteristics or other concepts - there is some debate about whether new conceptions are even useful, especially when they are difficult to measure. The emphasis on the skills that are measurable, can fix the political agenda to value some skills over others and risk overlooking important skills that could be developed through policy intervention. Some new skills of interest (e.g. team work, communications) are context dependent - their content or value depends on their application in a particular situation - and this relationship to context is not easily captured with standardised assessment measures. Skills measurement is carried out on individuals, but skills also apply to groups or to organisations. Simple aggregation of individual scores will not reliably reflect the composition of skills at these higher levels.

#### *Uses and Usefulness of Assessment*

Results of international comparisons easily turn into political debate especially if those results are controversial. Experience with the IALS and other international assessments indicates that assessment results can be interpreted as embarrassments for governments, negative views of certain social groups, and important drivers for policy. The potential uses and reporting of any adult assessments will need to be debated early on, and public reaction to national comparisons will need to be anticipated for any strategy to move forward.

Of particular importance here are degrees of interpretability and comparability. Some national initiatives, for example, attempt to evaluate a range of skills using self-assessment methods, which have questionable validity and reliability. In addition, the results from such measures may lack a common reference point from which relative performance can be absolutely judged. This limits their usefulness for policy purposes, as it is impossible to know whether a high, low, or average score is good or bad.<sup>1</sup> Usefulness improves when measures are theoretically-based and the theory defines difficulty levels. Results can then be associated with difficulty levels and sound judgements can be made about different levels of performance. Theoretically-based measures, which are then empirically supported through the assessment process, provide the best means of interpreting and comparing the scores. Even when measures are both theoretically-based and empirically verified, for example in the IALS or the US National Adult Literacy Survey (NALS), defining levels is a tricky, technical problem.

Finally, a useful assessment at the European level should have policy relevance. This consideration might eliminate the necessity or desire to assess some skills (e.g., social competence or self-management competence) as the implications of knowing levels of those skills in the population are unclear. Or perhaps the policy problem can only be informed through a longitudinal study, in which case the inclusion of a specific skill measure on a population based national assessment would not yield useful results. These issues of relevance for policy purposes need to be openly discussed and debated.

#### *Ethical and cultural issues*

Even if there were no scientific problems to resolve, assessment of some skills, especially those that move into the personal domain (e.g., social skills, learning-to-learn), may be controversial simply because their assessment brings them into the sphere of international comparison and public policy.

---

<sup>1</sup> For example, if a self-report assessment finds that Greeks have “higher” social scores than the Dutch, but the Dutch have “higher” communication scores, is this a problem for the Greeks, the Dutch, or both? If so, why?

When personal characteristics are reformulated as skills, the implication is that such characteristics are malleable. Therefore, it may be possible for policy to intervene to change these “skills”. This intrusion of policy into the private domain may be threatening to some. Their inclusion may raise questions about legitimacy that would need to be fully debated in advance. Any controversy after the fact might limit the usefulness of the measure, for example if it makes some analyses politically charged.<sup>2</sup>

## **2.2. Practical Choices and Options**

In defining a strategy for direct assessment of adult skills at the European level, the study considered several issues: policy purposes, political feasibility, and methodological approaches. This section identifies the kinds of choices that need to be made and presents some recommended options to consider based on our review.

### **Assessment Objectives**

As discussed earlier, there are a number of arguments for launching a European-wide initiative to assess adult skills. The issues need to be debated and consensus about objectives should be reached before launching any initiative for at least two reasons:

- Member States have their own needs and priorities and their buy-in is essential for any European initiative to succeed.
- The rationale for the survey has important implications for survey sampling, design, and implementation

Three specific goals, all implied by the Lisbon strategy, could be considered:

- To establish indicators for the purpose of comparing the skills performances of EU countries
- To identify low level skills in adult populations across Europe, and among specific groups
- To improve understanding of how far skills (and which particular skills) matter for socioeconomic performance, both for individuals and for nations; hence to provide an evidence-based rationale for skills policies

The relevance of the first two goals is that basic skills are important for social inclusion, and therefore it is useful to understand the stock of skills in the populations across countries, and across population sub-groups. The main decisions to be made involve the types of skills to measure, and whether the sample should be representative of the population or focus on special groups. For example, if policy makers are most concerned with assessing skills for individuals with low levels of literacy, then the sample would target these individuals and might require special methods to detect variation in low-level skills. Alternatively, the interest may be in contrasting different age groups (or occupational groups), which would suggest a strategy to use successive waves of data collection focused on different groups. The main choice is whether to survey the whole population at once or whether to adopt a progressive strategy that will eventually cover the whole population.

---

<sup>2</sup> The Finnish initiative on assessing learning-to-learn, is a case in point. This assessment aimed at school children has been interpreted as an intelligence test, and for this reason the relationship between scores on the assessment and success in later life is not explored.

Addressing the third goal would require a different emphasis, but this need not be inconsistent with the achievement of the first two goals. To assess the impact of skills on individuals' economic performance requires the measurement of multiple skill domains, and good background information including data on pay and employment. Ideally, a longitudinal study in some countries would produce more robust findings at this micro level.

To assess the impact of skills at a national level, and with a single cross-section, it is best if the sample represents a more heterogeneous group of countries. It is important for the sample of countries to vary, for example, with respect to skill quality and distribution and investments made in education in order to tease out the relationship between skill and socioeconomic outcomes. Findings would be made much more robust with the repetition of the survey after intervals of between five and ten years, which would allow changes in the national skills base to be assessed and compared with national economic performance.

One might consider also the possibility of analyses at an intermediate level, that is, the industry or sector, in which links between industry productivity and skills could be investigated. Unfortunately, this option would require much larger survey samples in each country, in order to ensure sufficient representation of enough industries. At a conservative estimate this would require a doubling of the sample sizes necessary to achieve the other goals of the assessment. We do not recommend this course, on grounds of cost.

Prior experiences with international surveys indicates that reaching a consensus about purposes can be difficult, as countries do not always know what kind of information they need or would like to have and as any initiative must serve both international and domestic purposes. Some countries may even be sceptical about the value of surveys or of their use for assessing skills. Nevertheless, following the objectives of the Lisbon strategy, we recommend that the skills assessment strategy embrace all the goals identified at Lisbon.

### **General Implementation Issues**

The implementation phase poses the biggest challenge to validity, reliability, interpretability and comparability, and is the single largest source of error in international assessments. Respecting the autonomy of the involved parties, management of data collection will need to be closely coordinated, so that consistent oversight over methods can be assured. Strict guidelines, standards and quality assurance mechanisms must be put in place with respect to sampling and key implementation activities (e.g., pre-collection documentation and review of procedures, education processes to ensure that national study teams understand the intent and detail of activities, post-collection analysis and certification procedures for national study teams and the international consortium). Adequate forms of probability sampling must be deployed in every country.

To achieve higher response rates, procedures for repeat visits to secure interviews with selected respondents must be outlined. Techniques for improving response rates should be considered, including payment of respondents, as studies indicate that even a small payment (e.g. 15 Euro) can improve response rates and even lower average overall costs.

Standardised international assessments are expensive, partly due to costs of overheads (management, design, statistical analysis throughout the project). The IALS experience suggests that contributing financially to the undertaking has a salutary effect on countries' commitment. Participants in the ALLS, for example, are responsible for financing their own national item pilot

and the work of any experts they involve. They also contribute roughly \$150,000 US toward international overheads for the pilot and main study over a three or four year period.

### *Timing*

Given the likelihood of discrepancies between policy priorities and methodological feasibility and the uncertainty about shared objectives and financial commitment, implementation must first involve political agreement on content, purposes, and the like. Once this agreement has been reached, it would likely take another 3-5 years at minimum to field any assessment. The timing will also be partly dependent on the resources and level of staffing committed.

Any short-term strategy would necessarily limit the domains that could be assessed to those areas where reliable, valid, and comparable measures already exist. As a result, a short-term strategy will be unlikely to inform policy makers about those policy issues that have only recently appeared on the political agenda. Developing methodologically acceptable assessments of some skills that are currently of interest to policy makers (e.g., learning-to-learn, entrepreneurship) would take several years to accomplish. Although there are ongoing initiatives to grapple with assessment in new domains, such as DeSeCo, the final development of these assessments is still a long way off.

It is difficult to comment on the longer term. New developments are underway, for example with respect to measurement of other skills or in the use of adaptive, computer-based approaches, which might eventually be adopted for a European initiative. However, rather than wait for new, sometimes very speculative developments, we recommend an approach that reflects the current state-of-the-art. As the initiative moves ahead, new developments should certainly be monitored. We envision that the initial assessment strategy is not a one-off, but the beginning of a longer-term effort that will necessarily develop over time.

### *Collaboration*

Collaboration with other recent or ongoing initiatives is strongly encouraged in order to avoid duplication of efforts. In addition, collaboration might allow an initial limited scope of the assessment to be broadened and thus to include more skills or skills domains. The European Commission should consider taking a lead on ensuring cooperation with ongoing OECD efforts. Two relevant efforts include the ALLS, currently in the field, and the development of an OECD strategy for the assessment of adult competencies. The latter effort is in its beginning stages and aims to field an assessment by 2010. An expert group has been formed, had its first meeting in April 2004<sup>3</sup> and is set to meet again in Stockholm in November 2004.

It will be important for any European initiative to establish a high level, experienced technical working group to carry the work forward. This group should include individuals who have direct experience in designing and implementing international, household based assessments. Existing documentation from the IALS experience will be an important source of recommendations and lessons learned that any working group may wish to review and consider.

Our recommended approach (discussed further in the next section) includes use of items from existing international (IALS, ALLS) and national (UK Skills Survey) assessments. The Educational Testing Service (Princeton, NJ, USA) owns the IALS, and would need to be consulted as to their use.

---

<sup>3</sup> The draft phase 2 report from this study has been circulated to this group.

Statistics Canada/the Queen in Right of Canada owns the ALLS tests. Anyone may use the items from either assessment without royalty subject to the following three conditions: the use is for non-commercial ends, the findings are made available in the public domain, and the item pool is kept secure. Statistics Canada/ETS would determine which items are retained for the purposes of establishing trends in further assessments and the balance of items are placed in the public domain. This means that the European Commission would be free to use the test as is or to adopt some part of it. Statistics Canada would be willing to consult as to which items are put into the public domain. The UK Skills Survey is in the public domain.

### **Approach: What and How**

It is assumed at the outset that any European initiative must strive for the highest acceptable standards of validity, reliability, comparability and interpretability. As discussed, achieving this may depend as much on implementation as on design choices—what to assess and how to assess it. Here we present some specific recommendations for consideration that aim to reflect acceptable standards.<sup>4</sup>

#### *What skills?*

Our review indicates that a subset of the skills of interest to policy makers meets the standards for inclusion in an adult assessment initiative: literacy, numeracy, problem solving, ICT skills, job-related ‘generic’ skills, and language skills. We recommend that all be assessed, as a concern for these skills is in accordance with the Lisbon agenda.

Literacy (prose, document), numeracy (quantitative literacy) and problem solving skills can be directly assessed, using the ALLS tests. There is no need to develop new assessments in these areas.<sup>5</sup>

ICT skills may be estimated through self-report survey methods.<sup>6</sup> Two good sources of items are the UK Skills Survey and the ALLS background questionnaire. The UK Skills Surveys assessed computing skills (the importance and complexity and computer usage) via self-report within the job context. These survey items will need further development to establish their usefulness in other countries. The use of these items would also be limited to the working population or to the recently unemployed. ICT skills (incidence, complexity, frequency) at work and in everyday life were also assessed in the ALLS background questionnaire.

Job-related ‘generic’ skills may also be estimated through self-report survey methods, through adaptation of the UK Skills Survey. This survey assesses importance and sophistication (complexity) of computing skills and also the importance of several other generic skills: literacy, numeracy,<sup>7</sup> technical know-how, high-level communication skills, planning skills, checking skills, client communication skills, horizontal communication skills, problem-solving, and physical skills.

---

<sup>4</sup> A fuller rationale for some of these recommendations is presented in the phase 2 report. Some choices, of course, may require extensive review of both technical and political issues by the working group and other responsible parties.

<sup>5</sup> The French IVQ assessment is also an option, but it might require further development for use in other countries.

<sup>6</sup> Educational Testing Service developed some performance-based tests of ICT skills for possible inclusion in PISA, but these were not included as the PISA/OECD team felt they were too expensive. These may warrant further investigation to see if they might be suitable for an adult assessment.

<sup>7</sup> Our recommendation is to retain the literacy and numeracy items in the self-report portion of the assessment as a useful comparison to the direct tests. Eliminating these items would not save much time, and some are needed anyway as an ingredient of a measure of communication skills.

As with ICT skills, further development will be required to ascertain usefulness of these items in other countries. Also, while these measures have the advantage of measuring work skills in their context, it will also have to be accepted that the measures of these generic skills of individual job-holders can only be obtained in this way as an approximation, depending on the match between job and the job-holder.

Foreign language skills may be assessed through direct tests. The DIALANG system<sup>8</sup> rests on a common framework for assessment of foreign language skills in twelve languages and is presently delivered free of charge through the Internet. It would need some further development to adapt it to the survey situation. The full assessment covers 5 domains and takes up to two hours, which is too long as a component of the type of integrated data collection that we advocate. Three alternatives, each with advantages and disadvantages, have been identified: a) assess one skill domain (e.g. vocabulary), b) assess three domains (listening, reading, grammar) in an hour of testing time, and c) carry out a separate assessment from the main skills assessment. Our current view is that the second option is possible, if used in conjunction with a matrix-sampling strategy whereby some individuals would be tested in language skills as a substitute for assessment in the other domains discussed.<sup>9</sup> There are examples of language tests in the private domain that could be further explored.<sup>10</sup>

#### *How: Design and Implementation*

##### Length, mode, and composition

As indicated before, our review indicates that the integrated approach is most viable. It would involve a one-stop assessment with an average length of 1.5 hours. The precise timing of the different components recommended above is open for discussion and will depend on policy choices, results of pilot testing and other considerations.

We recommend that all respondents receive a core set of items. This core should include items assessing at least literacy and numeracy and questions to gather background information. This core should comprise 30-45 minutes of testing.

We recommend that the background questions be considered a separable component from the direct tests and the self-report items based on job context. This self-report section might comprise items to assess generic skills only or a combination of generic and ICT skills. The background component would include demographic items--such as information on the respondent's work force participation, education and training, and literacy activities--which would need to be specified in the design phase.

We also recommend adopting a matrix sampling design that would allow assessment of all skills discussed in Section 2.2. In this design all respondents would receive the core set of items, but the remainder of the time could be distributed to assess different areas. For example, some respondents

---

<sup>8</sup> DIALANG is carried out with the financial support of the European Commission) is based on the Council of Europe's Common European Framework (CEF).

<sup>9</sup> There is a recent European initiative to assess foreign language skills of pupils at the end of compulsory schooling. One consideration is whether this assessment is seen as sufficient for policy purposes, thereby making superfluous an assessment of foreign language skills in the adult population.

<sup>10</sup> Only recently, more sophisticated (but proprietary) language testing techniques have been implemented in the US and Canada. These language tests use speech recognition systems to assess the fluency of a person's listening and speaking abilities within a much shorter time period<sup>10</sup>. If such techniques can become available at reasonable costs, the foreign language assessment could be easily included without crowding out other skills or without requiring additional assessments.

may have more language testing and others may have more assessment of job-related generic skills. Two examples follow:

**Example 1:**

Core items plus problem solving: 45 minutes  
Foreign language (one domain): 20 minutes  
Job-related generic skills/ICT skills: 25 minutes

**Example 2:**

Core items: 30 minutes  
Foreign language (2 domains): 40 minutes  
Job-related generic/ICT skills: 20 minutes

It may also be necessary to administer a screening questionnaire if, for example, the assessments include tests for individuals with low levels of literacy. A screening test would identify those individuals and direct the interviewer as to which subtests to administer.

The expert group will also need to consider the mode of testing. The literacy, numeracy and problem solving tests, and self-report surveys discussed here have used interview, paper and pencil methods. The foreign language test would likely need to be adapted to suit this method.

However, it may be feasible and cost effective if some portions were computer-based: this option is recommended. In addition, some alternatives to the straight interview method are being tested in some national projects, and these might be further investigated. For example, web-based test of literacy and numeracy (using IALS items) are under development by Educational Testing Service and Statistics Canada.

Instrument design process

As mentioned, the instrument design process will require some development of national assessments in preparation for usage in an international assessment. Even those instruments considered suitable for inclusion in the assessment without amendment (e.g., literacy, numeracy, problems solving items from ALLS) will need translation into all national EU languages.

Items from the UK Skills Survey will need further development, estimated to take up to 12 months. There are two issues to explore. First, language and translation issues, which can be rather straightforwardly handled with existing techniques. Second, there is a question as to whether the norms of assessment and reporting in countries vary. “Importance” scales may be inappropriate for some international comparison, while “frequency” scales are not suitable for all skill domains. Both would need evaluation and adaptation through the pilot study.

All self report survey items and test items need to be translated in ways that do not impair the validity of the cross-national comparisons. Translation from a reference language can generate differences in meaning, and double translation--where two translators generate independent versions in the target language and a third translator reconciles any differences—is recommended over back translation.

Care will also need to be taken to adapt items carefully in order to minimize the potential for cultural bias in the item pool and to ensure that item difficulty is not compromised.

Experience suggests that there will be a demand for national differentiation in the design of the assessment. One way to accommodate national interests is to allow some choice in the inclusion of specific blocks of questions, e.g., countries might decide to include or exclude the foreign language element. However, it is our strong recommendation that differentiation in implementation should not be permitted.

### **3. THE POTENTIAL AND LIMITS OF A EUROPEAN INITIATIVE**

This section discusses in more detail the overall potential of a European adult skill assessment initiative, as well as some limits and risks that a strategy for such an initiative must address.

There are both promises and potential drawbacks from developing and implementing a European skills assessment initiative. Our rationale for raising a number of critical concerns is not intended to discourage a European adult skills assessment initiative. On the contrary, there are many good arguments to justify such an effort. However, a European adult skills assessment initiative should be well founded, well focussed, and self-consciously aware of the risks and possible unintended effects that could result from it.

#### **3.1. *The Need for a European Initiative for Assessing Adult Skills***

There are a number of arguments that support the need for adult skills assessments on a European level, some of which have been touched upon already:

- 1) It is commonly agreed that skills, competencies and knowledge are becoming ever more important in modern societies, both in relation to international competitiveness and in relation to social inclusion and the prevention of social marginalisation of specific groups.
- 2) Life-long learning is important in skill development; it is not enough to improve the formal education systems that serve children and young persons.
- 3) Presently little is known about which segments of the populations are in need of continuing education and training or what that education and training should cover.
- 4) Systematic information on the outcomes of the formal education systems in equipping citizens with necessary skills, given different levels of investment, is incomplete.
- 5) Little is known about the extent to which existing competencies are effectively utilised in different occupational sectors and contexts, and how and why rates of utilisation differ.
- 6) International comparisons of adult skills have earlier proved both useful and efficient for providing some of this knowledge and information.

#### **Skills, the Knowledge Society and the Need for Lifelong Learning**

The common assumption that skills, competencies and knowledge are becoming ever more important in modern societies is partly based on observable changes in the economy. The view that we are leaving an “industrial economy” and entering a “learning economy” is illustrated in table 1 below.

Global competition in a networked society, with highly flexible and adjustable forms of production and rapidly changing technology, poses specific challenges to the workforce. Traditional, fixed professional identities are being broken down, and new professional categories and job profiles are continuously emerging and being reshaped. The share of knowledge intensive jobs is growing and new types of skills requirements are emerging. These changes lead to a need to constantly update the qualifications of the labour force (Lundvall 2000; Bainbridge et al. 2004).

However, competitiveness is not to be achieved at the cost of ever growing social and economic inequality. Employability of the workforce should be employability on good conditions, with high wages reflecting high productivity. In the European context, this is the basic political premise upon which the learning economy is to develop. To recall the words of the Lisbon Strategy, the objective is not only more jobs but “better jobs”.

**Table 1: From the Industrial Economy to the Learning Economy**

ISSUE	The Industrial Economy	The Learning Economy
Markets	Stable	Dynamic
Scope of Competition	National	Global
Organisational Form	Hierarchical	Networked
Organisation of production	Mass production	Flexible production- embedded services
Key drivers of growth	Capital/labour	Access to knowledge/innovation systems (triple helix model)
Key technology Driver	Mechanisation	Digitalisation, miniaturisation
Source of competitive advantage	Economies of scale	Time to market, innovation
Relations with other firms	Single mover	Alliances and collaboration, "co-petition"
<b>The workforce:</b>		
Policy Goal	Full employment	Employability
Occupational profile	Fixed professional identity defined in the national/regional trade context	Converging and continuously emerging and being reshaped tied to globalising contexts and technological advance
Skills	Job-specific	Multidimensional (deep and broad foundation skills)
Requisite Education	A skill – A degree	Lifelong Learning

*Adapted from Atkinson (1998).*

### Limited Information about Lifelong Learning Needs and Achievements

As skills are developed over a lifetime, it is important for policy makers to understand the needs for lifelong learning and the outcomes of lifelong learning activities. Available information on lifelong learning in the EU is presently focused on activities rather than on learning outcomes. The 2003 progress reports on lifelong learning (Commission 2003b, 2003c), for example, focus on various forms of activities in relation to lifelong learning, reforms of educational systems, participation levels in different forms of learning, and so on. Information about learning outcomes and more generally about skills levels and skills shortages in various segments of the countries’ populations was not gathered in the questionnaires completed by participating countries.

Although the promotion of lifelong learning is one of the key elements in the Lisbon Strategy, the structural indicators which are intended to assess the progress towards achieving comprehensive lifelong learning are inadequate. Again, they emphasise learning activities rather than outcomes. Although preparatory work to develop quality indicators of lifelong learning has been carried out (Commission 2002), the only indicator which specifically addresses the question of lifelong learning concerns the share of the adult population aged 25 to 64 who state that they received education or training in the four weeks preceding a survey.<sup>11</sup> Other types of statistical information on lifelong learning in relation to the Lisbon strategy or more generally focus on educational attainment levels (e.g., Brainbridge, et al. 2004) or on public spending on training and education.

<sup>11</sup>Eurostat Structural Indicators website, Life-long learning activities:  
[http://europa.eu.int/comm/eurostat/newcronos/queen/display.do?screen=detail&language=en&product=STRIND&root=STRIND\\_copy\\_890046314483/strind\\_copy\\_817397594099/emploi\\_copy\\_934568948015/em051\\_copy\\_727901795419](http://europa.eu.int/comm/eurostat/newcronos/queen/display.do?screen=detail&language=en&product=STRIND&root=STRIND_copy_890046314483/strind_copy_817397594099/emploi_copy_934568948015/em051_copy_727901795419)

### **Limited Information on Returns on Investments**

Even though the Member States differ significantly in their level of public investment in education and training and in the scope of continuing education activities, little is known about whether these differences in spending and activities are related to differences in skills levels and characteristics. In 2000, public expenditure on education in Denmark, for example, amounted to almost 8.5 per cent of the GDP. In Germany and Ireland, the figures were 4.5 and 4.4 respectively.<sup>12</sup> Yet, there is very little information available to support a conclusion that the workforce in Denmark is much better equipped in terms of skills than the workforces in Germany and Ireland. Data on youth education attainment levels are not very informative in this respect (the level in Ireland is higher than in Denmark), and not much can be said about the strengths and weaknesses of formal education systems.

### **A European Initiative?: The Relevance of International Comparison**

Even if there is a good case for better understanding the learning outcomes and skills profiles of different national populations, the argument could be made that this information should be collected and dealt with at the level of the Member States. The Lisbon Strategy and the voluntary system of open coordination have set the stage for viewing training and education policies of the Member States as a key European concern.

At the Lisbon European Council and at later summits, the Member States have agreed to pull new domains of national public policy into a system of voluntary coordination with a goal toward creating a more dynamic and competitive knowledge-based economy for Europe. It is logical that feedback systems on progress towards stated objectives are also established on a European scale, as has been done in the comprehensive system of structural indicators. It also follows that any future skills assessment initiative must incorporate the concerns and priorities laid down in this Strategy.

Furthermore, other international skills assessment initiatives have proved important in influencing public debate and political agendas, and indeed in initiating policy reform. For example, the OECD's Programme for International Student Assessment (PISA) provided data on competencies of 15-year olds that policymakers have used as indicators of the success of their education systems. Scores for Denmark, for example, prompted a comprehensive overhaul in the primary school system. Similarly, the International Adult Literacy Survey (IALS) provided important information on skills in adult populations in participating countries and has had impact on policy in various ways (For a further discussion cf. Murray, 2003, and the examples provided in the IALS case study in Appendix A).

### **The Need for a European Skills Assessment Initiative**

In sum, several arguments seem to justify the need for a European skills assessment initiative. Skills assessments on a European scale are relevant in the context of the Lisbon Strategy and its focus on lifelong learning and increased investment in human resources. Lifelong learning and human resource investment are, in turn, essential if Europe is to become more competitive and create more high-skill, high-wage jobs. However, too little is systematically known about the current situation in all the Member States about learning *outcomes* and the results of human resource investments.

---

<sup>12</sup> Eurostat Structural Indicators website, public expenditure on education as percentage of GDP: [http://europa.eu.int/comm/eurostat/newcronos/queen/display.do?screen=detail&language=en&product=STRIND&root=STRIND\\_copy\\_890046314483/strind\\_copy\\_817397594099/inmore\\_copy\\_680414916384/ir010\\_copy\\_598439700219](http://europa.eu.int/comm/eurostat/newcronos/queen/display.do?screen=detail&language=en&product=STRIND&root=STRIND_copy_890046314483/strind_copy_817397594099/inmore_copy_680414916384/ir010_copy_598439700219)

International skills assessments promise to fill some gaps in the knowledge base and offer a potentially powerful tool to influence public agendas and inform policy.

### **3.2. *The Limits of Adult Skills Assessment***

Although there are a number of good arguments for supporting a European-wide assessment of adult skills, there are also risks and limitations to consider. One set of risks involves the practical problems associated in designing and carrying out such an assessment. These issues are explored more fully in Chapter 5. Another set of risks are perhaps more political in nature, but must be considered as well. Chapter 4 will discuss the political feasibility dimensions in more detail. This section briefly discusses some more fundamental challenges that have mainly to do with the definition and characteristics of “skill” as a concept.

#### **“Skills” as an Open and Contested Concept**

Even though the word “skill” is widely used, there is no consensus as to the precise meaning of the term. Moreover, skill as a label, which carries with it the prospect of labour market rewards, has historically been a contested concept among employers and employees, and sometimes between men and women, in the definition of jobs. “Skill” as a concept has also evolved in public discourse, starting from a fairly narrow definition used to refer to quite specific high-level educational qualifications and analytical capacities or to “hard” technical abilities or vocational competencies associated with particular occupations.

Over the past decade or more, the term has increasingly been used to refer to a broader range of qualities that individuals possess that are thought to be germane to workplace productivity. Skills may refer to technical, job specific competencies, but also to personal characteristics, attitudes, “generic” competencies (e.g., problem-solving, teamwork, communication) and even “aesthetic skills”(Stasz, 2001; Payne, 1999). This changing definition creates measurement problems (discussed further in the Section 5) that may be difficult to overcome.

#### **“Objectification”: Skills Assessments Threaten to Fix Attention on Measureable Skills**

The Lisbon Strategy and its system of quantitative targets and structural indicators calls for standardised data that support systematic comparisons between the Member States. This requirement can create pressures for Member States to live up to the set objectives (cf. Haahr 2004). It also creates a need for quantitative assessments that provide valid and reliable estimates of skill for a population. Skills that are not easily measured by quantitative methods may not be assessed at all.

While it is a methodologically sound approach to seek to measure only what is measurable, the approach is risky. Many relevant skills and competencies may be overlooked simply because they are not measurable. A European skills assessment initiative may, if successfully implemented, stir considerable interest. It may therefore also fix attention and the political agenda exclusively on the skills included in the initiative.

#### **Overlooking the Context Dependency of Skills**

Research indicates that many skills associated with work productivity are context dependent (cf. Bjørnåvold 2000; Stasz, 2001). Communication skills, for example, can differ according to

audience, purpose, style and mode. Some jobs involve communicating with external audiences or the public (e.g., flight attendants, sales persons), while others involve primarily internal communications with colleagues (Stasz, 1997). However, accounting for such context differences is difficult in standardised measurements. It is unlikely that the European adult skills assessment initiative can capture the context dependency of many skills. It will therefore to some extent inevitably paint a distorted picture of the character and distribution of skills.

### **Group-related rather than Individual Skills**

A fourth point is that many skills of interest to policy makers are tied to groups or organisations rather than to individuals. Problem-solving skills, for example, can apply to individuals and to teams or groups. The problem-solving capacity of the team rests not on the problem-solving capacities of each participating individual but on the capacities of the group or on the distribution of different types of skills within the group (Stasz, 1997). Research also indicates that a group works better if a number of different skills profiles are represented within it (Belbin 2003). Similarly, and by implication, the “problem-solving” capacity of organisations cannot meaningfully be said to reflect the sum of the problem-solving skills of its employees.

### **Ethical Concerns**

A final point concerns the rights and responsibilities of the individual or the family versus public authorities in developing skills. The proliferation of broader notions of skills and competencies, and the predominant belief that “soft” skills will be ever more important in the modern economy, raises new questions. The desire to measure and compare European populations’ “problem-solving competencies”, “critical thinking competencies”, “physical/health competencies”, “communication competencies”, “aesthetic competencies”<sup>13</sup> or other “soft” skills may add a new dimension to the way in which governments and international organisations (the EU and the European Commission) become involved in personal life.

It is not a question of governments or agencies directly interfering with specific individuals’ lives based on their performance on an assessment (such problems can be prevented through confidentiality procedures), but rather that of new skill domains being pulled into the sphere of public policy and political intervention via measurement and international comparison. In developing a strategy for European adult skills assessment, it is important to consider what governments should be permitted to measure. This is also a question of political feasibility since skills assessment on a European scale may be controversial.

---

<sup>13</sup> All examples of skills/competencies that are frequently mentioned as key competencies in the country contribution reports of the OECD’s DeSeCo-initiative for selecting and identifying key competencies for a “successful life and a well-functioning society”, cf. Salganik and Stephens (2003); Trier (2003).

## 4. THE FOCUS OF A EUROPEAN SKILLS ASSESSMENT INITIATIVE

Adult skills assessment at a European level is likely to be a both costly and risky venture. Consequently, any future initiative should be designed to be as useful as possible for policy development purposes.

This chapter first examines briefly the relevant policy statements in order to identify the strategic priorities that have already been set out by the European Council. A new initiative should arguably be aligned with these policy directions to the extent possible, recognising that methodological or other limits may prevent close alignment. Then, the chapter raises certain issues regarding political feasibility. Based on prior experiences in assessing skills at the national or international levels, a number of guidelines are formulated, which aim to enhance the political utility and usefulness of a European skills assessment initiative.

### 4.1. *Political Priorities: Which Skills Are in Focus?*

The two sides of the Lisbon strategy – economic competitiveness and social inclusion – have significant consequences for the way skills development is viewed in a European policy context. Skill development is important not only for the sake of creating a more productive labour force, but also to optimise chances for individuals in their lives as a whole. Developing the appropriate skills goes beyond the domain of education policy, as it is seen as a trajectory of lifelong learning, involving labour policy and social policy as well.

This broad policy context will be described below in terms of the official decisions made by the Council. What kinds of skills are defined in the objectives derived from these decisions? And consequently: what should be the overall focus of a future European skills assessment strategy? Which skills should be assessed against this policy landscape?

#### **Strategic Priorities set by the European Council**

The need for improved skills development among Europeans can be observed in a number of European Council decisions, especially in connection with the Lisbon Strategy. These decisions demonstrate a broadly shared awareness of the importance of skills and of the need to strengthen skills in Europe. They also demonstrate growing attention to the need for the assessment of skills.

Three summits are of particular relevance here: Lisbon, Stockholm, and Barcelona. The Presidency conclusions are general agreements and thus provide few specifications about what kinds of skills should be developed and assessed. However, the most frequently mentioned skills are literacy, numeracy and e-skills, with some attention drawn to entrepreneurial and social skills – skills seen as being developed through trajectories of lifelong learning.<sup>14</sup>

#### *Lisbon: Literacy and the Prevention of “E-Exclusion”*

Most central to the European policy context on skills development are the decisions made during the Lisbon Summit of March 2000, resulting in the Lisbon Strategy. It concluded the following:

---

<sup>14</sup> In the following, quotes are from the respective Presidency Conclusions, available at [http://europa.eu.int/european\\_council/conclusions/index\\_en.htm](http://europa.eu.int/european_council/conclusions/index_en.htm).

“Every citizen must be equipped with the skills needed to live and work in this new information society. Different means of access must prevent info-exclusion. The combat against illiteracy must be reinforced.” As discussed earlier, this emphasis on skills has two sides, social and economic: “The new knowledge-based society offers tremendous potential for reducing social exclusion, both by creating the economic conditions for greater prosperity through higher levels of growth and employment, and by opening up new ways of participating in society.” Still, skills development can serve both sides simultaneously as “the best safeguard against social exclusion is a job.”

#### *Stockholm: Emphasis on eSkills*

Growing attention for eSkills can be discerned as member states gathered at the Stockholm European Council of 23-24 March 2001. There it was stated that “With almost half the working population using computers at work, rising to three quarters for white collar workers, formal PC training either at home or in the workplace may be a powerful tool to boost digital skills. However, less than 22 per cent of the workforce has had any formal computer training. The number of people whose employers have paid for that training is even smaller.”

#### *Barcelona: Focus on Skills Recognition and Basic Skills*

At the Barcelona Summit (March 2002) it was decided that the European education and training systems should become a world reference by 2010. The importance of skills development is recognised again, while the issue of skills assessment gains more prominence: skills need to be assessed and certified in order to stimulate (job) mobility. In order to achieve this “the Council welcomes the content of the Action Plan on Mobility and Skills and has agreed on the importance of the three challenges of the Plan: (a) to improve job mobility; (b) to promote geographical mobility; and (c) to establish adequate channels of information on work and training opportunities in the EU. To this end, it stresses the need to develop and recognise qualifications and skills, including those acquired informally, to invest in human resources, to pursue efforts to ensure lifelong learning and to modernise Public Employment Services, particularly the EURES network.”

One of the calls for action in the decision also specifies the kinds of skills that are seen as important, although it is primarily aiming at formal education and not lifelong learning: “to improve the mastery of basic skills, in particular by teaching at least two foreign languages from a very early age; establishment of a linguistic competence indicator in 2003; development of digital literacy; generalisation of an Internet and computer user’s certificate for secondary school pupils.”

#### **Skills Focus in EU Education and Labour Market Policy**

The strategic priorities set out in these declarations illustrate the importance attached to skills development by European governments and suggest the general focus on certain skills. Some additional information on the perceived importance of various skills can be derived from various Commission communications and reports and concrete activities developed in response to the various priorities defined by the Presidency Conclusions. We highlight a few illustrative examples below.

#### *Personal Competencies and “Learning to Learn”*

During the Barcelona Summit, an existing Commission report (Commission 2001) was accepted which gives a more precise description of skills under scope, through a set of concrete future objectives for education systems.<sup>15</sup> This report sets out an overall approach to national education policies for the member states. Relevant objectives in the context of skills policy are, among others:

---

<sup>15</sup> <http://europa.eu.int/comm/education/doc/official/keydoc/com2001/com2001-059en.pdf>

- Improvement of foreign language teaching, seen as essential if Europe is to achieve its potential – whether this is its economic potential, its cultural or its social potential.
- Updating the definitions of basic skills and professional skills.

Basic skills are in this connection seen as skills that “cover the vocational or technical skills as well as those social or personal competencies which enable people to work together and to lead happy and fruitful lives”.

Accordingly, as the report argues, the range of relevant personal competencies is today very broad: “The range of skills used in the work place is constantly widening. Similarly, society as a whole is less uniform than in the past, so personal competencies (such as adaptability, tolerance of others and of authority, team work, problem-solving and risk taking, independence, etc.) are more widely required if people are to live together in tolerance and respect for each other”. Specific emphasis is placed on “learning to learn”-abilities: “The most important of these competencies is the ability to learn – maintaining the curiosity and the interest in new issues and skills – without which lifelong learning cannot exist.”

#### *Long-standing and Continuing Focus on Language skills*

A simple search on Cordis for “language skills” results in 8 programmes and 258 projects, ranging from promoting foreign language learning in general to basic language skills for specific vulnerable groups. More projects can be named, but it may suffice to state that in a multi-lingual community as Europe, possessing language skills is self-evidently a key skill, as is foreign language policy.

#### *The eSkills Forum*

The eSkills Forum provides a very clear example of a priority given to defining and assessing a specific skill. Following up on the eSkills Summit (Copenhagen, October 2002) and activities of the ICT Skills Monitor Group<sup>16</sup>, the forum was established in mid-2003. According to their terms of reference, this forum will “bring together all relevant stakeholders to catalyse discussions and actions to narrow the gap and to address mismatches of ICT and e-business skills.” A more specific attempt to assess e-skills on a European level is the European Computer Drivers Licence (discussed further in Chapter 5).

#### *The European CV: A Comprehensive List of Skills*

The European CV is an attempt by CEDEFOP (the European Centre for the Development of Vocational Training) to make the European labour market more transparent. The CV is not intended in any way to facilitate aggregate skills assessment, but it is illustrative of the kinds of skills that are considered relevant from a European labour market perspective. Skills that are described in the European CV include:

- Language skills (mother tongue and foreign languages): reading skills, writing skills and verbal skills.
- Social skills and competencies: “Living and working with other people, in multicultural environments, in positions where communication is important and situations where teamwork is essential (for example culture and sports), etc.”

---

<sup>16</sup> The ICT Skills Monitoring Group was established in September 2001 by the Enterprise Directorate-General with representatives of Member States and Norway in the scope of the eEurope GoDigital initiative.

- Organisational skills and competencies: “Coordination and administration of people, projects and budgets; at work, in voluntary work (for example culture and sports) and at home, etc.”
- Technical skills and competencies: “With computers, specific kinds of equipment, machinery, etc.”
- Artistic skills and competencies: “Music, writing, design, etc.”<sup>17</sup>

### **Scope for Defining Useful Skills Assessment Strategy**

The policy statements above and the highlighted examples of activities and communications illustrate that there is no firm European agenda for adult skills assessment. A range of skills is being discussed, with specific attention being paid to language skills, digital skills (e-skills/ ICT-skills), and literacy. Nevertheless, skills such as entrepreneurial skills and the broader “soft” skills such as social skills, learning-to-learn abilities and broad personal competencies are also seen as important.

In our view, the breadth of the political agenda on skills in Europe reflects several things. First, it reflects the dual economic and social ambition found in for instance the Lisbon Strategy. Skills that are seen as immediately or exclusively relevant for competitiveness and productivity are considered important, as well as skills that contribute to peoples’ ability to lead “fulfilled lives” and being an active part of society.

Second, the variety of different skills in different contexts seems to reflect an uncertainty about the kinds of skills that are required in modern society, seen from both an economic, labour market-oriented perspective and a social inclusion perspective.

The implication for the development of a strategy for a European skills assessment initiative is that there is considerable scope for identifying the kinds of skills that might be included in a European-level assessment.

## **4.2. Informing Policy Making: Lessons from Previous Efforts**

In addition to considering the interests of policy makers in defining the skills that might be assessed, it is useful to consider past experience. Prior efforts may help inform policy by suggesting possible advantages or problems. Here we highlight some lessons learned that were identified in our critical review of various initiatives.

### **Skills Must be Relevant for Specific Policies**

The so far somewhat frustrating experiences of the Danish National Competence Account are revealing in several respects (see Annex A for a case study). The Danish NCA was conceived within the framework of the OECD’s DeSeCo project as a first operational expression of DeSeCo’s goals. At the same time, it was also characterised by a certain euphoria surrounding the “knowledge society”/“lifelong learning” agenda around 1999-2000, in Denmark and elsewhere. Against this background, the original project was quite ambitious. The basic idea underlying the initiative was that a national competence account could be used as a tool to fine-tune training, education, HRD and other policies in order to enhance the position of the Danish economy and society, both in terms

---

<sup>17</sup> [http://www.cedefop.eu.int/download/transparency/cv\\_format\\_en.doc](http://www.cedefop.eu.int/download/transparency/cv_format_en.doc)

of international competitiveness and in the more general capacities of the population for leading a fulfilled and active life.

This idea, however, has turned out to be quite unrealistic, as the data does not lend itself to developing specific policy responses. At the time of our interviews, the project staff was struggling to identify some predominant themes in the data at all, which would allow them to draw some interesting and relevant conclusions. This problem may partly be resolved if the NCA survey was repeated at a later point, allowing comparisons over time and the identification of trends. However, it seems questionable whether specific policy objectives could in principle address some of the 10 key competencies covered by the initiative (e.g., “self-management competence”, “cultural competence” and “social competence”), within the formal education system or elsewhere.

### **Need for Evaluation Criteria**

The problems of the Danish NCA appear to stem not only from its focus on a number of skills for which public policy intervention seems questionable. Even if the key competencies are viewed as appropriate within the policy domain, they lack an external reference point or evaluation criteria which could lead to policy recommendations. What level of “democratic competencies” denotes a “good” level of competence? At what level is there a problem? When is the workforce’s level of learning competencies or social competencies problematic, when is it not? When does it call for a political response? Without clear answers to such questions, staff at the NCA are faced with the challenge of squeezing out some conclusions that are not entirely trivial, rather than drawing conclusions on the overall direction of educational policy or framing precise policy recommendations.

Even if the NCA design permitted longitudinal and cross-national comparisons, it would not solve the basic problem-- for most of the skills covered in the initiative there are no precise and generally accepted external yardsticks to measure which levels of skills are unacceptable, acceptable or desired. Notions of “the knowledge economy” and the DeSeCo project’s identification of 10 “key” competencies are too vague. There is no conclusive understanding of the precise mixture and level of skills that would ensure a competitive and employable workforce in a competitive economy, not to mention “a successful life and a well-functioning society” (DeSeCo) for citizens.

This type of problem can only be avoided if required levels of skill are theoretically based and subsequently verified empirically. This has been done, for example, with direct measures of literacy and numeracy in IALS, but was not achieved with the proposed “practical cognition” assessments in ALLS. Even when levels have been theoretically and empirically defined, the reporting of those levels can still generate controversy (See NAALS review, Annex A).

In the context of a future European skills assessment initiative the lesson to be drawn is that any strategy should carefully consider which skills to assess, not only with respect to methodological concerns, but also with respect to usefulness. What policy conclusions, for example, can be drawn from an observation that Greeks have more social skills than the Dutch, and that the Dutch have more communication competencies? Is that a problem for the Greeks, the Dutch or for both of them? If so, why? As Member States often choose to evaluate their own performance against others, this issue of the meaningfulness of the measure needs to be addressed head on.

### **Focus on Essential Skills**

The OECD experience suggests that skills assessment initiatives may have policy repercussions to the extent they highlight conditions that are seen as important problems or threats to either economic competitiveness or social inclusion.

If a European skills assessment initiative is to have an impact in policy terms, it should therefore focus as much as possible on those skills which are generally agreed as crucial for an individual's participation and performance in working or social life. Emphasis on skills in which there is a high degree of consensus has the further advantage that it will be easier to agree on external evaluation criteria or assessment levels (as has been demonstrated in the OECD's literacy studies). It is rather uncontroversial to view "illiteracy" or "poor reading skills" as a problem that must be addressed. While the DeSeCo project is an attempt to identify other "key" competencies, this effort can only be seen as a first tentative step in developing a consensus on the competencies valued in modern society.

### **Potential Controversy: Respecting the Limits of the Personal**

There may be considerable controversy in assessing certain "soft" skills and personal competencies. Emphasis on cognitive competencies, learning skills, or other more personal domains, for example, may be seen as intrusive.

Experiences from the Finnish "Assessing Learning-to-learn" initiative, for example, are illustrative (See Appendix A for further discussion). This initiative seeks to assess school childrens' "learning to learn" competencies, which are defined by a number of different cognitive, behavioural and motivational dimensions. The assessment has been controversial in some respects, however, as it can very easily be interpreted as a form of intelligence testing. For this reason, the initiative carefully avoids analyses that could be seen as examining the significance of learning-to-learn "intelligence" scores for success in later life.

Such controversies and concerns may multiply if assessments measure skills related to deeply personal traits, not only among adults and outside a classroom context, but on a European scale. Assessment of such skills would require a considerable pool of legitimacy and trust from citizens towards their governments and the EU.

### **4.3. Political Priorities and Possibilities**

This chapter has described the political desirability and feasibility of measuring certain types of skills. Will the measurement of skills and, for example the international comparison thereof, inform policymakers about the achievement of the specific goals that they have set? Can this information be useful for a wide variety of future policy challenges and objectives or are they restricted in their policy relevance?

Our review of policy and current activities indicates that European policy makers focus predominately on certain skills, namely on literacy, languages and digital skills, with several other "soft" skills also being highlighted.

Our review of lessons learned from previous skills assessment initiatives highlights the pitfalls and risks associated with assessments of a very broad range of skills. If adult skills assessment is to be

relevant for policy development purposes, it seems that the following should be taken into consideration:

- For the skills assessed, it should be both possible and realistic to address skill levels through policy intervention.
- It should also be possible to formulate rather clear evaluation criteria or “acceptance levels”, otherwise results are likely to be irrelevant.
- To maximise policy impact, skills assessment should focus on skills that are generally considered indispensable, from an economic or social perspective. Information on skills that are not considered essential, or are of questionable significance, is unlikely to significantly affect policy.
- Assessing skills that can be seen as deeply personal and sensitive may not be politically feasible.

Therefore, it appears that *language skills*, *literacy* and *e-skills* could comprise key elements in a European skills assessment initiative, at least from the perspectives of policy goals and political feasibility. These skills domains are prioritised in European Council conclusions, and appear to live up to the feasibility criteria. Other skills that are currently in focus, for instance *entrepreneurial skills* and certain *social skills*, could also potentially be elements in an initiative provided that adequate answers to crucial questions can be found (e.g., can acceptance levels be defined? Can general agreement on significance and definitions be established?).

Whether it is methodologically feasible to develop a European skills assessment initiative that generates valid and reliable information on these and other skills domains is another question. We turn to this question in the last chapter.

## 5. METHODOLOGICAL ISSUES

The overall issue to be addressed in this chapter concerns the question: how can a European direct assessment initiative best be designed so as to produce valid and reliable results?

First, a description and evaluation is provided of the various potential methods of assessment. In the next section, we consider several important aspects of the data collection process. In light of these considerations, the options available and in particular the advantages offered by a one-visit, integrated, data collection process, several methodological issues arise in respect of which skills domains might be assessed, and how one might carry out the assessment. Finally, we discuss some methodological considerations relating to the design of the instruments to be used for the assessment.

### 5.1. Ways of Measuring Skills in an Adult Population

Table 2 summarises three potential ways in which the skills of the adult population might be directly assessed. Before discussing these methods, however, it is useful to recall the main indirect method of assessing skills that has traditionally been used, namely through measures of attainment in national education systems.

**Table 2: Ways of Direct Measurement of Skills in the Whole Adult Population**

Approach	Example(s)	Advantages	Disadvantages
<b>1. Occupation</b> The proportions in “higher-skilled” occupations	TOWES Machin and Van Reenen (1998); Howell and Wolff (1991).	Occupational titles easily available from labour force surveys or censuses; sometimes internationally comparable	No European occupational classification system; skills change within occupations; the level of skill varies within occupations; “expert” assessments are imperfectly reliable
<b>2. Tests</b> Scores from literacy numeracy, language or ICT tests	IALS, ALLS, DIALANG, IVQ. OECD et al. (1997); Freeman and Schettkatt (2001)	Objective; international comparisons possible	Narrow range of skills; expensive to administer; no successfully tested and reliable ICT tests.
<b>3. Survey-Based</b> <b>a) Self-Assessment of Competencies</b> Individuals’ own reports about themselves	Elements of SKILLS FOR LIFE SURVEY Bynner (1994)	Wide range of skills	Subjective, and skill assessment associated with self-esteem; literacy/numeracy self-assessment proven unreliable
<b>b) Self-report of Job requirements</b> Individuals’ reports of generic activities	1997 & 2001 SKILLS SURVEYS Ashton et al. (1999); Felstead et al. (2002)	Wide range of skills; intimately connected to jobs	Job skill requirement could differ from person skill; does not measure skills of non-employed people.

This proposal begins from the proposition that measures of educational attainment – whether qualifications-based or related to years in school and college – are too loose as measures of the skills of adult populations and as indicators of the skills being deployed in workplaces. They provide for no assessment of the capacity of educational systems to generate the skills needed for work and for life in general.

Nevertheless, educational attainment indicators (such as ISCED levels) are potentially informative. If qualifications could be made internationally more comparable in a reliable and accepted manner, particularly in certain occupational fields, measures of the proportions of qualified adults would be a very useful addition to the policy-makers knowledge. Nothing in this proposal, therefore, should detract from efforts to develop comparability of occupational and academic qualifications across Europe.

### **The Occupational Titles Approach**

One direct method of assessing skills, shown in the first row of the table, is via occupational titles. The idea is that if the skills required to be used in any “job” defined by its occupational title could be measured reliably and precisely, and if such measures were available for all occupations, then the skills of the employed part of the population could be assessed through a simple count of occupational employment, as obtained through censuses or labour force surveys.

Unfortunately, neither of the two conditions in the previous statement prove to be satisfied. The skills needed for particular occupations have been assessed directly through interviews with job incumbents, for example in the TOWES project in Canada in a number of occupations. The purpose was not to gain a national skills assessment, but to assist in recruitment for employers and in career planning for individuals. It is doubtful whether the experts’ assessments are reliable enough to command acceptance as a reliable method in a national assessment. Often, experts’ judgements of qualification requirements are simply done from the achievements of incumbents, which is not necessarily a good guide to the present and future requirements for new recruits to the occupation. Moreover, only a proportion of occupations have been assessed.

There is also considerable variation in the skills requirements even within fairly narrowly defined occupations: the variation derives from how individual jobs are configured. Sometimes it is very difficult even to assign individuals to occupations. The skills requirements are also changing at a pace likely to be faster than any regular re-assessments by the experts. Even if all these issues did not arise, a European assessment using this approach would require availability of a detailed breakdown of the occupational composition of employment in each member country using a common coding system – something which does not currently exist. For these reasons, the occupational titles approach is not part of our recommended strategy.

### **Testing: Advantages and Disadvantages**

The second row of Table 2 concerns tests. There is a general presumption in favour of testing as the best method of assessing individuals’ skills, because tests are “objective”. They do not suffer from the potential biases that can come from dependent parties, especially the individual being assessed. Tests therefore start with a presumption of superior reliability over other methods of assessment.

This claim to the high ground of objectivity underpins the history of testing in the field of international skills assessment from IALS onwards, and accounts for the considerable impact of the limited number of such assessments that have hitherto taken place. There remain, however, several issues surrounding the use of tests:

- The validity of the tests needs to be considered: what exactly are they testing? For example, the “quantitative literacy” tests in IALS were criticised as being too far removed from an

adequate test of numeracy, because they required reading skills to be able to do the tests. The literacy skills tests in the Skills for Life Survey have been criticised as testing an indeterminate range of skills, including others besides basic literacy, so that it was unclear whether poor performers lacked the literacy skills or other types of skill.

- A second issue concerns the transferability of the skills assessed in the rarefied context of a test. Do skills tests carried out in the home predict performance in the workplace? To reduce the context dependency of test scores, tests so to speak create a standardised context. The question which cannot always be answered is whether this context is relevant outside the test (cf. IBF 1998; TSER 1999)?
- A third issue is that the reliability of the test procedure for gaining a national-level skills assessment is limited by low and differential response rates, and by respondents becoming bored or tired during long tests.
- A fourth issue is the precision of the tests and their ability to discriminate between groups. The precision can be made to differ according to the general skill level of the individual to meet particular policy needs. For example, accurate measurement of basic skills is likely to be most important for the development of policies to enhance individuals' employability.

If the degree of reliability and the validity of the tests are thought to vary significantly across countries (for example, through differential data collection mechanisms leading to greatly-varying response rates), the international comparability of the skills measures, and hence the overall assessment strategy, is compromised.

#### *Positive Lessons*

The history of direct assessment testing in national and international contexts in the recent period has taught us a lot about the consequences of these issues. These lessons need to be assimilated as part of any European-level strategy. On the positive side, the series of tests carried out under the auspices of the OECD and some national governments – we refer to the NAAL, IALS, ALLS, IVQ and Skills For Life Survey (see Annex A for case studies) – have shown that certain skills are testable, and that they do measure up to reasonably acceptable standards of international comparability and have therefore achieved sufficient reputation and respectability.

That the measures of skills have been found to have real effects on employability and other indicators such as pay and social inclusion suggests that the skills tested do transfer at least to some extent to work contexts. The findings have been considered sufficiently robust in most countries to influence educational policy, particularly in respect of the resources devoted to teaching basic literacy.

#### *Negative Experiences*

On the negative side, however, the experience of these studies suggests that testing is limited to a relatively narrow range of skills. For certain other skills, for example team-working or communication skills, it has not (till date) proved possible to devise tests that are sufficiently valid, reliable and transferable to the workplace, and equally so in many countries.

A second negative aspect of these existing tests is that the development and execution of skills tests has been expensive. The design of tests occupies quite a large part of the central cost. For a European strategy this could be minimised by drawing on existing survey designs, but some development work will be necessary if only to translate items into all the European languages.

### *The Methodological Limits and Possibilities of Testing*

From our evaluation of all the existing tests and our reading of the efforts towards test development on a wide range of skills, we draw the following general methodological conclusions:

- Testing must unfortunately be restricted to a narrow range of skills. These certainly include literacy and numeracy. They might also include foreign language skills and ICT (see more below). The development of tests for other skills, such as communication skills, would be costly and likely not successfully pass the criteria of validity, reliability and transferability.
- Existing tests of literacy and numeracy have been refined to a reasonable extent. We believe that there is little point in developing new instruments for a European assessment. The European strategy should consider using and translating instruments from either the recent ALLS or the French IVQ. The Skills for Life Survey also presents a possible array of instruments for literacy and numeracy measurement, but these are based on multiple-choice methods which are normally considered inferior, especially at the lower end of the skills spectrum, and their validity has not been established.
- The choice between the sets of items used in ALLS and IVQ does not seem to us to be decisive at this stage from the point of view of methodology. The advantage of IVQ lies in its design, which permits more differentiation at the lower end of the scale; the possible advantage of ALLS lies in the overlap with the OECD's approach to skills assessment, which will almost certainly involve a considerable element of continuity with previous studies.
- As more findings are published, it will be easier to evaluate the content validity of IVQ and ALLS, and this may assist in the choice of a testing methodology as part of a European strategy. An additional consideration will be the length of time taken for the assessment. As part of the design of a European questionnaire, an objective should be to choose items with the greatest validity to shorten testing time on literacy and numeracy, thereby freeing up time for assessing other skills.
- Efficient use of interview time, as well as ensuring good response rates, means using some form of early sorting of participants to direct them to the appropriate level of skill testing.
- A most important lesson that has been learned from the existing test-based studies concerns data collection processes (cf. below). The consistency of these processes must not be compromised, otherwise the findings will not carry respect and all other efforts towards international comparability with regard to survey and test design will be futile.

### **Survey-Based Self Assessments**

The alternative and potential supplement to testing is to survey nationally representative samples of individuals. One way that has occasionally been used to assess skills through surveys is through self-assessment, as listed in row 3a of Table 2. The approach has been to ask individuals to say "how good" or "how effective" they are at certain activities.

The advantage of this indirect method is that one can ask about a broad range of skills. The disadvantages, however, are that the assessment is severely subject to social-esteem bias and to a lesser extent that individuals may have unreliable knowledge of their own skills. Though in one survey (the Skills for Life Survey) individuals were found to report their ICT skills reasonably well, individuals' judgements about their literacy and numeracy skills have been found to considerably exceed the results of tests. Self-assessment would not command acceptance in any international assessment, and this method, though perhaps useful in special contexts, is not recommended for further consideration as part of this proposal.

### **Survey-Based Self Reports**

A more satisfactory survey-based method of skills assessment is through self-report of individuals' jobs, as listed in row 3b. The idea is to question respondents directly about their work activities and the requirements of their jobs. Because individuals are directly reporting what they do, rather than how good they are or what skills they have, they are more likely to reply without social esteem bias.

This difference is, of course, a matter of degree. Some respondents may still "talk up" their jobs to make themselves sound more important. Nevertheless, tests of inter-rater consistency are encouraging, and suggest a reasonable degree of reliability. Moreover, arguably the individual jobholder knows most about his/her own job. Thus, self-reports suffer from less measurement error than other methods such as experts' judgements.

Survey-based job analyses are not, of course, the only way in which work skills can be assessed directly. Work consultants and academic ethnographers will typically use a range of methods, including employee surveys, participation, observation, focus groups and so on, to assess from many angles what each job involves; the object of investigation may extend beyond skills to behavioural traits and employee relations.

These other methods, however, are typically tailored to the particular needs of the occupations prevalent in a small number of establishments, and could not be generalised to populations. The skills measures would often not be truly generic (a "generic" skill could be utilised in a wide variety of contexts). Moreover, the costs of extending such an intensive investigation of skills in particular establishments to some representative sample of the national population would be prohibitive. Thus, job analysis-based measures of skills are viable as part of a national (and European) strategy only by adaptation to the context of the social survey.

#### *The British Skills Surveys*

Such an adaptation is exemplified in the 1997 and 2001 Skills Surveys in Britain. The approach to measuring skills can be shown by the following example of a question. After some pre-ambles, respondents are asked: "In your job, how important is making speeches or presentations?" They are given a 5-point scale to respond on. There are 37 such questions as well as others designed to look in more detail at the usage of computers.

Statistical methods are used to reduce the large amount of information about people's jobs contained in the responses to a manageable and theoretically coherent set of skill scores. In practice, eleven detailed generic skills are identified and measured in this way, including computing skills, literacy, numeracy, technical know-how, high-level communication skills, planning skills, client communication skills, horizontal communication skills (such as working within a team), problem-solving and checking skills. In addition, broad measures of the skill level required in jobs can be obtained through measures of the education requirements of new recruits and of the cumulative training and learning time of jobholders. Repeated cross-sectional surveys of representative populations generate a description of change in the utilisation of skills in the workplace.

This method of skills assessment is able to generate reliable measures of the skills utilised at work. Compared to testing methodologies, it has two advantages that are relevant to the European strategy. First, it can be used to create a much broader range of skills estimates, at the comparatively low extra cost of time during a social survey. Second, the job analysis method measures skills, by definition, in the context in which they are being used. Therefore, there is much less concern about

the transferability of the skills. With tests performed in the home, there is a worry that these may not reflect people's competencies in real world situations. With job analysis based measures, however, there remains only the lesser concern that the skills used in one work context might not transfer fully to those in another organisation.

#### *Disadvantages of the Job Analysis Method*

There are two important disadvantages of the job analysis method. First, as an indication of the skills of individuals, this method gives us only approximate measures. Jobholders may have too many or too few skills for the job. If too many - and they have relevant under-utilised skills - the jobholders become dissatisfied, and the statistics would underestimate the stock of skills in the working population. If too few, the employer would become dissatisfied with the employee's performance. It is a plausible presumption that in the long term individuals become matched to their jobs. Those with insufficient skills lose their jobs or are not hired in the first place; those with under-utilised skills move on out of their own choice, if they can. This presumption may be more valid with respect to some skill domains than others. For example, in respect of computer use it is expected that the match between usage and personal human capital is quite close. There will be few persons skilled in programming, for example, who do not use these skills in their work. In the case of other skills such as writing, there may be a significant number of individuals who would be very capable of writing long reports, but whose work does not require it. The match might also vary across occupations. In particular, there is a reason to expect a better match for higher-level jobs: for managerial and professional jobs, where the jobholder often has the opportunity to mould the job to his/her own taste, the activities of the job will become aligned even more to the jobholders skills.

Nevertheless, it should be recognised that the national-level measures of skills obtained in this way would be strictly measures of the skills utilised in workplaces of each nation. The skills used are the outcome of complex forces operating on both the demand and supply side of the labour market. On the demand side, the utilisation of skills is affected by the cultural approach and strategic objectives of employers in the industries they operate in, which affect the way that human capital is deployed and work is organised. On the supply side, the skills utilised are those which job-holders bring to the workplace, supplemented by the skills they develop through doing the job and through additional learning in the workplace.

To make inferences from the job-requirements data about the skills of the job-holders, it would be necessary to make explicit assumptions about the match between the job and the job-holder. As just argued, these assumptions may be more valid in some occupations, and with some skills, than others. Additional items asking respondents about skills mismatches between the individual and the jobholder can add to our knowledge of over- and under-utilisation of skills. These items may improve confidence in whatever assumptions are made. Nevertheless, the reliability of such items in an international context would need to be evaluated.

The second disadvantage is that the job analysis method cannot be applied to the economically inactive population (who - by definition - do not have jobs). There may be difficulties applying it even to those who are unemployed, though it is possible that an adapted method could be applied to those recently unemployed (This would require a pilot study, as there is currently no precedent). Thus, no direct inferences could be made about the skills of the whole adult population from the job analysis approach to measurement. Nevertheless, the use of this method for the employed part of the population, and perhaps also for those unemployed people who are recently out of work, is a viable means of measuring the skills of the majority of the population who are in employment.

## **5.2. The Data-Collection Process**

An assessment of the skills of an adult population involves selecting a representative sample of individuals, and for each respondent carrying out an interview involving either a test or a survey or some combination of the two. The interview is likely to take place in the individual's home. Though the workplace is an alternative in principle, there can be problems of securing a random draw of individuals in workplaces, and there is a risk of gleaning distorted replies from respondents feeling the influence of supervisors or other colleagues.

### **An Integrated Process?**

An important issue for early resolution in developing a European strategy for skills measurement concerns whether the different skills domains are to be measured through an integrated process, or whether two or more separate data collection exercises are to take place.

There is a strong argument in favour of using an integrated process, namely that there are expected to be robust associations between skills of different types. These associations can only be fully investigated if there are skills measures in multiple domains for each individual. The links between skill types can affect the evaluations of education and training systems. In addition, a significant part of the analyses to be carried out will seek to investigate the associations of skills with labour market and other outcomes. These associations should be looked at using multivariate techniques, rather than just examining simple correlation between individuals' skills and outcomes.

A second advantage of an integrated process is that any further data collection exercises would add significantly to cost. One could expect the costs of two separate data collections to be not far short of twice the costs of one.

Against this, however, must be noted the limitations of interview time that are inherent in any assessment of an adult population. Adults are going to be voluntary participants in this exercise, and they have constraints on their time and willingness to participate. The situation is quite unlike a national assessment of school children's skills (as, for example, in the PISA project), where the assessment can take place in school time under the supervision of teachers. Between one and two hours is the maximum time that one should conceive of taking for the assessment process in the home. There are, therefore, limits to the number of tests and survey questioning that could be carried out in this time.

One alternative, while keeping within the aim of an integrated data collection process, might be to consider second visits to the same individual, with the aim of carrying out a further batch of tests and questions, thus splitting each visit into manageable lengths. The experience of the Skills for Life Survey in Britain suggests, however, that this is a risky strategy: the response rate for the second phase of the assessment was down to unusually low levels, and for understandable reasons. Deploying a second visit also means incurring the extra cost. We think that integrated data collection should take place in one visit, in which effort and resources are devoted to ensuring maximum response rates.

A more viable option, while keeping to one visit, is to cover a variable range of skill domains within each interview. The aim is to make the single-visit integrated data collection process go further.

One can target a large sample, then randomly split the sample into two or more groups, and administer interviews assessing different combinations of the skill domains to each group. The interactions between skill domains could then be assessed through analyses within those groups. To obtain adequate aggregate descriptions of each skill domain, each group needs to be large enough to sustain reasonably precise results for the whole population. Within this strategy, one can also choose to administer a core domain (e.g. literacy) to everyone, in order to obtain a fuller set of disaggregated analyses of this domain.

Despite these considerations, it remains conceivable that a skill is of great policy interest, not very highly correlated with other skills, and not easily squeezed into the available interview time. In that case, the European strategy could contemplate a separate data collection process for this particular skill. Below, we consider whether language skills might be a suitable candidate for such separate treatment.

### **Quality and Management Considerations**

There is a serious risk that reliability in an international context is easily compromised by inconsistent data collection strategies. Not only might the sampling methodologies differ across countries, the experience of the IVQ in France implies that the sequence and manner in which questions are presented to respondents significantly influences the results. It is suggested that a one-by-one delivery of questions (as opposed to presenting a whole batch of questions in a booklet) elicits “better” responses: with a more congenial relationship between interviewer and respondent, probably more correct answers, fewer non-responses on individual items. Overall, the IVQ also claims a higher response rate, though this could also be due to the interview time being shorter overall. Whatever data collection strategy is adopted, several commentators have urged, and we agree, that the uniformity of the data collection process will be absolutely vital to the success of the European strategy.

It follows that the management of the data collection needs to be closely coordinated and to the extent possible centrally managed, so that oversight over methods can be assured. Though surveys will be nationally managed, this must be done to precise specifications of the body responsible for managing the overall data collection process. This injunction might seem obvious, and no doubt applies to all cross-national data collection exercises. However, its importance cannot be over-emphasised in this context, since seemingly minor variations in methods could have major consequences, and the results will be used to make significant statements about the skill levels of national populations.

In addition to this general requirement for international consistency, it goes almost without saying that sample selection, testing and interviewing must be of the highest quality in an exercise of this sort. This means that an adequate form of probability sampling must be deployed in every country, and that there would need to be considerable emphasis on the training of interviewers to carry out the quite complex interviews. Survey companies would also have to demonstrate that they have proper quality assurance mechanisms in place. We think that any high-quality survey company should be using computer-assisted-based interview techniques, and that these should be adopted uniformly through the assessment.

To support higher response rates, survey companies would need to be required to schedule several repeat visits to households to try to secure interviews with selected respondents. There also needs to

be a uniform policy concerning the payment of respondents. There is some evidence of improvement in response rates as a result of offering a small payment (perhaps 15 Euro).<sup>18</sup>

While survey data can be gathered in a standard way through face-to-face interviews, there is an issue to be decided about the administering of the tests. Should they be administered orally, i.e. with the interviewer asking the questions and recording answers, or should respondents be asked to write their responses to written test materials? The choice of oral or written methods will depend on the chosen instruments.

### **5.3. Methodological Issues Concerning the Skills Domains**

#### **Literacy and Numeracy Skills**

From the above we have concluded that literacy and numeracy skills can and should be assessed by means of a test carried out as parts of an interview. There remain two related issues to be discussed. One concerns whether to adopt the ALLS approach or the IVQ approach to literacy and numeracy testing. The other issue is how to avoid duplication of effort with others being urged by the OECD on some or all EU member countries. We think that the European Commission should consider taking a lead on ensuring cooperation.

#### **ICT Skills**

The importance of information and communication technologies throughout the “knowledge economy”, and the diffusion of these technologies into the large majority of occupations, means that ICT skills should be included as part of a European assessment. There is evidence that these skills increase the chances of gaining employment, and earn a premium for those employed in modern labour markets. Governments need to support their policies for enabling adequate development of computing skills with adequate knowledge of where the distribution and growth of such skills in the population, and of the use of those skills in workplaces.

#### *No Solid Tests Identified*

Whether ICT skills at the national level can be assessed through the preferred methodology of testing is, however, questionable. We have not so far uncovered an existing batch of ICT tests that could simply be applied to samples of adults in member countries. The UK Skills for Life Survey is an example of a test, but the validity of the questions does not seem to have been evaluated, and it is limited to a Windows environment.

Existing certification tests, such as for instance the European Computer Drivers Licence (ECDL) are also problematic. For one thing, most certification assessments are mostly summative, measuring whether a set of predetermined knowledge elements or skills was mastered. Summative tests are generally poor predictors of performance outside the specific knowledge area being measured. They may verify an individual’s internalisation of certain rules and procedures, and they may measure, to some extent, an individual’s ability to do something well by showing that the individual has mastered the rules and steps required to do something well. However, they do not provide meaningful data about whether an applicant “knows why”- i.e., application of knowledge is in context.

---

<sup>18</sup> In the U.S. NALS, a \$20 incentive was found to increase the response rate and decrease overall survey costs. The incentive was particularly effective for surveying harder to reach populations.

Therefore, any strategy for testing ICT skills would certainly require a programme of development. In favour of a strategy of developing a test of ICT skills for the adult populations is the standard argument that tests can provide reliable evidence.

### *Formidable Methodological Difficulties*

There are, however, some formidable methodological difficulties. Computing skills need to be tested in the context of varied software environments that may not easily be compared. Thus, respondents used to an Apple-Mac or Linux environment might struggle when assigned tasks to be carried out in a Windows environment.

Moreover, all these environments are changing quite rapidly. It is unlikely that the same test could be applied in successive assessments separated by, say, five years; it would therefore make it difficult to measure progress. A further consideration is that any test of ICT skills would need to take up a substantive amount of time in the one-visit integrated collection of data advocated above.

### *The Alternative: Assessing ICT Skills Through Usage*

In view of these difficulties, an alternative should be considered. ICT skills could also be evaluated through their usage. Self-reported usage in jobs, along the lines advocated above, would generate indicators of the level of computer skills. The time taken to go through the necessary items would be very much shorter than a test, and would elicit adequate responses even from those with limited ICT skills. Indicators of usage at work can be supplemented by indicators of home usage of computers.

In addition to the fact that this approach would be much more economical on interview time, there are two aspects of existing studies which speak in favour of using this approach. First, the computer usage measures in the Skills for Life Survey correlated well with the Windows-based tests of computing skills carried out in that survey. Second, the range of questions asked in the 1997 and 2001 Skills Surveys provided some valuable indicators of computer skills, which correlated as expected with labour market indicators. For example, both the importance of computer usage at work and the degree of sophistication with which they are used have been shown to be associated with higher wages. There is some evidence and good reason to believe that this association occurs because the importance and sophistication of computer usage proxies the extent and depth of computing skills which the jobholder possesses.

### **Learning-to-learn skills / learning skills**

“Learning skills” or “learning-to-learn”-skills have attracted considerable attention during the past years. Most likely this is so since such skills are seen to be of potentially far-reaching importance in an increasingly knowledge-intensive economy which requires a continuing upgrade of the skills of the workforce in lifelong learning processes to retain its employability.

The Finnish “Assessing learning-to-learn” initiative is, according to our knowledge, presently the most advanced attempt at devising a way to assess learning-to-learn skills. However, the Learning-to-Learn system only partly consists of tests, other important parts being self reports of beliefs and attitudes, with the methodological challenges this poses. Test modules concern verbal-argumentational comprehension, quantitative-relational comprehension, logical reasoning and reflective abstraction (cf. Annex A for a fuller description of the initiative).

In our assessment, the validity of the test modules can be questioned. Validity testing has only taken place as internal, structural consistency testing. Predictive validity has not been tested. Testing takes place via the multiple-choice method, which is methodologically problematic (random answers will generate a minimum share of correct answers but there is no effective way of establishing whether answers are made at random or not). Moreover, the domains that are being tested can very well be seen to refer to much broader competencies than just those associated with learning-to-learn or learning capacities. The logical reasoning and reflective abstraction domains would seem to border on problem solving skills. Verbal-argumentational comprehension and quantitative-relational comprehension seem to border on literacy and numeracy respectively.

In addition to these methodological questions, issues of feasibility must also be mentioned. The Learning-to-Learn test modules in many ways resemble traditional intelligence tests, and it would probably be highly controversial if such tests were to be carried out among adults on a European scale.

### **Other Generic Skills**

There has been little success in developing valid tests for a wider range of generic skills. There is also nothing to suggest that if the EU were to attempt further development it would be successful at an early date, or even after a long exploratory programme of development.

#### *Problem-Solving: Elements of Test May be Available*

One exception is the test of analytical problem-solving skills developed for ALLS. This test passed through quite stringent piloting, and the results from three countries suggest that different levels of problem solving can be identified and are distinct from the other scales. However, the measurement may not be interpretable at the lower levels as the measures are reading dependent. The assessment involves the ability to apply cognitive processes toward determining a solution when that solution is not immediately obvious to the problem solver. Depending on time constraints, this domain may be suitable for direct testing also in an EU assessment (However, cf. IBF 1998 and TSER 1999 on the problems of developing valid tests in the domain of problem-solving).

#### *Self-Reports for Other Generic Skills*

Other generic skills can be assessed through the method of self-reported job requirements discussed above. In this way, it would be possible to derive, for those in work or only recently unemployed, assessments of their literacy, numeracy, technical know-how, high-level communication skills, planning skills, client communication skills, horizontal communication skills, problem-solving and checking skills.

In the case of the first two domains, literacy and numeracy, the assessments would be supplementary to the test measures. Though the levels or units of assessment between the two approaches (test or usage) would not be commensurate, individuals could be ranked according to each scale, and the comparison between the two would provide an indication of how well skill possession and skill usage are matched in respect of these two domains.

An alternative would be to omit the literacy and numeracy domains from the set of items on job requirements. However, this would save only small amounts of time in the interview, and some items on literacy would need to be covered anyway as an ingredient of a measure of communication skills.

## Foreign Language Skills

The DIALANG assessment system has been described elsewhere (See Annex A). Because this system has developed a common framework for assessment of foreign language skills, it is suitable in principle for further development for use in a European assessment. We consider that the validity and the reliability of this system to be both adequate for such a purpose. Notwithstanding, this conclusion holds the fact that national assessments were not part of the design objectives of the system.

### *Adaptation Needed*

The major element of further work would be to adapt the process of administration of the DIALANG tests to the context of the survey situation, with an interviewer, a respondent, and a limited amount of interview time. The latter constraint poses a methodological dilemma for any strategy to use this system in national and international skills assessment. For each language, there are five skill domains to be assessed: listening, reading, grammar, writing and vocabulary. For each of these domains, something like a minimum of 20 minutes interview time is involved. If all five domains were to be assessed, the test would take up to two hours, with serious consequences for response rates and no room for assessment of other skills, in the context of a one-visit, integrated process of data collection as we have advocated.

There is no easy solution to this dilemma. If foreign language skills in the adult population are to be assessed, three alternatives could be considered, each with advantages and disadvantages:

- A. Just one skill domain (perhaps vocabulary) could be assessed, alongside the other non-language skills, in the integrated process that we advocate. This might need to be used with one section of the sample, or could be considered a voluntary element for part of the sample. By including the language skill with just one part of the sample, this leaves room for assessing other skills for other parts of the sample, but the associations of language skill with the other types of skills can still be estimated. The advantage of this option would be that it would maintain commitment to a single-visit integrated process of data collection. The disadvantage is that the assessment would lay itself open to criticism as measuring too narrow a set of foreign language skills, especially since there is evidence that no single domain correlates well with the others.
- B. Three foreign language skills could be assessed, in which case we would recommend that these be listening, reading and grammar, since the combined proficiency in these is known correlate well with vocabulary and with speaking. These three foreign-language skill domains would occupy approximately one hour of interview time, and would inevitably crowd out other skill domains from the assessment. Adopting this option would follow if the policy and research value of national-level data on foreign language skills were to be given a very high priority.
- C. A third alternative could also be considered if national-level assessment of foreign language skills were to be afforded a high priority: a separate assessment could be carried out, apart from the main skills assessment. A separate assessment could use all five foreign language domains, and facilitate a thorough assessment of the effectiveness of foreign language testing in the educational systems of member countries. However, such a process would sacrifice the ability to estimate directly the association of foreign language skills with other skills, and would add greatly to the overall cost.

A further issue for resolution would be the choice of foreign language, which will vary across countries. We would expect that in most cases, the language would be English. For English-

speaking people the language of choice would be French. An alternative would be to allow respondents to be assessed in the foreign language of their choice, taken from among the official European Union languages. The DIALANG initiative offers unique opportunities in this respect, as it already allows testing in 12 different languages

#### **5.4. Instrument Design Process**

While we see no case for any EU institution to embark on a major programme of instrument development in order to carry out a European skills assessment, there will be a need for some development of items from national assessments in preparation for usage in an international assessment. Even those instruments considered suitable for inclusion in the assessment without amendment from the form taken in the originating project will need translation into all national EU languages.

##### **Development of Items from National Assessments**

With regard to the test elements, many have been developed with international comparisons in mind, and therefore do not require further development to make them suitable for a European assessment. The ALLS fall into this category, and if the literacy and numeracy test instruments in ALLS were adopted in some form, the main task of any project team would be to select a minimum number of instruments to provide valid measures in these domains. If, however, the IVQ were adopted, it may be necessary to develop some of the items that are currently specifically relevant only to France.

In the case of self-reported job requirement data, none of the above-discussed assessments have up till now been conducted in the context of an international assessment, and the survey questions have not been designed with the object of cross-national, and cross-language consistency. These surveys would need some adaptation to apply to a European assessment, to ensure a uniform interpretation of terms and of the response scales. We consider that response scales indicating the frequency of activities are the least likely to suffer from variations in interpretation, subject to proper translation (see below). A preliminary small validation project could be built into the piloting of the survey instruments.

##### **Translation and National Specificity**

Both survey items and test items will have to be translated into all European languages. It will be imperative to ensure as far as possible that translation does not impair the validity of the cross-national comparisons. Equally, care will need to be taken that cultural differences in data collection processes do not introduce distortions. This last risk reinforces the need for the management of consistency across nations, but also emphasises the need for genuine participation of national representatives in the design of the core instruments.

Translation from a reference language, likely to be either English or French in most cases, can generate differences in meaning. To minimise or eliminate these, back translation to the original, by an independent translator, is often recommended. However, there is evidence that back translation may not identify all such changes in meaning, especially where both translation and back translation are quite literal. A recommended alternative is double translation, where two translators generate independent versions in the target language. Any differences are then reconciled by a third.

There could be a demand for national differentiation in the design of certain parts of the assessment. One way of allowing discretion is through permitting some choice as to the inclusion of specific blocks of questions. For example, countries could be permitted to include or exclude the foreign language element. If any such choice element were to be facilitated, the discretionary blocks should be those that occur last in the interview. If core elements were to be left to the end, then the responses could be distorted in unknown ways by the presence or absence of the discretionary blocks in the assessment. Other requests for national differentiation should be resisted, as they would jeopardise the acceptability of the international comparisons.

## **5.5. Conclusion**

In conclusion, there are a number of methodological issues that need to be considered in developing a strategy for an adult skills assessment. This chapter has reviewed the relevant issues and made some specific recommendations. However, it is not possible to weigh all the options and choices that will need to be made to move any European initiative forward. As discussed in previous chapters the precise design of the assessment awaits further discussion as to purposes and goals of a European initiative and a consideration of various political and ethical issues.

## 6. References

- Ashton, D., B. Davies, A. Felstead and F. Green, 1999. *Work Skills In Britain*. Oxford, SKOPE, Oxford and Warwick Universities.
- Atkinson, R., 1998. *The New Economy Index*. Progressive Policy Institute.
- Bainbridge, S, Julie Murray, Tim Harrison and Terry Ward, 2004. Learning for employment. Second report on vocational education and training in Europe, CEDEFOP Reference Series; 51.
- Belbin, Meredith, 2003. *Management Teams. Why They Succeed or Fail*, London: Butterworth Heinemann.
- Bjørnåvold, J., 2000. Making learning visible. CEDEFOP.
- Bynner, John Morgan, 1994.
- Commission, 1993. White Paper on Growth, Competitiveness, Employment, COM(93) 700 final, Brussels
- Commission, 1995. White Paper on Education and Training -Teaching and Learning - Towards the Learning Society, COM(95)590, Brussels.
- Commission 2001, The Concrete Future Objectives of Education Systems” Brussels, January. COM (2001) 59 final.
- Commission 2002. “European Report on Quality Indicators of Lifelong Learning”. Report based on the work of the Working Group on Quality Indicators, Brussels, June.
- Commission, 2003a. “Education and Training 2010. The Success of the Lisbon Strategy Hinges on Urgent Reforms”, SEC(2003) 1250.
- Commission, 2003b. ”Implementing lifelong learning strategies in Europe: progress report on the follow-up to the Council resolution of 2002 on lifelong learning, EU and EEA/EFTA countries”; Drawn up with the assistance of CEDEFOP.
- Commission, 2003c. ”Implementing lifelong learning strategies in Europe: progress report on the follow-up to the Council resolution of 2002 on lifelong learning, Acceding and candidate countries”; Drawn up by the European Training Foundation at the request of the European Commission.
- European Council 2000. Presidency Conclusion, Lisbon.
- European Council 2002. Presidency Conclusions, Barcelona.
- Felstead, A., D. Gallie and F. Green, 2002. *Work Skills In Britain 1986-2001*. Nottingham, DfES Publications.
- Haahr, Jens Henrik, 2004. ”Open Coordination as Advanced Liberal Government”, *Journal of European Public Policy*, vol. 2, pp. 209-230.
- Hamilton, Mary, 2003. ”What Does IALS Measure? Literacies. Researching Practice. Practising Research, Fall issue.
- Howell, David R. and Edward N. Wolff, 1991. "Trends in the Growth and Distribution of Skills in the U.S. Workplace, 1960-1985," *Industrial and Labor Relations Review*, 44(3), pp. 486-502.
- IBF, Institut für Bildungsforschung, 1998. ”Assessment of Problem-solving. The Application of the Problems Solving ‘Project Approach’ of IBF in the Framework of ILSS. Unpublished paper, Bonn.
- Lundvall, B. Å., 2000. “Europe and the learning economy- on the need for reintegrating the strategies of firms, social partners and policy makers”. Dept of Business Studies, Aalborg University.
- Machin, Stephen & John Van Reenen, 1998. "Technology and changes in skill structure: evidence from seven OECD countries," IFS Working Papers 98/4, Institute for Fiscal Studies.

- Murray, T.S., 2003. "The Assessment of Literacy, Numeracy and Language at the International Level: A Review". Paper prepared for the National Academy of Sciences (NAS), March.
- Payne, J., 1999. All things to all people: Changing perceptions of 'skill' among Britain's policy makers since the 1950s and their implications. SKOPE Research Paper No. 1. ESRC funded Centre on Skills, Knowledge and Organisational Performance, Oxford and Warwick Universities.
- Radaelli, Claudio M., 2003. The Open Method of Coordination: A new governance architecture for the European Union? Stockholm: Swedish Institute for European Policy Studies.
- Salganik, Laura Hersh and Maria Stephens, 2003. "Competence Priorities in Policy and Practice", in Rychen, D. S. and L. H. Salganik (eds.), *Key Competencies for a Successful Life and a Well-Functioning Society*, pp. 13-40.
- Shapiro, H., 2000. "Knowledge show me the way", In: *Learn Now – New Skills for The Information Society*, Stu Rosenfeld (ed.).
- Standing Group on Indicators, 2003. "Implementation of 'Education and Training 2010' Work Programme. Final List of Indicators. Results of the Consultation with the Working Groups", Brussels, European Commission.
- Stasz, C., 1997. Do Employers Need the Skills they Want?: Evidence from technical work. *Journal of Education and Work*, 10 (3), 205-223.
- Stasz, C., Ramsey, K., Eden, R., Melamid, E., and Kaganoff, T., 1996. *Workplace skills in practice: Case studies of technical work*. Santa Monica, CA: RAND.
- Stasz, C., 2001. Assessing skills for work: Two perspectives. *Oxford Economic Papers*, 3, 385-405.
- Threlfall, Monica, 2002. "The European Employment Strategy and Guidelines: Towards an All-Working Society", paper presented at the Workshop on the European Employment Strategy, University of Loughborough, 26 April
- Trier, U. P., 2003. "Twelve Countries Contributing to DeSeCo: A Summary Report. In D.S: Rychen, L.H. Salganik and M.E. McLaughlin (eds.), *Selected Contributions to the 2<sup>nd</sup> DeSeCo Symposium*. Neuchatel, Switzerland: Swiss Federal Statistical Office.
- TSER, 1999. New Assessment Tools for Cross-Curricular Competencies in the Domain of Problem-solving, Final Report of project ERB-SOE2-CT98-2042. CEC, DGXII. Project Coordinator Jean-Paul Reff.
- Working Group 2003. "Basic Skills, Entrepreneurship and Foreign Languages", Progress Report, November. Brussels, European Commission.

## **Annex A: Case studies of Selected Adult Skills Assessment Initiatives**

## **NAAL and NALS**

### **National Assessment of Adult Literacy (NAAL) & National Adult Literacy Survey (NALS)**

#### **1. Project description**

##### **Goals and Ambition**

NAAL is a nationally representative and continuing assessment of the English language literacy skills of American adults age 16 and older. The National Center for Education Statistics (NCES) has conducted assessments of U.S. adult literacy since 1985. The 2003 NAAL will provide the first assessment of the nation's progress in adult literacy since 1992. In addition to describing the status and progress of literacy in the nation and in each of the six participating states (Kentucky, Maryland, Massachusetts, Missouri, New York, and Oklahoma), the 2003 NAAL will provide information about background factors associated with literacy, the skill levels of the least-literate adults, and the application of literacy skills to health-related materials.

This is a long-standing initiative, begun in 1985 with a household survey of young adults aged 21-25 that was expanded to a survey of adults 16+ in 1992 (National Adult Literacy Survey, NALS) and 2003 (NAAL).

The goals for NAAL are to 1) report on the status of adult literacy; 2) assess the changes in adult literacy since 1992; 3) identify relationships between literacy and adults' background; 4) develop partnerships with states. The US Department of Education conducted the 1992 survey at the behest of the US Congress, which required an assessment of the nature and extent of literacy among adults in the nation as part of the Adult Education Amendments of 1998. The NAAL is owned by the NCES, who would have to grant permission to use the items.

##### **Competencies assessed**

NAAL continues to use the definition of literacy underlying the 1992 NALS assessment: using printed and written information to function in society, to achieve one's goals, and to develop one's knowledge and potential. Like the 1992 assessment, the 2003 NAAL focuses on a broad range of tasks that adults perform in order to function at work, at home, and in the community.

The 2003 NAAL features two new components that enhance its ability to measure the literacy of the least-literate adults: the Fluency Addition to NAAL (FAN) and the Adult Literacy Supplemental Assessment (ALSA). These components will provide important new data on the literacy skills of those adults with the poorest text comprehension skills. Other enhancements to NAAL include a more extensive background questionnaire and the ability to provide a health literacy score.

The framework builds and expands on the framework used for the 1992 assessment. The literacy tasks are organized into three content areas--prose literacy, document literacy, and quantitative literacy. The literacy tasks will form the basis for three literacy scales, each of which ranges from 0 to 500 points.

Prose literacy tasks require demonstrated knowledge and skills in understanding and using information from texts, such as editorials, newspaper articles, poems, and stories.

Document literacy tasks seek demonstrated knowledge and skills in locating and using information found in job applications, bus schedules, maps, payroll forms, indexes, and tables.

Quantitative literacy tasks ask the reader to perform different arithmetic operations, either alone or sequentially, using information embedded in both prose and document formats. Tasks include entering cash and check amounts onto a bank deposit slip, balancing a chequebook, completing an order form, and determining the amount of interest from a loan advertisement.

The new framework will also cover the Adult Literacy Supplemental Assessment (ALSA) and the Fluency Addition to the NAAL (FAN). The FAN uses speech-recognition software to assess ability of adults to decode and recognise words and to read with fluency. The ALSA assesses the ability of least-literate adults to comprehend simple prose and documents, identify letters and numbers and read and understand the meaning of words. It incorporates some unique features, such as the use of familiar stimulus materials (e.g. packaged food products) and is contextualised (e.g. supported by visual information, logos, etc.).

The NALS background questionnaire gathers information on educational attainment, labour force participation, income and literacy-related activities. The NAAL is also including information about health-related matters.

### **Methodology**

Every participant receives the same background questionnaire, core screening items, and FAN tasks. Very low performance on the core items identifies ALSA participants. Adults performing above the ALSA cut-off score on the core items take the main NAAL. The entire interview takes an average of 90 minutes and is conducted at the respondent's home. The background questionnaire is CAPI administered. Other parts of the assessment are conducted by trained interviewers. The paper and pencil items are presented in task booklets.

The breakdown of the 90 minute assessment is as follows: background questionnaire (26 minutes); core screening items (5 minutes); main NAL or ALSA, depending on screening results (44 minutes); FAN (15 minutes).

NAAL 2003 utilises computer-assisted personal interviewing (CAPI) technologies for the background questionnaire only. The FAN uses speech recognition software to assess the ability of adults to de-code and recognise words and to read with fluency.

### **Results**

Results for the NAAL will not be available until July 2005. A number of reports have been issued on the NALS data. Sum (1999), for example, reports on the literacy skills of America's civilian labour force, including the employed and the unemployed. With respect to literacy proficiencies in the labour force, the analysis found that persons participating in the labour force scored higher on all three literacy scales than persons outside the labour force. Most persons in the labour force (43 per cent) scored at the two lowest literacy levels, while only 3-5 percent scored at the highest levels (level 5). This finding produced significant alarm among policy makers.

Those employed full-time had higher literacy levels than those employed part-time and both of these groups had higher literacy levels than unemployed persons. Men and women had similar literacy levels, although women scored higher than men on prose literacy, and men higher than women on quantitative literacy. Mean literacy scores of the full-time employed were correlated with educational attainment. Literacy proficiency varied with occupation and industry: for example, highest literacy rates were obtained by workers in finance, real estate, public administration, and

real estate industries, while workers in goods producing industries (manufacturing, mining, agriculture, construction) had the lowest proficiencies, on average. However there was large variability across employment sectors. Literacy proficiencies of the employed were strongly and positively related to weekly and annual earnings.

## **2. Project evaluation**

### **Utility/usefulness for policy development or other purposes**

According to Murray (2003) the NALS was the first assessment to use several technical approaches together, such as employing a BIB design, IRT scaling, conditioning, plausible values, proficiency levels and model-based imputation to minimise bias associated with having insufficient cognitive data for some respondents. Because of the huge sample size (13,600 individuals), NALS provided very rich data for exploring the social and economic causes and consequences associated with levels of literacy. For this reason, the IALS was explicitly linked to the NALS scales and NALS performance standards were adopted.

The NALS has had some impact on public policy in the US. The finding that a large proportion of citizens had lower level literacy skills caused some alarm. This has resulted in a study to re-evaluate the performance levels in the NALS (discussed below) and to hire new contractors to carry out the NAAL. There is extensive research on the NALS results (many reports available on the website) and many of these discuss policy implications.

In addition to describing the status and progress of literacy in the nation and in each of the six participating states (Kentucky, Maryland, Massachusetts, Missouri, New York, and Oklahoma), the 2003 NAAL will provide information about background factors associated with literacy, the skill levels of the least-literate adults, and the application of literacy skills to health-related materials. Similar state-level reports are available for NALS as well.

The literacy items developed in NALS (later used in IALS) are being used in another study--the Adult Education Program Study (AEPS)--which aims to provide national level information about adult education programmes and their participants. It will administer a questionnaire to a representative sample of adult education providers to gather information about the characteristics of programmes and services. It will select a sample of these institutions and assess the literacy skills of adult participants. Assessments will be conducted in English and Spanish to compare literacy outcomes in Spanish and English for Spanish-speakers. This study is under way and data will be available later in 2004. It is the first study to provide comprehensive information adult education programmes and on the literacy skills of those enrolled in adult education.

### **Quality and relevance of skills definitions**

The NALS definitions were developed by Irwin Kirsch at Educational Testing Service and represent a step forward in definitional terms. The following definition of literacy (first uses in the 1985 young adult survey) was adopted: Using printed and written information to function in society, to achieve one's goals, and to develop one's knowledge and potential. This definition extended beyond simple decoding and comprehension to include a broad range of skills that adults use in accomplishing many different types of tasks associated with work, home and community (Kirsch et al., 2001). The definitions of prose, document and quantitative literacy adopted reflect

this notion that an ordered set of literacy skills appear to be called into play to accomplish diverse types of tasks. This idea also found empirical support in the IALS analyses.

### **Validity of assessment methods**

As yet, there is no information available for NAAL on validity. The NALS adopts Item Response Theory (IRT), which measures performance relative to the difficulty of the item, taking into account a person's ability and the item difficulty. Responses are evaluated as the probability that a person will give a correct response. Classical test theory assesses performance on a prescribed set of items. There is some controversy over the validity of estimates derived from IRT (summarised and discussed in the IALS reports, which used the same methodology).

The NALS adopted numerous procedures to enhance validity, including 400 trained interviewers; various quality control checks; validation of 10 percent of each interviewers' cases; observations of interviews; tape recording of interviews. Inter-reader reliability of coded items was 97 percent (A number of working papers are available on the website which review/report NALS results and technical issues in more detail.)

These are both very large surveys, with over 20,000 respondents in the 1992 survey (NALS), involving 400 trained interviewers plus field supervisors. The 1992 and current surveys incorporate extensive training and fieldwork procedures, validity checks, etc. The 1992 survey used some bilingual interviewers (English/Spanish). In addition specially trained interviewers conducted surveys with a sample of prison inmates. The prison sample required special data collection procedures. The methodology appears sensitive to these challenges.

Various goodness of fit models were used to test construct and discriminate validity of the three literacy scales. Analyses showed the 3 scales were highly related--i.e., performance across the three is highly correlated--but there there was still room for some differences in group profiles. However, some questions have been raised about the construction of the performance levels (Levels 1-5) associated with the literacy scales. When 1992 survey results were released, the reported performance levels led to the possibly incorrect impression that nearly half of American adults are illiterate. This interpretation of the survey results as well as the procedures that were used to develop the performance levels have been the subject of much controversy and criticism. In particular, there are questions about whether individuals who have the skills described for the two lowest levels are in fact "illiterate."

As a result, the Department of education supported an initiative to explore alternative standards for the performance levels as well as alternative methods for setting the standards to be carried out by a committee appointed by the Board on Testing and Assessment, National Research Council. This work began in 2002 and is ongoing.<sup>19</sup> Although the NALS profile was not intended to answer the question of how much literacy is enough, the U.S. National Education Goals Panel adopted the percentage of adults at or above Level 3 on the NALS prose literacy scale as an indicator of progress toward Goal 6: Adult Literacy and Lifelong Learning, thus creating a *de-facto* NALS-based performance standard for adult literacy (NCES, 2000). Many in the adult literacy community as well as in other countries have subsequently adopted this as the "standard" of literacy.

---

<sup>19</sup> The Committee on Performance Levels for Adult Literacy: Letter Report was completed in 2003 but has not yet been publicly released by the National Research Council.

### **Reliability of assessment methods**

NALS employed a nationally representative four-stage stratified sample, plus various statistical weightings to prevent bias in estimates. Corrections for missing data for both not reached and omitted responses were conducted as well as analyses of reasons for missing data/non-response. There may be questions about the comparability of the trend data as the items in NAAL will be modified versions of NALS. Otherwise, the NALS survey used sophisticated/extensive methods to enhance reliability (for details see Kirsch et al., 2001).

### **The quality of data collection**

The NALS hoped to enhance response rates by offering survey respondents a \$20 incentive. Two experiments were conducted in NALS to assess the various impacts of monetary incentives on the quality of the data. A \$20 incentive was found to increase the response rate and decrease the survey costs. Incentive was particularly effective for harder to reach populations. The NAAL survey will also use an incentive.

## **3. Lessons Learned**

One lesson from the NALS initiative concerns the political sensitivity that may be attached to the results of population surveys. The popular press in the U.S. was such that the sponsors of the research are conducting further investigations into the validity of the performance levels used in NALS. Had the results indicated that 43 percent of Americans performed at Level 5 (not Levels 1-2), it is perhaps unlikely that the Department of Education would go to this trouble.

The NALS also provided useful data concerning the use of an incentive to increase response rates. The study indicated that even a small incentive of about \$20.00 is cost effective overall. NAAL makes use of some innovative methods, including speech recognition software to assess the ability of adults to de-code and recognise words and to read with fluency. It includes special measures for those with very low levels of literacy. This appears to be an advantage over the IALS, which was criticised by some as not suited to individuals with only rudimentary literacy skills. The success of these innovations, however, is unknown at this point in time.

The FAN and ALSA might be applicable in other cultures. The ALSA was initially designed for an assessment of ESL literacy learners in the United States, and most of the ALSA respondents are not native English speakers. It incorporates information about adult familiarity with items, as well as their ability to read text from the item.

## **4. References**

Persons contacted: Mark Kutner, American Institutes for Research, Washington, D.C..  
mkutner@air.org.

Kirsch, I. et al. (2001). *Technical Report and Data File User's Manual for the 1992 National Adult Literacy Survey*. NCEs 2001-457. Washington DC: U.S. Department of Education, Office of Educational Improvement.

Murray, T.S. (2003). *The Assessment of Literacy, Numeracy and Language at the International Levels: A Review*.

Sum, A. (1999). *Literacy in the Labor Force. Results from the National Adult Literacy Survey*. NCES 1999-470. Washington DC: U.S. Department of Education, Office of Educational Improvement.

[www.nces.ed.gov/naal](http://www.nces.ed.gov/naal) (contains information on current and previous surveys)

[www.nrc.org](http://www.nrc.org) (U.S. National Research Council).

# Adult Life Skills and Literacy Survey (ALLS)

## 1. Description

### Goals and ambitions

The Adult Life Skills and Literacy Survey (ALLS) is a comparative survey to measure the distribution in the adult population of a range of skills thought to be important to economic and social success.<sup>20</sup> It is patterned on the International Adult Literacy Survey (IALS) and involves the administration of direct performance tests to a representative sample of adults aged 16-65. ALLS is meant to "inform public policy in a number of related areas including education, labour market policy, human resource development and social development" (Giddings and Barr-Telford, 2002).

ALLS has three sets of objectives. The short-term objectives are to field a pilot study and a main study in a heterogeneous group of countries. The medium term objectives roughly paralleled those of the original IALS study: to profile the distribution of prose literacy and document literacy in the adult population, to profile the distribution of numeracy and problem solving and to determine the inter-relationship of each of these skills as well as their relationship to prose and document literacy; to determine the relationship of each of the tested skills to individual economic and social success; and to identify sub-populations whose performance places them at risk. The long-term objectives are to shed light on the causes and consequences of the observed skill distributions, contribute to the literature on the basis of human cognition, and foster continued international co-operation on the design, implementation and analysis of data on the distribution and co-variables of skills.

ALLS is a joint initiative of Statistics Canada, the US National Center for Education Statistics (NCES), the United Nations Educational, Scientific and Cultural Organisation (Regional Office for Latin America and the Caribbean), and the Organisation for Economic Cooperation and Development (OECD). The development work was initially funded by Statistics Canada and NCES but both have agreed to relinquish control of the longer-term R&D agenda in the area of adult skills to a more co-ordinated effort to be led by the OECD. According to the planning report, the OECD programme of work is to include two elements—one aimed at developing a coherent theoretical and conceptual basis for assessing skills (under the DeSeCo project) and a second activity aimed at developing instrumentation for the next generation of adult skills assessment for implementation in 2010.

Statistics Canada/the Queen in Right of Canada owns the ALL tests and anyone may use the items without royalty subject to the following three conditions: the use is for non-commercial ends, the findings are made available in the public domain, and the item pool is kept secure. Statistics Canada would determine which items are retained for the purposes of establishing trends in further assessments and the balance of items are placed in the public domain. This means that the European Commission would be free to use the test as is or to adopt some part of it. Statistics Canada would be willing to consult as to which items are put into the public domain.

---

<sup>20</sup> Initially, the survey was called the International Life Skills Survey (ILLS) and some early documents retain this title.

### **Competencies assessed**

ALLS assessments covered either three or four domains: prose literacy, document literacy and numeracy on all assessments; a country could also choose to add problem-solving and reasoning. The numeracy domain was expanded from the IALS to refer to the knowledge and skills required to effectively manage the mathematical demands of diverse situations to interpret, apply and communicate mathematical information in commonly encountered situations. The problem solving assessment involves the ability to apply cognitive processes toward determining a solution when that solution is not immediately obvious to the problem solver.

Originally the ALLS planned to study prose literacy, document literacy, numeracy, team work, problem-solving, practical cognition, and working with information technology. Additional development work was conducted on several of these skill frameworks to enhance the chance of generating valid, reliable, and comparable skill profiles at the international level. Teams responsible for numeracy, problem-solving, teamwork, practical cognition and ICT were funded to refine their frameworks and to collect sufficient empirical data to demonstrate the measurement properties of the proposed measures trans-nationally. In meetings held in 1998, it was judged that while the proposed frameworks for problem-solving and attitudes toward teamwork were adequate, the approach to measurement failed to yield data of sufficient quality. The instrumentation for measuring computer literacy was also judged inadequate.

As a result, new development teams were recruited and funded by Statistics Canada, NCES, and the governments of Sweden and Luxembourg. Small scale piloting conducted in pairs of countries failed to yield measures of sufficient reliability for practical cognition, team work, and ICT. Extensive research was carried out to develop a conceptual framework and assessments of teamwork skills, but the results were judged to not reach a useful standard of reliability or validity for use in the final survey (For more details see reports available at [www.ets.org/all](http://www.ets.org/all)). In the end, measures of familiarity and use of ICT were incorporated into the background questionnaires. (Items related to attitudes toward teamwork were initially included, but later eliminated after further pilot study).

The background questionnaire collects respondent information, general and linguistic information, parental information, labour force activities, literacy and numeracy practices at work, household information and income, participation in education and learning, social capital, well-being, and information and communication technology literacy (ICTL).

Prose literacy and document literacy are measured on IALS scales to provide measures of trend for those countries that participate in both surveys. Like the IALS, the ALLS employs a cross-sectional study design that can provide a snapshot of the distribution of skills in the domains measured.

### **Applied Methods**

Like the IALS, the ALLS employs direct performance tests. The ALLS is administered through a household survey of adults aged 16-65 to include all civilian non-institutionalised persons residing in the country at the time of data collection. A country may include other sub populations provided that its sample design include any necessary augmentation of the sample size to accommodate the analysis requirements for the additional sub-populations.

Seven countries participated in first round of data collection-- Bermuda, Canada, Italy, Mexico (Nuevo Leon), Norway, Switzerland, and USA. These countries fielded the ALLS pilot study in 2002 and the main data collection in the first and second quarters of 2003.

The data collection components include a screening questionnaire, a background questionnaire, and a set of simulation tasks. Because the test is relatively expensive (average cost for IALS in Canada was \$160US per person), participating countries were given the option of fielding a three- or four-domain survey as follows: 1) a 28-page spiralled design booklet that assesses performance in four domains—prose literacy, document literacy, numeracy, problem-solving/analytic reasoning or 2) a 18-page spiralled design booklet that excludes the problem solving domain. The prose and literacy tasks are similar to IALS tasks. Assessment is carried out by trained interviewers.

As to timing, the background questionnaire is designed to take about 27-30 minutes. The test booklets were built to include two thirty-minute blocks per person with the block pairings designed to give the best information per value domain and for estimating a partial inter-skill covariance matrix.

Countries may choose a computer assisted interviewing method in the collection of the background questionnaire information or a paper and pencil method. Either in-person or telephone interviews were used to screen-in appropriate respondents; a computer assisted data collection method may also be used at this stage.

The study originally intended to administer a computer-based test to a nested sample of workers within firms so that explicit statistical linkages would be available to isolate the impact of observed skills on productivity and profitability. However, a second planning meeting determined that fielding a computer-based test was beyond the financial and technical capability of many of the prospective participants. It was determined that the assessment should use paper and pencil rather than computer-based tests, only six domains would be measured, rather than the eight originally proposed, and the test would be administered to representative samples of adults drawn from households rather than from workers within firms.

## **Results**

According to the planning report, the pilot survey was to be conducted from October to December 2003. The results from the pilot study are not yet available, but international comparative data from the first round of ALLS collection are due to be published on February 16, 2005. A second round of data collection planned with about 10 countries will be carried out in 2005 and 2006. After that, there is an agreement with OECD to end the ALLS in OECD countries to "make room" for an all-OECD adult assessment in 2010. Statistics Canada is continuing to work with non-OECD countries within the context of UNESCO's LAMP project.

## **2. Evaluation**

### **Utility/usefulness for policy development or other purposes**

ALLS will be comparable to a number of assessment efforts. The prose and document literacy scales in ALLS are measured on IALS scales, and thus provide a measure of trend for those countries participating in both assessments. ALLS is closely related to PISA, which measures

learning outcomes for 15-year-old students in all OECD member countries. PISA provides a basis for assessing and monitoring the effectiveness of educational systems at national levels and for making international comparisons. While PISA is concerned with the skills of those completing compulsory schooling for whom the goal is to reach an appropriate level of development of adults life skills, the ALLS is directly concerned with the assessment of skills of adults. There is correspondence between PISA Reading Literacy and ALLS Prose and Document Literacy. Prose literacy items have been embedded in PISA in an attempt to place 15-year-olds on the Prose Literacy scales. In addition, there is recognition of the importance of non-continuous texts or documents in PISA in that they represent some 33 percent of the Reading Literacy 'assessment in PISA. ALLS is also related to the Definition and Selection of Competencies (DeSeCo), which is lead by Switzerland with the support of the US, under OECD's Indicators of Education Systems (INES) project. DeSeCo seeks to conceptualise the different competencies needed for individuals to lead an overall successful and responsible life and for society to face the challenges of the present and the future.

ALLS has made a contribution with the addition of new direct measures of numeracy and problem solving skills. The study results will provide a profile of numeracy, problem solving, document literacy and prose literacy in the adult population of participating countries, determine the relationship of these skills to one another, and to individual economic and social outcomes. These analyses will be important for further understanding of the relationship between skills, economic growth and prosperity and to factors underlying social well being and social cohesion.

#### **Quality and relevance of skills definitions**

There was significant development work to develop the new domains, numeracy, problem solving, practical cognition, teamwork, and ICT skills. The ALLS website provides documentation for development of prose and document literacy, numeracy, problem solving, teamwork, and ICT.

Prose and document literacy have been extensively evaluated in IALS. Numeracy is defined as the knowledge and skills required to effectively manage the mathematical demands of diverse situations. Problem solving is defined as the ability to solve problems by clarifying the nature of the problem and developing and applying appropriate solution strategies. Teamwork is defined as the competencies needed for effective participation in a team, including interpersonal communication skills, adaptability, flexibility, mutual trust, and willingness to work with others. ICT skills were indirectly measured through the background questionnaire.

The ICT items are intended to provide information on four dimensions: incidence (i.e., does the person use a computer, the internet and various forms of ICT?), frequency (i.e., how often does a person use a computer?), criticality (i.e., how critical or useful is the application of ICT to everyday life?), and complexity (i.e., how complex are the tasks carried out?) (Lowe and McAuley, 2000).

Legitimacy is discussed as ensuring comparability across countries, and with reference to ensuring that measurement is linguistically, culturally and geographically appropriate. The prose and document literacy items included some new items to ensure cultural and linguistic representativeness of the populations to be assessed. There were a series of development meetings and various reviews by study managers in all participating countries to rate every item based on cultural concern and possible translation and adaptation problems in each country. These procedures expanded the item pool and were intended to address some of the criticisms directed at IALS items.

### **Validity of assessment methods**

Pilot studies were conducted for all new items and in new domains. Assessments for team working, ICT and tacit/practical knowledge failed to reach validity standards and thus were not included in the main study.

As with IALS, the IRT design provides a full covariance matrix for the literacy, numeracy and problem-solving scales, thus allowing the estimation of the discriminant validity among these scales.

The pilot testing of the new problem solving items showed that the difficulties predicted from theory were being confirmed empirically and that the results were not perfectly correlated with the other scales. Current data from Norway, Italy and the US indicate that the items are picking up something orthogonal to the other skills -- the correlation between problem solving and reading is about .80.<sup>21</sup> However, the developers have some concerns about the measurement, as there are only a few items per person and as problem solving relies on reading. For this latter reason, it may be difficult to interpret scores for poor readers. This means that it may be difficult to detect variance in problem solving scores for poor readers, and therefore difficult to conclude unequivocally that people with low reading scores have low problem solving skill. This may not be a problem educationally, since teaching of reading would likely come before teaching of problem solving. However, it causes problems with respect to interpretation and communication of the results -- a result of large percentages of persons lacking problem solving skills belies the fact that many would be successful solving problems every day that do not involve reading.

In the planning report, validity is also discussed with respect to developing background questions, for example in providing accurate and parallel descriptions of occupations, educational background, vocational training, and on the job training experiences required a careful analysis for each of the participating countries. Only variables that had the same meaning across all participating countries will be used in the comparative analysis.

A variety of statistical techniques will be used to establish the validity and comparability of ALLS performance data. These are fully discussed in the planning report, and included techniques to check the validity and results both within and across countries, traditional item statistics to identify initial problems in comparability, IRT scaling, differential item functioning techniques to see if any questions are operating differently across countries. Tasks that are not operating in the same way in all participating countries will not be included in summary statistics. Additional techniques are applied to describe the nature of results both across and within countries. There are an extensive set of standards that cover planning, data collection, sample design, etc. that participating countries must follow. (These are summarised in Appendix C of the Second Round Planning Report and available on the website).

### **Reliability of assessment methods**

As discussed above, assessments for some new domains were developed and pilot tested, but their reliability was judged insufficient. Measures of practical cognition were developed and pilot tested in Spain and the US but were not included in the ALLS study because of concerns with respect to the interpretability of the estimates flowing from the scoring system. Two different groups were commissioned to develop a framework and behavioural measures for team work skills, but these did

---

<sup>21</sup> Canada and Switzerland are also fielding the problem solving measures.

not provide items of sufficient quality to support valid and reliable scales. Large scale pilot testing by countries participating in the first round provide evidence that the ALLS instruments yield reliable, valid, comparable and policy relevant data.

Participating countries were given guidelines with respect to sample sizes needed to make different estimates of interest (e.g., at least three educational levels, at least three age groups, and two gender groups), with assumptions about response rates. Sample designs must be probability-based. National study teams are required to engage a survey statistician to participate in the development and implementation of the survey design.

Extensive procedures were put in place to ensure scoring reliability within countries, such as re-scoring of assessment data, quality control checks, etc. In our review of the documents, it appears that many efforts were taken to ensure high reliability.

### **The quality of data collection**

Many of the problems of possible comparability and possible bias in the IALS data can be traced back to a failure by countries to adhere to the agreed procedures for the survey administration. As a result the ALLS planning documents devote more attention to ensuring that minimum standards are met. These include standards and guidelines for study design and implementation and quality assurance specifications. These documents set out strict requirements for acceptable data collection agencies. As the ALLS is being implemented, there is no information yet available concerning the actual data collection or on response rates.

## **3. Lessons learned**

The ALLS attempts to build on the IALS by incorporating some of the IALS literacy measures (thus permitting some longitudinal analysis for those countries who participate in both) and by adding assessments in two new domains—numeracy and problem solving. Although the results of the first wave will not be available until next year, the initiative thus far points to some useful lessons.

The ALLS experience highlights the problems that can arise in developing direct measures of other skills. Originally ALLS planned to include measures of prose literacy, document literacy; numeracy, team work, problem-solving, practical cognition, and working with information technology. Additional work was conducted on several skill frameworks to enhance the chance of generating valid, reliable, and comparable skill profiles at the international level. As a result, teams responsible for numeracy, problem-solving, teamwork, practical cognition and ICT were funded to refine their frameworks and to collect sufficient empirical data to demonstrate the measurement properties of the proposed measures trans-nationally. The reports from this development work are publically available and can contribute to further efforts to develop measures of skills in domains that are of policy interest.

Cost of direct assessment of skills remain high, because international projects of the scale of ALLS and IALS incur significant overheads. These overheads cover management, design, and statistical analysis throughout the project. Participants in ALLS are responsible for financing their own national item pilot and the work of any experts they involve. The experience of the IALS suggests that contributing financially to the undertaking has a salutary impact on countries' commitment.

ALLS participants are required to contribute roughly \$150,000 US toward international overheads for the pilot and main study over a three or four year period. In seeking funding from international agencies to help offset costs, Statistics Canada is pursuing a strategy that might be useful to consider in future efforts. If such funding materialises, countries will be able to receive either a rebate or additional analyses.

A final lesson that is carried forward from IALS to ALLS concerns the necessity of direct assessments in some particular policy contexts. Although there are relationships between literacy proficiency and educational attainment, that relationship is complex. The IALS data indicate, for example, that many adults manage to attain a high level of proficiency despite low levels of education. Conversely, others have lower literacy skills in spite of high education levels. This finding suggests that education is a poor proxy for skill and that direct measures are important for understanding the relationships between skill and economic and social prosperity.

## 4. References

### Interviews conducted/persons contacted

Scott Murray, Director General, Institutions and Social Statistics, Statistics Canada  
17 “a” R.H. Coats Building, Tunney’s Pasture, Ottawa, Ontario, K1A 0T6, 23 October 2003, 10 and 17 May 2004.

### Literature

*Adult Literacy and Life Skills Survey International Planning Report*. Second Round. Ottawa: Statistics Canada, 2002.

Final Report for Analysis of Teamwork Skills Questionnaire (1999). ([www.ets.org/all](http://www.ets.org/all))

Giddings, D. and Barr-Telford, L. (2000). *Skill Development and Public Policy*. Statistics Canada, National Center for Education Statistics, and OECD.

Lowe, G. and McAuley, J. (2000). *Adult Literacy and Lifeskills Survey: Information and Communication Technology Literacy Assessment Framework* ([www.ets.org/all/ICTL-2nd-framework.pdf](http://www.ets.org/all/ICTL-2nd-framework.pdf))

Baker, D.P. et al. (no date) *Teamwork*. Status Memorandum. ([www.ets.org/pub/corp/team/pdf](http://www.ets.org/pub/corp/team/pdf))

[www.ets.org/all](http://www.ets.org/all) (Last updated September 2003: for results of the pilot test and other documentation, including reports on domains where new assessments were developed and piloted, but due to methodological problems were not included in the survey.)

# International Adult Literacy Survey (IALS)

## 1. Description

### Goals and ambitions

The goals of the International Adult Literacy Survey (IALS) are to: (1) provide comparable estimates of levels and distributions of literacy skills in adult populations; (2) understand the nature and magnitude of literacy issues faced by countries and to explore insights into the factors that influence the development of adult skills in various settings, contribute toward understanding the demand and supply of skills in the global, knowledge based economy. To provide a useful tool for policy makers for analysis and for crafting policies and programmes that can contribute to economic and social progress.

The IALS was a large-scale co-operative venture by governments, national statistical agencies and the OECD. Development and management of the survey was coordinated by Statistics Canada and Educational Testing Service (ETS). The US National Center for Education Statistics (NCES) provided substantial input along the way. (For a brief history of IALS see Murray, 2003).

IALS has been administered in three waves of data collection: 1994 (Canada, France, Germany, Ireland, the Netherlands, Poland, Sweden, Switzerland, USA); 1996 (Australia, Flemish community in Belgium, Great Britain, New Zealand, Northern Ireland; 1998 (Chile, Czech Republic, Denmark, Finland, Hungary, Norway, Italy, Slovenia, Italian speaking regions of Switzerland). Limited data on Portugal was included in round 3 analysis. France withdrew from study in 1995 over concerns of comparability. Japan, Malaysia, Mexico and Canary Islands (Spain) have experimented with IALS derived instruments, but are not included in reports because of limited and non-representative samples.

Educational Testing Service owns the items, and would need to be consulted as to their use. Our information indicates that the items can be used by others provided that the use is for non-commercial ends, the findings are placed in the public domain, and the item pool is kept secure.

### Competencies assessed

Literacy is defined as a particular capacity and mode of behaviour: "the ability to understand and employ printed in daily activities, at home, at work and in the community -- to achieve one's goals and to develop one's knowledge and potential" (OECD, 2000, p. x). Literacy is measured operationally in terms of 3 domains -- prose literacy, document literacy, quantitative literacy-- each encompassing a common set of skills relevant for diverse tasks. Prose literacy tasks assess the knowledge and skills needed to understand and use information from texts, including editorials, news stories, brochures and instruction manuals. Document literacy measures involved locating and using information contained in various formats, such as job applications, payroll forms, transportation schedules, maps, tables, and charts. Quantitative literacy is measured via application of arithmetic operations, either alone or sequentially, to number embedded in printed materials (e.g. balancing a chequebook, figuring out a tip, determining amount of interest on a loan)

For each domain, literacy proficiency is expressed by a score defined as the point at which a person has an 80 percent chance of successful performance from among the set of tasks of varying

difficulty included in the assessment. Individuals are assigned to one of 5 levels of literacy based on their scores, with level 1 indicating very poor skills and levels 4 and 5 indicating individuals who demonstrate command of higher-order information processing skills.

### **Applied Methods**

IALS employs a face-to-face household surveys, conducted by trained interviewers. Respondents are first asked a series of background questions (e.g., demographic details, work history), then are given a booklet containing six simple tasks. If a respondent fails to complete at least two tasks correctly, the interview is terminated. Otherwise, the respondent receives a separate booklet with a much larger variety of tasks. The assessment is not timed, and respondents are urged to try each exercise. Thus, respondents receive maximum opportunity to demonstrate their skills. In addition, individuals provide self reports: they were asked how frequently at work they engaged in literacy activities with various kinds of texts, about their writing and reading practices at home, and to rate the adequacy of their own skills.

Data collection was the responsibility of each of the participating countries in each wave of the survey. There was an extensive process of pilot testing, translation, cultural adaptation of items, with these processes becoming more specific with successive waves of the IALS .

Nationally representative samples of the civilian, non-institutionalised adult population aged 16-65 in each country were surveyed. Countries were encouraged to field sample sizes large enough to yield 3000 completed cases to ensure reliable estimates of literacy profiles. Individual countries were free to sample younger or older populations.

### **Results**

The IALS results have been widely reported (OECD and Statistics Canada, 1995; OECD and HRDC, 1997; OECD and Statistics Canada, 2000). As the data has been made available to academic researchers, a number of additional analyses have been carried out. In addition, several countries have made use of IALS data. These uses are discussed further in the next section.

Briefly, the IALS data illustrate how skills are distributed nationally and internationally, what determines the attainment of higher levels of literacy, and what are its broader social and economic outcomes and benefits. IALS permits analysis of the distribution of skills within and between nations. Scores on the three literacy dimensions are depicted in scales that range from 0 – 500 points. Using these scales, mean scores for countries can be compared, as well as the spread of scores.

Item response theory (IRT) scaling procedures were used to establish scales for a set of tasks with an ordering of difficulty that is essentially the same for everyone. First the difficulty of tasks is ranked on the scale according to how well respondents actually perform them. Next individuals are assigned scores according to how well they do on a number of tasks of varying difficulty. The scale point assigned to each task is the point at which individuals with that proficiency score have a given probability of responding correctly. IALS used an 80 percent probability of correct response. Thus, individuals estimated to have a particular scale score perform tasks at that point on the scale with an 80 per cent probability of a correct response. It also means they have a greater than 80 percent chance of performing tasks that are lower on the scale. It does not mean, however, that individuals at a given level can never succeed at tasks with higher difficulty values. Once scores are placed

along each of the scales using the criterion of 80 percent, it is possible to see to what extent interactions among task characteristics capture the placement of tasks among the scales. Analyses of the task characteristics reveal the information processing skills needed to perform the task, and the order of these skills. To capture the order, the scale is divided into five levels reflecting the empirically determined progression of information-processing skills and strategies. The levels were selected not as a result of any statistical property of the scales, but rather as a result of the shifts of skills and strategies required to succeed at various levels (see OECD, 2000, p. 94 for further detail).

Different types of results have been reported in IALS publications. For example, the following results are reported in OECD (2000). Scaled literacy scores are used to compare differences in literacy skills within and between countries. Low skills are found not only among marginalised groups, but among significant proportions of adult populations in all countries surveyed, even the most economically advanced. Differences are compared to education level. The largest differences between countries occur for people with the least formal education. Education attainment is the most important predictor of literacy proficiency—on average people increase their literacy scores by about 10 points for each additional year of schooling. In addition to education level, a person's home background and particularly the level of education of the parents, influences literacy. Literacy skills appear to be maintained and strengthened through regular use. While schooling provides an initial foundation, informal learning and the active use of literacy skills in everyday use are important for attaining higher levels of proficiency.

Across countries, higher levels of literacy in the workforce are associated with larger proportions of knowledge jobs in the economy. Literacy skills positively influence the probability of having a white-collar high skilled position and negatively influence the probability of being unemployed or have a blue-collar position. Literacy enhances career prospects. In most countries, low scores are associated with long-term unemployment. Benefits of literacy differentially accrue to individuals with different levels of education attainment—those with tertiary education experience higher benefits than those with secondary education.

Wage analyses indicate that educational attainment is the most important determinant of earnings in most countries. Literacy proficiency has a substantial effect on earnings, partly dependent on level of education. But in many countries literacy also has an independent net effect on wages. There are large differences between countries in how much education is rewarded in the labour market and how much they pay for skills and experience. A number of non-market benefits are associated with literacy skills—high literacy is associated with better health outcomes and public and civic participation.

Although the IALS results have attracted a great deal of interest from policy makers, analysts and the popular press, the study did not cover some important questions of interest, for example the relationship of literacy skills to other skills thought to be important to workforce productivity and labour market success.

## **2. Evaluation**

### **Utility/usefulness for policy development or other purposes**

IALS appears to have had some impact at international and national levels. A recent summary provides several examples (Murray, 2003):

- OECD: IALS has been used extensively in OECD publications, most notably in successive editions of *Education at a Glance* (2002) and in *Human Capital Investment: An International Comparison* (1998).
- OECD adapted the IALS literacy assessment framework for PISA to express PISA results on IALS scales. Except for 3 countries the PISA reading estimates for 15 year olds and IALS prose literacy estimates for 16-25 year olds are basically identical in 19 of 22 countries, and thus provide independent validation of the frameworks and methods. In three countries--France, Sweden and Germany--the estimates of skill are discrepant. IALS appears to underestimate French skills, while it appears to overestimate proficiency in Sweden and Germany. The latter seems to confirm suspected deficiencies in sample coverage and response rates in those two countries.
- UNDP has used IALS to produce an alternative index of human development for selected countries in their annual *Human Development Report*.
- UNICEF, through the Innocenti Center in Florence, has used IALS as a measure of equality in educational output across all social classes.
- The UNECE and ECLAC have used the data to inform educational labour market policies in their respective regions.
- UNESCO, through the UNESCO institute for Education in Hamburg, have published on the patterns of adult education and training using IALS data.
- World Bank and European Bank for Economic Development and Reconstruction have used the data to inform a range of educational policy and economic policies in the transition countries of Eastern Europe.
- The European Union has used IALS to fill important gaps in their data related to lifelong learning.

Academics have made extensive use of the dataset, with some 100 research monographs written to date. These tend to focus on the relationship of literacy skills to economic outcomes at the individual level and on patterns of participation in adult education and training and their relationship to skill (Murray, 2003 provides specific references for each topic).

Murray (2003) summarised the use and impact of IALS for different countries as follows:

- Of all countries, Canada has made the most use of IALS. The Canadian government was the first to produce a national report on IALS data and has published several research monographs. This has led to a doubling of federal and provincial government investments in adult literacy remedial programmes (Murray, 2003) For example, Statistics Canada used modified IALS assessments to assess the skills of the deaf and hard of hearing, of recent immigrants to Ontario, as well as for participants in a school to work transition programme in Nova Scotia. Human Resource Development Canada has created job profiles for 250 occupations that include IALS scales, thus showing relationships between IALS scale score

proficiencies and other job skills. Other efforts seek to bridge student test scores in some of the provinces with scores on both PISA and IALS in order to gauge how the school system is performing relative to these international standards.

- Sweden. IALS has had some impact on public policy despite not producing a national report. Although they have the highest average literacy of all countries and among the least variable skill level, Swedish authorities 'were sufficiently shocked at the proportion of level 1 to institute a program that offered a full year of paid educational leave to individuals without high school graduation.'
- Norway. The gov't. was obliged to field two separate variants, one in Nynorsk and one in Bokmol. The assessment performed identically in both languages, 'suggesting considerable language transfer is operating.' IALS data resulted in a modest increase in government sponsored literacy programming.
- Denmark. The data are reported to have resulted in significant changes in workplace learning programs (Jensen, et al., 2000).
- Finland. Results have been published by Linnakyla et al. (2000). 'Despite the strong performance of Finnish adults the authors report that the results have precipitated a strong policy response in Finland.'
- Germany has yet to publish a national report (as of March 2003), although Statistics Canada/OECD have published a comparative report in German. The PISA reading literacy data have had more impact than IALS.
- The Netherlands. IALS seems to have affected public policy, as 'the equivalent of some \$70 million USD is reported to have been reallocated from the initial education system to literacy programs in the adult education system as a direct result of the IALS findings.' Data for Holland have seen extensive use by Dutch labour economists at the Universities of Amsterdam and Gronigen.
- Belgium. IALS was restricted to the Flemish community living outside Brussels. Researchers from the University of Ghent have published analyses that have been used by the Flemish Education authorities in planning adult education initiatives.
- France. Despite having withdrawn their data from the international comparative report, the government has devoted much effort to understanding IALS methods and analysing the study results. France has also experimented with alternative methods that explores the basic literacy and numeracy skills of French adults. The national statistics institute is testing assessment methods that are loosely related to those in IALS (INSEE, 2002).
- Ireland. IALS has had significant impact, as government funding for adult literacy programmes was increased. Data have been published by Morgan et al., 1997.
- Northern Ireland. Data published by Sweeney et al (1998) reveal significant differences in the proficiency of the Catholic and Protestant populations.

- Great Britain. IALS has had impact on public policy and opinion following the second comparative report and the UK's national report (Carey et al, 1997). 'Interest seemed to wane following the publication of Sir Claus Moser on the state of literacy and numeracy...and the subsequent rapid expansion of adult literacy programming and the development of an elaborate national assessment and certification system.' The ONS played a central role in the EU's evaluation of the IALS methods; the report failed to identify any significant weaknesses.
- Switzerland is the only country to have fielded IALS in three languages: German, French, and Italian. Various reports have been published.
- Italy's sample was problematic, so only a limited amount of data was included in the third international comparative report. Researchers from CEDE, the government agency responsible for the study produced a national report that contained comparative data.
- Czech Republic used IALS data to support a range of labour market development initiatives aimed at smoothing transition to a market economy. This was aided by the fact that Czech results were comparatively good when judged by international standards.
- Hungary. 'IALS results largely ignored in Hungary as they did not fit well with national beliefs about the performance of the Hungarian education system.'
- Slovenia. IALS had a 'profound impact' on a range of labour market and educational programmes, in large measure because the government wanted to assure that Slovenian workers could compete in the European labour market.
- Australia. The Australian study had the largest sample size and realised the best response rates of any country. The data have had widespread use 'despite the fact the Prime Minister of the day announced that the nation's teachers were solely responsible for Australia's mediocre performance.'
- New Zealand. Data has seen wide use by policy departments and NGOs, despite some data collection problems that under-reported demographic variables.
- USA. Although the US national adult literacy survey (NALS) precipitated a good deal of public opinion and political debate, response to the IALS findings has been muted. This may be attributed to the fact that key messages from IALS were already known from NALS and studies such as TIMSS, as well as to the fact that the US failed to publish a national report.

### **Quality and relevance of skills definitions**

IALS is a performance test that directly measures three types of literacy skills—prose, document and quantitative literacy. Test items on the IALS were drawn to represent a wide range of difficulty and everyday domains. Theoretical item difficulty was based on research on the attributes of text and tasks, and actual item difficulty and proficiency estimates were derived using advanced psychometric and statistical techniques. The IALS was built on a skill model, which relies on theories of item difficulty to support generalisations beyond the items selected for inclusion on the test. It was built upon four North American studies that embodied skill models: the Functional

Reading Study (FRS) conducted by Educational Testing Service in the 1970s; the Young Adult Literacy Survey, conducted by ETS in 1985; the Survey of Literacy Skills Used in Daily Activities conducted by Statistics Canada in 1987; and the 1990 National Adult Literacy Survey conducted by ETS for the US Department of Education.

A balanced-independent block (BIB) design was employed, so that no individual respondent took the entire pool of assessment items. This design has the advantage of providing maximal coverage of both the ability and difficulty distribution, the design does not support reliable estimates of individual scores. Rather, it provides statistically reliable estimates for population sub-groups. The BIB design also creates a problem of missing data—performance must be imputed for those items that are not taken.

### **Validity of assessment methods**

A number of quality control measures were implemented throughout the course of the IALS in order to insure that high-quality data would be obtained. Enhanced measures were also taken to further improve data quality and comparability during subsequent phases of the survey.

Validity has been discussed in several ways. First is the approach to direct assessment. Three approaches can be characterised: 1) item models make no attempt to generalise beyond the test items themselves; 2) competency models assume that general performance is perfectly correlated with performance on the items selected for inclusion in the test; and 3) skill models which rely on explicit theories of item difficulty to support generalisation beyond the items selected for inclusion on the test. IALS adopted the third model and was built upon theoretical and methodological insights from four large-scale North American (NA) surveys that embodied skill models. Nine countries participated in a pilot study to validate the instrumentation and data collection methods. Items were taken from the NA surveys as well as from assessments used in other European countries (about half of the text sources were from NA). Participating countries translated the texts and quality assurance and psychometric studies were undertaken to identify problems, such as poorly performing items. Statistical procedures were used to determine empirically whether some items performed differently in some countries, most often due to difficulties in translation and adaptation. In the IALS analyses, deviating items were permitted to have unique difficulty parameters—a strategy which minimised any bias in the overall literacy levels of the country that would have been due to shifts in relative difficulty of the items. Overall, 92 percent of the items satisfied the criterion of consistent relative difficulty across all countries and thus were assigned international difficulty parameter estimates.

From this review, it appears that validity was addressed in multiple ways and many measures were taken to enhance it. However, one weakness noted in the study was that the collegial approach depended ultimately on the ability of participants to adhere to agreed specifications and procedures. Countries did not always choose to follow recommendations to alter instruments after problems had been identified or lacked the technical expertise to comply with procedures. This necessarily weakens validity and comparability.

The issue of legitimacy was raised when France withdrew its results from the study before the first round of data were published. The withdrawal came despite French participation in the design, development and testing of the study, and was motivated by concerns about the comparability of the results about to be published. France felt the results tended to underestimate the true literacy skills of the population relative to other participating countries. Three specific objections were raised: 1)

items were biased in favour of "Anglo-Saxon" cultures at the expense of Latin cultures due to the origins of the survey in North America; 2) translation and adaptation of some items into French had increased their difficulty level; 3) French respondents were less motivated than respondents in other countries. Independent analyses of these issues, however, could not identify any impact on the French IALS results. Other studies indicate that differential motivation is also not a factor in explaining differences in national proficiency levels. Although the French claims could not be "proven" the IALS reports indicate that issues such as potential cultural bias were taken seriously and that it is important to guard against the possibility of linguistic, cultural or geographic biases in comparative studies of this type.

An interview with a representative from the French INSEE provides their perspective. In the view of the interviewee, results from the first IALS were misleading. There was no single factor explaining the deviating French results, but on the basis of the IVQ-data (the literacy test now being field in France) the INSEE has identified two important factors: problems of interpretation and the method of data collection. IALS respondents had to complete a booklet in writing, while IVQ permits oral responses. The results from these methods differ significantly, even with identical test questions. The IVQ also used improved sampling techniques, shortened the length of the questionnaire in order to increase motivation and changed coding to allow for more choices. The IVQ reportedly has a better response rate than IALS.

### **Reliability of assessment methods**

All 12 first-cycle countries used probability sampling for most of the stages of their sample design; ten used it in all stages. Two countries used non-probability sampling methods. For round two, enhanced data quality procedures were imposed, thus all countries implemented statistically sound designs. The exception is Portugal, which conducted a literacy survey as part of an EU-sponsored research project undertaken independently of IALS cycles but using similar methodology and identical test instruments (details are documented and available).

In our assessment, the IALS covered many different aspects of reliability, as noted. There are extensive technical and published reports by the IALS team and independent researchers. In addition, the EC commissioned a review of IALS after France withdrew from the study (see below), which was undertaken by the United Kingdom Office of National Statistics (ONS). A review of IALS methods, conducted by three independent international experts prior to the first publication, recommended the results be published but identified several ways in which survey methods could be improved. The second and third waves of the data collection were enhanced to improve methods and thereby enhance validity, reliability, and comparability of data. The final report (OECD, 2000) provides details and references to other studies and also addresses methodological issues raised by others.

### **The quality of data collection**

Several measures to ensure reliability were imposed. Survey administration guidelines specified that each country should work with a reputable data collection agency or firm, preferably one with its own professional, experienced interviewers. Rules were established concerning supervision, quality checks, etc. Precautions were taken against non-response bias and specified in Administration Guidelines. Countries were required to capture and process data files using procedures to ensure logical consistency and acceptable levels of error and to map national datasets into highly structured, standardised record layouts. Statistics Canada ran various checks on range,

consistency, etc. Test scorers received extensive training; re-score reliability was calculated, with re-scoring of one country's tests done by another participating country. In-depth analyses were conducted to assess data quality, and cases which present problems for international comparability were noted.

Feasibility is an issue with respect to implementation, in particular the willingness and technical capabilities or capacity of the participating countries to follow the guidelines and technical specifications required to enhance validity, reliability and comparability. As noted, the collegial nature of the undertaking meant that countries had some leeway in their level of compliance, and to the extent that they deviated from recommended procedures may contribute sources of error into the measurement effort. Although numerous steps were taken to correct possible biases or errors, e.g., corrections for non-response bias, some variability is unavoidable in this type of study.

### **3. Lessons learned**

Murray (2003) offered a number of suggestions for future measurement, as well as lessons learned.

1. Empirical evidence from IALS suggests that the attributes which underlie the the difficulty of adult reading tasks are stable and predictable over a broad range of languages and cultures. However, it is also important for assessments to have face validity—to be seen and accepted as culturally, linguistically, and geographically unbiased. The IALS was criticised for its Anglo-Saxon origins, and future assessments need to be more systematic in seeing to achieve more diverse item pools. (The ALLS moved towards this goal by involving researchers from 26 countries and stimuli in 16 languages).
2. Future assessments of literacy skills, the underlying theory regarding the determinants of difficulty of adult reading tasks should be tested in other languages, through linguistic research.
3. The results from IALS underscore the dynamic nature of literacy which develops throughout the life course. Further studies should attempt to employ longitudinal designs that can directly assess this development.
4. To understand relationships between skills and the economy, assessments will need to include additional skill domains. Research needs to continue to develop valid and reliable direct measures in skill domains thought to be of economic, educational and social import. (ALLS and DeSeCo are examples of this development effort).
5. Studies can better understand the demand for skill by including questions to determine how employees are called upon to use their skills (as does ALLS and the UK Skills Surveys). These surveys would also have to be conducted within a representative sample of firms to understand links to productivity, interfirm variability in the demand for and utilisation of skills, etc.
6. Improve quality assurance strategies to include explicit standards and guidelines for each key activity; pre-collection documentation and review of procedures for key activities; education processes to ensure that national study teams understand the intent and detail of

activities; and post-collection analysis and certification procedures for both national study teams and the international consortium.

7. Theoretical development of adult literacy should continue. The IALS assumption about the uni-dimensionality of scales is open to question, particularly at the lowest region of the scales.

Scott Murray emphasised three types of error that need to be considered in large scale surveys-- psychometric, sampling of the content domain, and implementation. In his view, the last posed the biggest challenge for international assessments—participating countries must agree to adhere to guidelines and standards.

Similarly, OECD (2000) identified several factors that will help reduce variability in multi-country surveys of this type: "presence of clear and realistic standards, consortia of skilled and experienced institutions, sufficient budgets to fulfil the complex statistical and operational demands imposed in such a study, well developed quality assurance procedures to minimise deviation from specification and to identify problems at a stage where they can be dealt with, and finally, and perhaps most importantly, a willingness on the part of participating countries to adhere to agreed standards and guidelines"

Finally, an international survey like IALS is expensive. The development of the initial test was largely borne by ETS and the governments of Canada and the US. European participants in the first wave agreed to match a Eurostat contribution of \$100,000. In reality international overheads of the first round of IALS collection totalled more than \$1 million, leaving the US and Canada to absorb the balance. In the second round, participants were required to cover a much larger fraction of the international overheads, on average \$60,000 per country. By the third round, the charge was \$75,000. The rising costs are reflected in the further refinement and extension of quality assurance procedures as lessons learned from each survey were used to improve subsequent waves.

## 4. References

### Interviews conducted/persons contacted

Meeting with Scott Murray, Director General, Institutions and Social Statistics, Statistics Canada 17 "a" R.H. Coats Building, Tunney's Pasture, Ottawa, Ontario, K1A 0T6, on 22 October 2003 (C. Stasz).

Meeting with INSEE, France (Fabrice Murat), on 17 December 2003 (F. Green and J.H. Haahr)

### Literature

Carey, S., Low, S. and Hansbro, J. (1997). *Adult literacy in Britain: A survey of adults aged 16-65 in Gready Britain*. The Stationary Office, London.

INSEE (2002). *Enquete Information et Vie Quotidienne*. Paris.

Jensen, T.P., Holm, A., Andresen, A., Hastrup, S., Heinskov, M.B., and Jabocsen, J. (2000). *Dasnskernes lease-regne-faerdigheder-udvalgte resultater*. Kobenhaven: Amternes og Kommunernes Forskningsinstitut.

Linnakyla, P., Malin, A., Blomquist, I. and Sulkunen, S. (2000). *Lukutaito Tyossa ja Arjessa: Aukuisten kansioanvaliein lukutaitotutkimus Suomessa*. Koulutuksen tutkimuslaitos, Jyvaskylan yliopisto, Tilastokeskus, Petusministerio Opetushallitus ja kiroittajat, Jyvaskyla.

Morgan, M., Hickey, B., Kellaghan, T. (1997). *Report to the Minister for Education on the International Adult Literacy Survey: Results for Ireland*. St. Patrick's College, Dublin: Educational Research Centre.

Murray, T.S. (March 2003). *The Assessment of Literacy, Numeracy and Language at the International Level: A Review*. Paper prepared for the National Academy of Sciences (NAS).

OECD (2000). *Education at a Glance*. Paris.

OECD and Statistics Canada (2000). *Literacy in the Information Age: Final Report of the International Adult Literacy Survey*.

OECD (1998). *Human capital investment: An international comparison*.

OECD and HDRC (1997). *Literacy Skills for the Knowledge Society: Further Results from the International Adult Literacy Survey*.

OECD and Statistics Canada (1995). *Results of the First International Adult Literacy Survey*.

## **IVQ – Information et Vie Quotidienne**

### **1. Description**

#### **Goals and ambitions**

The IVQ is a specific French version of the IALS and consists of tests and questionnaires concerning literacy and numeracy. It was developed in response to France's dissatisfaction with some methodological aspects of the IALS-survey.

The French cited three objections: 1) because of the origin of the survey in North America, the items had an "Anglo-Saxon" cultural bias, 2) the translation and adaptation of items into French had increased their difficulty levels, and 3) the French respondents had been less motivated than respondents in other countries. OECD, 2000).

The project description lists three overall objectives:

- To evaluate the educational system
- To refine the analysis of the labour market (direct assessments provide a better understanding of the skills of the labour market than proxy measures such as education qualifications)
- To guide remedial policies for improving literacy.

#### **Competencies assessed**

The survey assesses literacy and numeracy, with primary focus on literacy. Moreover, even though there is a national reading curriculum defined in the French school system, it was decided not to take this as an outset: The way reading is taught in schools corresponds poorly to the needs of everyday life, and a household survey poses specific challenges: People are interviewed at home, not in a class room, and often many years after they have left school. Hence, the decision was that exercises must look as little "school-like" and academic as possible, and exercises were built on everyday supports (newspaper, CD, statistics, etc.). At the same time, an attempt was made to take into account the latest scientific insights in psychology, and theoretical testing frameworks for exercises were defined by researchers.

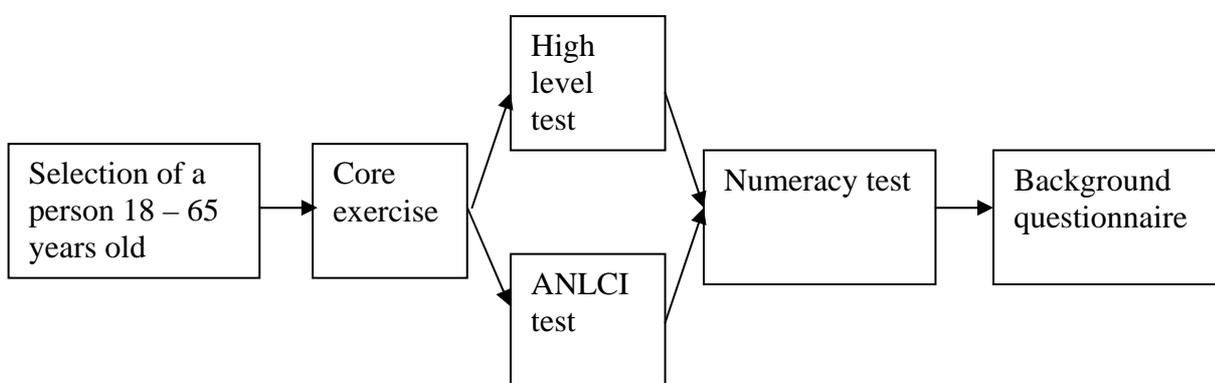
#### **Methods**

The IVQ was a performance test carried out through a household survey. Background information was collected through a traditional questionnaire. The literacy test material was developed with the involvement of a range of experts representing different institutions.

Literacy tests were divided into several steps, cf. figure 1. The first step was an orientation module, making use of a TV guide magazine as test material, something everybody has encountered at some point (if the respondent could not read or speak French, the interviewer would still attempt to obtain answers to the background questionnaire with the assistance of a family interpreter). If the results of the respondent were not good enough (less than 11 points of 15 on reading or less than 11 points of 19 for an exercise on understanding), the respondent was lead on to a specific questionnaire with the easiest exercises: The "ANLCI Test". If the respondent had good results, he/she was proposed one of the two sets of the "High level test" (one of these two sets contains six items from the IALS).

Next step for the respondent was the numeracy tasks, where the respondent was presented with short sentences read aloud by the interviewer to avoid reading difficulties. Finally, in the background questionnaire, the respondent was asked to provide some information on his/her family, level of education, occupational situation, reading practices and - if the respondent did not do well in the core exercise - any difficulties encountered in daily life because of reading difficulties. The interviewer was allowed to stop all exercises if it seemed too hard or too easy for the respondent in question.

**Figure 1: Overall test design, Information et Vie Quotidienne**



As for the core exercise, the test item consists of an extract from a TV guide, and the test begins with very general questions: "What is this? What is its use? How have you found out what it is". Next, the respondent has to read some words (the title of a movie or the names of guests in a show). Finally, the respondent reads the review of the movie (about fifteen lines) and answers tens questions about it.

After this exercise, the respondent has to read two numbers (/10 008 and 45 000 016) and to solve three easy problems: If he/she answers correctly to none of these three problems, he/she will later begin the series of numeracy tasks with question no. 1. If the respondent succeeds in solving 2 or 3 of the easy problems, he/she will go directly to question no. 18 in the numeracy test (numeracy exercises are ranked by difficulty).

As for the ANLCI test (named after the institution having developed it), it is proposed for people who did not do well in the core test. First, in order to make the respondent feel more confident, he/she listens to a radio text and answers to some questions. Then a shopping list is dictated to the respondent to write and the module ends by an exercise rather similar to the core exercise (reading and understanding written words on a CD).

There are two versions of the high level test. One includes six exercises from the IALS-test; the other is a new pool of literacy tasks, which also includes the same oral comprehension exercise as the ANLCI-module. A third of the respondents, randomly selected, are to complete the IALS-test.

The background survey includes questions on the job level and education of parents, respondents' level of education, employment, reading habits, and language spoken in different contexts

Two pilot tests were implemented in December 2000 and April 2002, in order to assess whether the literacy survey could actually be implemented. Following the experiences with IALS, this was not sure.

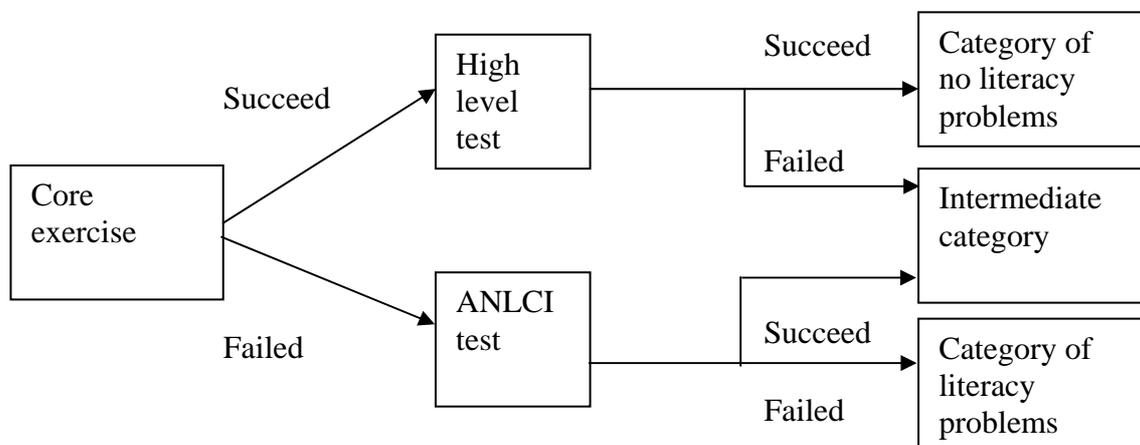
Full-scale data collection was carried out in November 2000 among a representative sample of adults aged 16-65 in 10 French regions. 4000 dwellings were selected in the 'master sample' from the last French Census. The non-response rate was 48 per cent. Half of the non-response was refusals or absences; the other half was demolished or vacant dwellings. Benchmarking on variables such as educational level, gender, age, etc., corrects the non-response bias. 2086 persons agreed to take part in the survey. Of these, 1997 actually completed the exercises, the remainder answered only the background questionnaire. A more comprehensive survey (target 7000 responses) will be carried out in 2004.

**Results**

The IVQ has produced data on literacy levels in the French adult population, and since it includes six IALS test items, it has also contributed to the debate about the validity of the IALS test results.

As for results concerning literacy in the French adult population, the first conclusions were published in April 2004 (Murat 2004). The report identified the significance of various background variables for literacy (gender, education, age), but in particular it focussed on the literacy difficulties faced by non-native French speakers: Whereas 12 per cent of the total adult population are said to face literacy problems, this is the case for only 7 per cent of the native French speakers, born in France, but for 30 per cent of the population born outside France with another mother tongue than French, but having learned to read and write in French. For persons born outside France with another mother tongue than French, and having learnt reading/writing in a non-French language, almost 2/3 (64 per cent) of the respondents face problems in reading and writing in French.

**Figure 2: Categorisation framework, Information et Vie Quotidienne**



These results are based on a categorisation of respondents test scores into three groups, cf. Figure 2: A category of persons with no literacy difficulties, an intermediate category and a category of persons with literacy difficulties. As for the last category, it consists of respondents who obtained less than 11 points of 15 on reading or less than 11 points of 19 for an exercise on understanding in the introductory core exercise, and who subsequently answered less than 8 of 10 questions correctly in at least one of the three sub-domains tested in the ANLCI-test (reading, writing, comprehension). The high level category, as opposed, consists of persons who passed the core exercise and who solved at least 1/3 of the tasks of the high level test. The intermediate group consists of two groups of persons: Those who did not succeed in the core test, but who answered more than 80 per cent correctly in the ANLCI tests, and those who succeeded in the core exercise but failed to answer correctly in at least 1/3 of the tasks of the high-level test.

As for the results of the IVQ in comparison to the IALS-test scores for France, the IVQ data seems to suggest that at least the French IALS test scores were problematic (Murat 2003b). The IALS data showed that the proportion of low-skilled readers among the adult population in France was 40 per cent, leading the French government to withdraw from the study, whereas the IVQ data suggest that this proportion is 12 per cent. This may be due to differences in test items.

However, when comparing the test scores of the 1994 IALS survey to the 2000 IVQ survey on identical test items, significant differences appear: 13 test items are identical in the IALS and the IVQ survey, but the proportion of correct answers is significantly higher in the IVQ than the IALS survey for 11 of these 13 test items (Murat 2003b: 10).

Several explanations are possible. However, differences in *test design* and *data collection procedures* from IALS to IVQ seem likely to be significant:

- In the IVQ, each exercise was presented one by one to the respondent, and not as in the IALS in the form of a rather large booklet. This may have ensured that the respondents' attention was more focussed in the case of the IVQ. The fact that omitted answers are very rare in the IVQ would seem to suggest this.
- The relation between the interviewer and the respondent was different in the IVQ. The use of CAPI (computer assisted personal interviewing) and the segmentation of tests allowed a more convivial interview situation. Less artificial and less "scholarly", the situation seems likely to have facilitated the respondent's ability to concentrate.
- The use of CAPI and the segmentation of tests may also have reduced stress for the respondent, facilitating better answers. In the IALS, the entire test was conducted as a written questionnaire, often lasting more than two hours. During this period, the situation of the interviewer was uncomfortable, as he/she had nothing to do. This may have increased stress for the respondent and the feeling that test time was limited, although the interviewer clearly specified that it was not.
- The shorter duration of the IVQ test may have allowed respondents to remain more concentrated. The IALS was, as mentioned, a quite long survey, the test booklet (which existed in 7 versions) containing 15 different exercises, and total test duration often exceeding two hours. As opposed, the IVQ contained just 6 of the 15 IALS exercises, which presumably allowed respondents to remain more concentrated, and the average testing time of the total IVQ survey is significantly lower than for the IALS.

Other differences between the IALS and the IVQ may also be emphasised when explaining the score differences: Items scoring in IALS was rather crude, allowing only three scores, correct, incorrect and omitted, and there may have been sampling problems with the IALS survey (sampling in IALS used phone numbers, randomly chosen. To avoid the problem of hidden lines, the interviewer used the “random route method”. Replacements were allowed when and appointment could not be arranged, and this happened for 45 per cent of the households).

## **2. Evaluation**

### **Utility/usefulness for policy development or other purposes**

The first results of the IVQ survey have only recently been published (Murat 2004), and there have been few reactions from behalf of policy makers and the political system. Hence, the survey has not yet been put to practical political uses.

However, based on an assessment of the IVQ test design and the preliminary results, the initiative would appear to be of a certain potential usefulness. At present, it seems useful for one particular purpose:

- providing information on the overall magnitude of literacy problems in different segments of the population, thereby potentially focussing energy and resources towards the most needy groups.

This means that the IVQ at present has the potential of achieving at least one of its overall objectives, namely to guide remedial policies for improving literacy.

As for the IVQ’s usefulness in relation to its’ two other explicit objectives (to evaluate the educational system; to refine the analysis of the labour market), much will depend on the research resources which will be devoted to analysing the IVQ survey results and their policy implications, and whether additional surveys will be implemented in the future, allowing inter-temporal comparisons.

One reservation in this connection pertains to the contents of the background questionnaire: The IVQ’s usefulness could be improved if a wider range of background variables were to be included. If the IVQ is to inform policy making more directly in relation to the educational system, it would also seem important to obtain information about for instance fields such as individual beliefs, motivations, preferences and experiences in relation to education, training and measures to improve literacy. As it stands now, the IVQ can provide some overall information on the character and significance of “the literacy problem” in France, and the significance of this problem in specific groups (age, gender, mother tongue etc). But this information would seem insufficient for evaluating the education system, at least as long as there is no longitudinal data.

### **Quality and relevance of skills definitions**

The IVQ focuses on literacy and numeracy. The initiative appears in this connection to be built on experiences from previous assessment initiatives, not least IALS. Against this background, definitions seem rather clear and relevant.

### **Validity of assessment methods**

With its outset in the experiences from the IALS in France, it seems that considerable efforts have been put into designing methods and test items that are valid. One question has to do with the placement of respondents into three groups of proficiency, high, intermediate and low. Whereas the test items would seem to be valid for providing information on proficiency, the conceptual validity of the categorisations and their cut-off points is open to question. The thresholds defined for accepting a test score as “successful” or “failed” respectively are just one among many other possibilities, and they ultimately rely on assessments of the involved researchers and experts.

### **Reliability of assessment methods and the quality of data collection**

Considerable efforts seem to have been devoted to the development of reliable data collection procedures. Not least, much attention appears to have been paid to the fact that conditions for carrying out household surveys involving direct performance testing among adults differ considerably from other types of surveys or, for instance, school based testing. A number of relevant lessons can be learnt from this:

## **3. Lessons learned**

The IVQ points to the significance of developing data collection and testing procedures which are specifically targeted at adults who have in many instance left the formal education system many years ago.

### **The importance of data collection and test design**

Both on the basis of the IVQ itself, and on the basis of the differences in the test results between the IVQ and the IALS, the following experiences can be highlighted:

- The precise context of the interview/test situation is important. There seem to be significant advantages, both for response rates and for data reliability, in creating an interview situation which is convivial, minimises stress for the respondent and maximises the respondent’s ability to concentrate on the task in question.
- Presenting test items one after the other, and posing questions in oral form are two seemingly useful approaches in this respect. Computer Assisted Personal Interviewing made this possible, but it may also be possible using paper-and-pencil data collection mechanisms. As opposed, large booklets constructing a “school-like” testing situation, and requiring up to two hours of testing would appear to increase the risk that respondents are stressed and loose concentration. The IVQ staff itself stresses in this connection that motivation should be measured as part of the survey, in order to be able to analyse its impact on test scores subsequently.
- The IVQ staff emphasises the significance of adequate training of interviewers. Testing adults that have more often than not left the formal education system decades ago is a sensitive situation, which requires a specific behaviour and code of conduct of the interviewer. This presupposes considerable training and instruction.

### **International comparisons as a political tool**

There is also lesson of a different kind to be learned from the IVQ. Thus, in terms of the potential political impact of IVQ, the absence of the international comparison seems to have restricted its

importance. At least at the time of writing, the publication of the first results appear to have aroused little interest, and it seems clear that the absence of international benchmarking plays a role here.

Comparing the IVQ and IALS hence underlines the opportunities and indeed the risks involved in international comparisons in the field of adult skills. International comparisons is a powerful tool for obtaining the attention of the media, policy makers etc., and in many instances also for initiating political action. However, the power of this tool is also its risk: If political attention is focussed on the wrong set of skills, and possibly on a questionable basis, considerable damage may be done.

The IVQ also provides an example of the type of political reaction than can result in a particular state if test results of international comparisons are not deemed valid and reliable: France in this connection withdrew from international efforts and launched its own national initiative. This highlights the requirement to ensure high quality data, both in terms of validity and reliability.

#### **4. References**

INSEE 2002. "Information et Vie Quotidienne", Questionnaire and test items, Paris.

Murat, Fabrice, 2003a. "Principes de L'Enquete Information et Vie Quotidienne", 12 Juin, Paris: INSEE.

Murat, Fabrice, 2003b. "Assessment of Adult Literacy. After IALS a French Experiment", Paris: INSEE.

Murat, Fabrice, 2003c. "Application de la Méthodologie IALS aux Données de L'Enquête IVQ, 12 Juin, Paris: INSEE.

Murat, Fabrice, 2004. "Les difficultés des adultes face à l'écrit", INSEE No. 959, Avril.

# DIALANG

## 1. Description

### Goals and ambitions

DIALANG is an assessment system intended for language learners who want to obtain diagnostic information about their language proficiency.

DIALANG is aimed at adults who want to know their level of language proficiency and who want to get feedback on the strengths and weaknesses of their proficiency. The system also provides the learners with advice about how to improve their language skills and, furthermore, it attempts to raise their awareness of language learning and proficiency. The system does not issue certificates.

The primary users of the system are or will be individual learners who have learned languages through informal settings, and those who study languages independently or on formal language courses. However, language teachers have also found many of the features of the system useful for a variety of purposes. Individual enterprises are also taking an interest in the system, using it as a tool to identify foreign language training needs.

The DIALANG project is carried out with the financial support of the European Commission, Directorate-General for Education and Culture (SOCRATES Programme, LINGUA Action D). The system has been developed since 1997. The beta version of the system came out in 2003 and Version 1 was launched in March 2004.

DIALANG's Assessment Framework and the descriptive scales used for reporting the results to the users are directly based on the Council of Europe's Common European Framework (CEF).<sup>22</sup> A recent European Union Council Resolution (November 2001) recommended the use of this Council of Europe instrument in setting up systems of validation of language competencies, thus in effect making the framework the European standard. The self-assessment statements used in DIALANG are also mostly taken from the CEF and adapted whenever necessary to the specific needs of the system.

DIALANG is presently delivered via the Internet free of charge, although a special software application must be downloaded and installed for users to be able to use the system. However, the key members of the Consortium behind the project (Lancaster University UK; Freie Universität Berlin, D; Centre for Applied Language Studies, Jyväskylä University, FIN; and CITO groep, NL) are currently in the process of forming a European Economic Interest Group which will own and manage the system for commercial purposes, while keeping the public system available for use free of charge. Negotiations are being carried out with large international providers of web-based testing platforms on delivering the system in a way that corresponds with the needs of institutions. The chosen company would act as a host for the system and would integrate the DIALANG database into its system, making it possible to carry out the assessment through a normal web-browser.

---

<sup>22</sup> [http://www.coe.int/T/E/Cultural\\_Co-operation/education/Languages/Language\\_Policy/default.asp#TopOfPage](http://www.coe.int/T/E/Cultural_Co-operation/education/Languages/Language_Policy/default.asp#TopOfPage)

### **Competencies assessed**

The system consists of self-assessment, language tests and feedback, which are all available in fourteen European languages: Danish, Dutch, English, Finnish, French, German, Greek, Icelandic, Irish, Italian, Norwegian, Portuguese, Spanish, and Swedish.

Using the system, the user can obtain assessments of five different foreign language skills: reading, listening, writing, vocabulary, and grammar. The system's assessment scores can be compared with the user's initial self-assessment.

### **Methods**

The DIALANG assessment procedure has the following steps:

1. Choice of administration language (14 possible)
2. Choice of test language and skill (14 possible languages; 5 skills: reading, listening, writing, vocabulary, structures)
3. Vocabulary Size Placement Test
4. Self-assessment (only in reading, listening, and writing)
5. System pre-estimates learner's ability
6. Test of appropriate difficulty is administered (three versions available)
7. Feedback
8. Begin a new test or exit the system

On entering the system, the learners first choose the language in which they wish to receive instructions and feedback. Then they choose the language and skill that they want to be tested on. The users are then presented with an optional placement test which also estimates the size of their vocabulary. Following the results of the placement test, where learners have to identify existing and non-existing words for the language in question, the learners are also presented with a set of self-assessment statements, if they choose a reading, listening or writing test. These self-assessment statements cover the skill in question, and the learner has to decide whether or not she or he can do the activity described in each statement. After taking the placement test and the self-assessment statements the learners are led onto either an easy, intermediate or hard version of the test, the selection of the test bank having been based on the results of the two pre-test services.

After the test, as part of the feedback, the learners can discover whether their self-assessed level of proficiency differs from the level of proficiency assigned to them by the system on the basis of their test performance. Users are also offered an opportunity to explore potential reasons for a mismatch between self-assessment and the test results in the Explanatory Feedback section.

Self-assessment (SA) statements are used for two reasons in the DIALANG system. Firstly, self-assessment is considered an important activity in itself. It is believed to encourage autonomous learning, to give learners greater control over their learning and to enhance learner awareness of their learning process. The second purpose of self-assessment in DIALANG is more 'technical': the system uses the Vocabulary Size Placement Test and self-assessment results to pre-estimate the learners' ability and then directs them to the test whose difficulty level best matches their ability.

Feedback on the level of proficiency assigned by the system is given in two ways:

- the more concise version accompanies the test score,
- the more extensive version is part of Advisory Feedback.

When learners get concise feedback on their test score, they are given a result on the CEF scale, A1 (near beginner) to C2 (near native-speaker proficiency), and the meaning of this score is described using these reporting scales.

The Advisory Feedback section of the assessment system uses scales which contain more extensive descriptions of proficiency in reading, writing and listening. The section provides the users with more detailed accounts of what learners can typically do with the language at each of the skill levels. The learners can also compare the description for a particular level with the descriptions for adjacent levels. These more detailed scales are also based on the CEF scales, but the descriptors were elaborated further with the help of other sections of the CEF and also other sources.

## Results

Evidence as to the results of the system is, as of now, scant. This however is understandable taking into account that the system has only recently been made widely available. It is anticipated that in the years to come it will be a source of useful research. Nevertheless, the system is finding uses in a number of different context, cf. below. It has not been attempted to use it as a tool for collecting data on the language proficiency of large scales samples, neither has the system until now had the ambition to contribute to the evaluation of educational systems or programmes, or to provide policy input in these connections.

## 2. Evaluation

### Utility/usefulness for policy development or other purposes

The DIALANG system and has only quite recently been made publicly available. Information on the practical usefulness in different contexts is therefore rather scattered. However, it seems clear that the system is finding uses in a number of different settings. Typically, the system is now used for placement purposes in "low stakes" situations, for instance as a placement tool for determining at which level a learner should enter a foreign language evening course or courses targeted at non-language majors at universities. Since February 2001, there have been over 90,000 visits to the DIALANG website, and from April 2003 to 9 May 2004, the test has been accessed more than 75,000 times, cf. table 1 below.

**Table 1: DIALANG tests commenced since April 2003, distribution on test languages**

Test Language	Number of tests commenced
DA	2780
DE	8303
EL	374
EN	37505
ES	3191
FI	3421
FR	8970
GA	224

Test Language	Number of tests commenced
IS	158
IT	1968
NL	3603
NO	1537
PT	579
SV	2872
<b>TOTAL:</b>	<b>75485</b>

It is also used as a guide to certification or examination levels. Using DIALANG, learners can obtain an assessment of whether it is realistic or not to submit him- or herself to a specific examination. DIALANG is also used in connection with self study programmes, where learners can for instance test their progress or the relation between their self-assessed proficiency and the test results deliver by the system. Private enterprises have also started to use the system to assess the level of language proficiency of employers and to identify training needs.

DIALANG is deliberately being described as a diagnosis-tool rather than a test tool. One of the reasons for this is the existence of a number of commercial foreign language test tools, and the risk discussed at the onset of the project that the DIALANG could undermine the market for language testing with public financial support. However, an initial decision was made to stay clear of the entire certification field. Since then, development was undertaken to address the needs of non-specialist language learners. Nevertheless, in its core, the system *is* a test, but one which is presently only relevant for "low stakes" situations. The framework and set-up of the test could be applied to higher stake situations, however this would be a costly exercise taking into account the need to guarantee continuous refreshment of item banks with calibrated and fully piloted items.

Using DIALANG as an element in a household survey for general policy purposes would be to use it in a low stakes situation, since the person's life would not be greatly affected by the test result.

In principle, the system therefore provides a good starting point for foreign language proficiency assessment in connection with a household survey. It is based on today's standard point of reference for defining language proficiency. A large number of items have been developed in addition to the 300 complete and tested items presently available. These items are in a base form and need to be reviewed and pilot tested. However, items that have already been tested could be used as an "anchor" for pilot testing new items, this would facilitate the process. Still, however, a number of questions must be resolved:

- The question of time requirements. Among language proficiency testing experts, there is agreement that it is necessary to ask at least 30 items for each type of language proficiency skill if assessment results are to be reliable and valid. It is estimated that this takes at least 20 minutes at present, often more. If several language skills are to be assessed, the foreign language module could in itself require between one and two hour of testing. This would affect response rates dramatically in a negative direction.
- The question of the delivery platform. At present, DIALANG presupposes that a specific software application is downloaded from the internet and installed onto the computer where the test is to be taken. The test may become available through normal internet browser within

the next year or two. Stand-alone versions of the test, requiring no internet connection, could also in principle be developed. There does not seem to be principal problems related to a transfer of the test to a paper-and-pencil version, but it will of course require some development work.

### **Quality and relevance of skills definitions**

DIALANG has chosen to use the Council of Europe Common European Framework for Languages as its foundation. Whereas the definitions provided in this framework are not beyond discussion, they have been considered useful and sufficiently elaborated by the DIALANG project team. Previous experience of the project members shows that even if quite considerable time and resources are devoted to developing language skills definitions, the results are no better than those delivered in the framework of the Council of Europe. The skills definitions used in DIALANG must be considered of relatively high quality and relevance.

### **Validity of assessment methods**

It is relevant to consider both the validity of the self-assessment statements and the test items of the various language skills tests.

#### *Self-assessment Statements*

Most of the self-assessment statements used in DIALANG were taken from the English version of the Common European Framework (Draft 2, 1996). In this respect, DIALANG is a direct application of the Framework for assessment purposes.

The DIALANG Working Group on Self-Assessment<sup>1</sup> reviewed all CEF statements in 1998 and chose those which appeared to be the most concrete, clear and simple; North's (1996/2000) empirical results on the statements were also consulted. More than a hundred statements were selected for reading, listening and writing. In addition, statements about speaking were chosen but as speaking is not part of the present DIALANG system, they were not included in the validation study described below and are thus not presented in this appendix. The wording of the statements was changed from 'Can do' to 'I can' because they were to be used for self-assessment rather than teacher assessment purposes. Some of the statements were modified to simplify them further to suit the intended users; a few new statements were also developed where there was not enough material in the CEF to draw on. All statements were audited by Dr Brian North, the originator of the statements in the CEF, and by a group of four language testing and teaching experts before the final wording of the statements was agreed.

Because DIALANG is a multilingual system, the self-assessment statements were then translated from English into the other thirteen languages. The translation followed an agreed procedure. Guidelines for translation and negotiation were agreed; comprehensibility to learners was a prime quality criterion. Initially, two to three experts per language translated the statements into their language independently and then met to discuss differences and to agree a consensus wording. The translations were forwarded to the Self-Assessment Group whose members had the linguistic proficiency to additionally cross-check the quality of the translations in nine languages. The translators were contacted and any questions related to wording were discussed and modifications agreed.

Subsequently, self-assessment statements have been calibrated, using the data was analysed with the OPLM software (Verhelst et al. 1985; Verhelst and Glass 1995). (Calibration is a procedure in which the level of difficulty of items, statements, etc. is determined statistically and a scale is constructed of them).

#### *Test items*

A large number of test items for each type of language skills have been developed. In total, approximately 30.000 items were developed (about 800 – 3.000 per language). Only 300 of these are now being used per language, i.e. 4.200 in total.

For the larger languages, the difficulty level of the various items have been tested through pilot studies on a test population, where approx. 100 - 300 responses were obtained for each item, and through subsequent statistical analysis of responses, following Item Response Theory and applying the OPLM software package. For the smaller languages, for instance Icelandic, a 2<sup>nd</sup> best solution has been applied, where a panel of experts have estimated the level of difficulty for each item.

A separate question has been to determine the cut-off points, i.e. where to place the cut-offs for each of the levels on the six-point CEF scale, in order to determine how to convert the test score onto the scale used for reporting the test result to the user. This question has also been determined using data from pilot tests and panel expert judgements. On the basis of these two sources, a specific software programme generates cut off points.

It seems clear that considerable effort has been put into securing the validity of self-assessment and test items. Whereas not all questions of validity have been resolved, the validity of assessment methods should be considered high in most respects.

#### **Reliability of assessment methods**

DIALANG is presently a self-assessment system for diagnosing language proficiency. It is therefore in part, up to each individual learner to provide reliable data to the system if credible test scores are to result. At present, the system is hence mostly relevant for "low stake" situations, e.g. for placement purposes in connection with language learning courses. If it were to be applied in "high stake" situations, for instance as a test in connection with entry exams or in connection with job interviews, there would need to be stronger external control of the assessment situation, and regularly refreshed item banks if data were to be reliable.

#### **Quality of data collection**

This question is not relevant for DIALANG in its present form. It is worth noting, however, that the completion of the test, when testing all the language skills that can be tested in the system presently, normally takes between 30 and 45 minutes, depending on the skill tested, the time taken to think through an answer and to some degree whether the learner has been placed at the most relevant level of test difficulty.

### **3. Lessons learned**

The DIALANG project is directly relevant for a European adult skills assessment initiative. Foreign language proficiency must be considered among those skills that can be directly and efficiently

affected by reforms of educational systems and political decisions and prioritisations within educational policy.

### **Potential contributions from DIALANG**

The DIALANG system provides a relevant, high quality starting point for including a foreign language proficiency element in a future European adult skills assessment initiative. This is so in several respects:

- It is based on the standard point of reference as regards definitions of language skills proficiency, the Council of Europe Common European Framework of Reference for Languages.
- Considerable effort has been put into developing a system that delivers equally valid and comparable test results for different languages.
- A large number of items have been developed in addition to the 300 complete and tested items presently available for each language. For English, an extra set of 300 items have been thoroughly reviewed but not pilot tested. These items could be refined and tested using items that have already been tested as an "anchor", thereby facilitating the process.
- The system can most likely be made compatible with a computer (lap top) based data collection procedure. At the same time, the format of the tests allows it to be transposed to a paper-and-pencil version, should this turn out to be necessary.

### **Requirements to development and adaptation**

There are also, however, a number of questions that must be resolved if the DIALANG system is to be useful in the context of a European adult skills assessment initiative.

#### *Choice of languages*

In one respect, DIALANG is too ambitious: Only a few of the foreign languages covered by DIALANG seem likely to be relevant in the context of a European skills assessment initiative. One likely prioritisation would be for respondents' skills in the most important foreign language to be assessed. For most countries in the EU, this will be English. For the UK and Ireland, it would seem likely to be French. However, since the system allows the respondent in any case to select from the range of languages on offer, a different option would be to let respondents themselves choose in which foreign language (among the 12 currently covered by DIALANG) skills are to be assessed.

#### *Choice of skills*

The question of time requirements. It is necessary to ask at least 30 items for each type of language proficiency skill if assessment results are to be valid. It is estimated that this takes at least 20 minutes at present. It may, however, be possible to shorten the test a bit if the items used are very reliable.

If several language skills are to be assessed, the foreign language module could in itself require between one and two hours of testing. This would affect response rates dramatically in a negative direction.

One possible solution is to focus on just one foreign language skill. However, this is a questionable option, and is not recommended by key experts, as the correlation between different language skills is typically not high enough (good reading skills does not necessarily mean good speaking skills

etc). Even if this is so, a decision can be made to prioritise just one foreign language skill, for instance *vocabulary*, and to test just this skill, but this would open up the skills assessment initiative to significant criticism.

Alternatively, it is recognised that *listening*, *reading* and *grammar* are three key foreign language skills, and that there is generally a high correlation between the combined proficiency of these three skills and other foreign language skills such as speaking, writing and vocabulary. Hence a second option is to test for these three language skills. The final option is to test for all five language skills where items have presently been developed in DIALANG.

#### *Several rounds of data collection?*

If three to five foreign language skills are to be assessed, however, testing time for language testing alone would exceed 1 hour. If very low response rates are to be avoided, this may require that several rounds of data collection are carried out with different samples, where one round would focus almost exclusively on language proficiencies. This would not only allow testing to take place for several language skills, it would also in principle make it possible to test proficiency levels for more than one foreign language, for instance English and a second foreign language, typically German, Spanish or French, in the non-Anglo-Saxon EU Member States.

#### *Delivery platform*

A final question concerns the delivery platform. At present, DIALANG presupposes that a specific software application is downloaded from the internet and installed onto the computer where the test is to be taken. Using elements of DIALANG in connection with a household survey on adult skills requires that a stand-alone version of the test, requiring no internet connection, must be developed, or alternatively that the test is transferred to a paper-and-pencil version. This will take time and resources, but there do not seem to be irresolvable problems related to these tasks.

## **4. References**

### **Interviews and persons contacted**

Alderson, Charles, Professor, Lancaster University, United Kingdom.

Huhta, Ari, Researcher, Jyväskylä University, Finland.

McLaughlin, Gillian, DIALANG Project Manager, Freie Universität Berlin, Germany.

Oscarson, Mats, Göteborg University Sweden.

Minna Scheining, Turku Polytechnic University, Finland.

### **Literature**

Alderson, J.C. (forthcoming). Diagnosing foreign language proficiency - the interface between learning and assessment.

Alderson, J.C. 2000. Technology in testing: the present and the future. *System* 28, 4, p. 593-603.

Council of Europe. 2002. The Common European Framework of Reference for Languages. Appendix C (The DIALANG Scales).

Floropoulou, Chryssoula. 2002. Foreign language learners' attitudes to self-assessment and DIALANG: A comparison between Greek and Chinese Learners of English. MA thesis. Department of Linguistics and Modern English Language. Lancaster University.

Huhta, A. and N. Figueras. 2004. The DIALANG Project: Using the CEF to promote language learning through diagnostic testing. In K. Morrow (ed.) Insights from the Common European Framework. OUP.

Huhta, A., S. Luoma, M. Oscarson, K. Sajavaara, S. Takala, and A. Teasdale. 2002. In J.C.Alderson (ed.) Common European Framework of Reference for Languages: Learning, Teaching, Assessment. Case Studies. Council of Europe. 130-145.

Kaftandjieva, F. and S. Takala. 2002. Council of Europe scales of language proficiency: A validation study. In J.C.Alderson (ed.) Common European Framework of Reference for Languages: Learning, Teaching, Assessment. Case Studies. Council of Europe. 106-129.

Yang, Rui. 2003. Investigating how test-takers use the DIALANG feedback. MA thesis. Department of Linguistics and Modern English Language. Lancaster University.

Also C. Alderson 2000. Assessing Reading. CUP. contains extracts from Dialang test specifications.

## **LAMP: Literacy Assessment and Monitoring Program**

### **1. Project description**

LAMP (Literacy Assessment and Monitoring Program) is a 'direct assessment of the literacy level of the adult population of different nations', which will measure 'a spectrum of literacy levels, from very basic reading and writing, to the higher level skills needed to participate fully in a learning society'. In a sense, LAMP is an extension of IALS on the lower levels of literacy.

The UNESCO Institute for Statistics in Montreal leads a group to develop and implement the survey. According to their project leaflet, the World Bank has shown particular interest in involvement. Others include international agencies such as UNICEF, UNDP, ILO and experts in literacy assessment such as Statistics Canada, the Educational Testing Service (USA), NEIPA (India) and HSRC (South Africa).<sup>(8)</sup> Fund-raising is continuing. Two donors, the UK Permanent delegation to UNESCO via DFID and UNESCO EFA Extra-budgetary, are already providing substantial financial support. However, more is needed to cover all the methodological development and to provide in-country support. <sup>(6)</sup>

The LAMP project organisation currently consists of a steering committee, Quality Assurance, a users network and Management Team. The work is carried out by six groups: Definition & contextual framework, Development of basic levels, Survey & instrument design, Survey implementation, Analysis & dissemination and Capacity building & methodology package. <sup>(2)</sup> Scott Murray (Director General, Institution and Social Statistics Canada) is involved as an external consultant to align LAMP with IALS.

The actual assessments will be carried out by national Project Teams in each participating country. An organisational framework is proposed by the LAMP team <sup>(12)</sup>, describing tasks and requirements of a National Project Leader. They also propose the Minister responsible for education in each country would take responsibility for the project and involves other relevant ministries, such as Labour/Employment, Health, and Social services. The minister would appoint the National Project Leader.

According to Levesley, Director of the UNESCO Institute for Statistics, the project team has currently agreed on the broad methodological principles, while they are already testing in some areas. The survey manual is worked out and some samples are drawn. In the preceding phase, they gathered data and experience from developing countries, developed an estimation method, design tests and the survey instruments and selected the samples. <sup>(2)</sup> Options for obtaining hardware and for developing the appropriate software tools are evaluated. The questionnaires of several literacy related surveys were reviewed and a database of questions has been assembled for reference in designing the survey questionnaires. <sup>(6)</sup>

In the beginning of 2003, work started on defining how the lower levels of the International Adult Literacy Survey (IALS) literacy scale (i.e. level 1 of IALS) could be extended substantially to allow for greater differentiation of basic skills, by assessing some of the component skills needed to reach higher levels of literacy. A technical meeting was held in Montreal on April 16-17 to discuss this. There was consensus that the following five components should be used to differentiate lower levels of literacy: oral comprehension, letter/grapheme recognition, word recognition, sentence reading

and passage reading. A locator test is built to direct respondents to a higher (based on the IALS methodology) or lower (components) level assessment module. (6)

Currently the pilot phase is taking place in a small number of developing countries. Morocco and Mongolia have signed up, while Jamaica and Kenya have shown interest and are negotiating. (9) The proposed timetable states pilot data collection and review will take place in August and September 2004, to start main data collection in these countries in October. Analysis and dissemination is planned for January up to July 2005. (12).

### **Goals and ambitions**

According to Scott Murray the primary goal is to provide national policy makers with reliable data on: the level and distribution of literacy skills; rates of adult education; social distributions of skills; the factors that influence the observed skill distributions; and the social and economic consequences of skill. He noted that repeated monitoring would allow for monitoring of trends while analysis of 16-24 year olds will provide an approximation of the current educational output. At the international level, LAMP will be useful in allocating funds based on need and allow for the monitoring of progress towards international goals including the Millennium Development Goals (MDGs). (1) This main objective can be split into three activities:

1. Develop a methodology for assessing literacy in developing countries, measuring a spectrum of literacy levels from very basic reading and writing to high-level skills needed to participate fully in a learning society;
2. Collect literacy data to inform the participating countries' policy-making and literacy programme design, and to help international monitoring and policy-making;
3. Build statistical capacity of the participating countries in the areas of surveys and of literacy assessment. (10)

UNESCO sees a huge need for a project like LAMP project "... because most current data on adult literacy in developing countries are not sufficiently reliable to serve the needs of national and international users. For example, the data generally relies either on individuals' self-declaration of their literacy or on "proxy" indicators such as their educational level. With literacy at the top of the development agenda, good data are needed in order to help target and design appropriate actions." (5)

### **Methodology and competencies assessed**

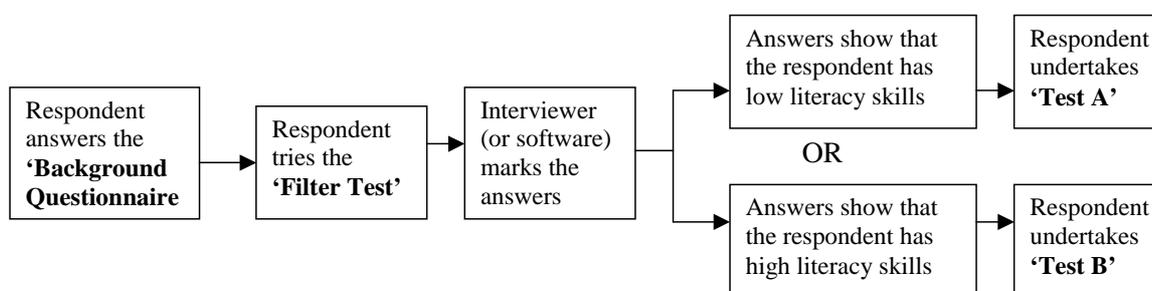
LAMP is a 'direct assessment of the literacy level of the adult population of different nations', which will measure 'a spectrum of literacy levels, from very basic reading and writing, to the higher level skills needed to participate fully in a learning society'. LAMP seeks to assess 'real-life literacy' and states that literacy is more than just elementary reading and writing. It involves a continuum of learning, and has many aspects and contexts such as: locating information in a bus timetable, reading for learning mechanics, writing to keep track of spending on a community project, or filling in government health forms. (11)

The assessment is based on a sample survey of adults (people aged 15 or more). The respondents will normally be visited at home and asked to complete literacy tasks in order that their individual proficiency in literacy can be measured. LAMP will assess three core aspects of literacy: reading, writing and numeracy. (10)

The questionnaire starts with questions on the respondents' educational background and methods of acquisition of literacy (formal, non-formal, informal), their use of the skills, their linguistic, social and economic situation, their perceived skills and needs, among other characteristics. The exact topic areas and questions are determined by the participating countries, with a core of common questions and separate countries modules. (7)

A locator test directs respondents to a higher (Test B, based on the IALS methodology) or lower (components) level assessment module (Test A). At the lower levels of literacy, the following five components will be used to differentiate: oral comprehension, letter/grapheme recognition, word recognition, sentence reading and passage reading. (6) The threshold for the filter test corresponds to the border between levels 1 and 2 in IALS.

**Figure 1: routing through the Literacy test (6)**



Test A asks the respondents to do some 'easier' exercises. The assessment of the lower literacy levels focuses on four components (letter, word, sentence, paragraph), and three aspects (understanding, accuracy, and speed). A profile of the respondents will be built based on their performances on Test A. For example, whether they still have problems with recognising isolated letters, whether they can read and write simple isolated words, whether they understand simple sentences and paragraphs, how quickly they can read them. The five components provide a hierarchical structure for profiling the literacy skills of individuals at the lower levels of literacy and provide information to group these individuals according to training needs. Each component should bring new information not provided by other components. Measuring speed and accuracy was deemed essential because it is critical for effective reading, motivation and progress. It will provide information on the fluency and sustainability of the literacy skills.

- Component 1: listening comprehension (in the language of assessment) – comprehension of vocabulary in context and comprehension of the overall text. Listening comprehension is related to memory, which is related to reading ability (or potential reading ability). There are also non-literacy reasons to include oral language comprehension such as measuring an individual's understanding of abstract and de-contextualised language (how they comprehend the oral message). It will therefore provide information useful for improving the effectiveness of oral communications, for example, radio messages in an AIDS campaign. It was noted however that assessing literacy skills in the same language as used in writing would not measure oral cognitive skills if the written language differs substantially, in words or grammar, from the oral dialect.

- Component 2: recognition of grapheme (letter, syllable, word component symbol or other depending on the writing system) – speed and accuracy. This component is of policy interest because it will show the extent to which individuals have learned basic elements of literacy but may not have acquired reasonable reading speed and fluency. It can help identify where improvements are needed in the education system, whether formal or non-formal.
- Component 3: word recognition – speed and accuracy. This component is also of relevance for policymaking, for the same reasons as mentioned for component 2. It was noted that components 1, 2 and 3 may be correlated and this should be investigated further in the feasibility studies.
- Component 4: sentence reading – speed and accuracy. Sentences provide a transition between words and texts.
- Component 5: passage reading – speed, accuracy and comprehension. Enables measurement of speed and accuracy on actual text reading and checks the most basic understanding. (7)

Test B is based on the IALS methodology. The focus is again on reading, writing and numeracy although for test B, due to time and resources constraints, as well as a need to avoid complexity, these might not be differentiated into separate scales. (10)

To project the results from the sample to the whole population LAMP uses model-based estimation. Direct estimation will be used for the overall figures, and model-based estimation will be used for smaller subgroups. This approach means that only a small-scale survey is required. Indeed the technique requires fewer interviews yet provides estimates of the same accuracy as the more traditional direct estimation methods. It does so by using knowledge of the relationships between literacy and demographic and socio-economic characteristics, and of auxiliary data available from larger surveys. The same technique could be used to derive literacy estimates in years where no assessment is conducted. (11)

## **Results**

The main aim of LAMP is to monitor levels of literacy in different countries. LAMP took a major step forward in assessing (il)literacy through direct tests instead of self-report or proxy methods. Still, as the study is still in its pilot phase it is hard to assess its success. The fact that only two countries have officially signed up and two intentionally, is a somewhat worrying sign, even though they just serve as pilots. A longitudinal study in a certain country could be interesting, but for an international comparison of progress, many more countries need to adopt the LAMP survey.

## **2. Project evaluation**

### **Utility for other purposes**

The LAMP method is aimed at (inter)national comparison over time. Whether this is possible needs to be proved once the pilot phase is finished. To balance between global comparison and national specific contexts, the monitor consists of two parts: one core which is the same in each national survey, but in a different language and one which is country specific. As in IALS, LAMP recognises the influences of language differences on the literacy assessments. These influences are even more problematic in the lower levels of literacy. (See discussion on validity) It was agreed that in developing the lower level assessments for LAMP, the aim should be to obtain comparability across languages in construct, recognising that it will not always be possible to ensure equivalence

of the tests. (7) Currently, Morocco and Mongolia have already signed up, while Jamaica and Kenya have shown interest and are negotiating. (9)

While LAMP is focussing on literacy, other aspects are measured too. For example component 1, on listening comprehension, is related to memory, which is related to reading ability (or potential reading ability). Also, measuring oral language comprehension will provide information useful for improving the effectiveness of oral communications, for example, radio messages in an AIDS campaign. It was noted however that assessing literacy skills in the same language as used in writing would not measure oral cognitive skills if the written language differs substantially, in words or grammar, from the oral dialect.

### **Utility for policy making**

In constructing the components of the LAMP survey, one of the design criteria clearly states that: “They must relate to the acquisition of literacy especially in developing countries context and provide information useful for designing policy interventions (e.g. curriculum development).” For example, components 2 and 3 (recognition of graphemes and words) are of policy interest because they will show the extent to which individuals have learned basic elements of literacy but may not have acquired reasonable reading speed and fluency. It can help identify where improvements are needed in the education system, whether formal or non-formal.

Therefore, “LAMP will provide global and regional benchmarks of literacy against which individual countries can measure progress and in particular in relation to the EFA targets.” (8) The monitor addresses some specific policy issues, such as: ‘In 2000, the Education for All (EFA) programme set the aim of achieving a 50% improvement in adult literacy by 2005. Literacy is also a key indicator for the Millennium Development Goals. The UN Literacy Decade started in 2003. Yet without good measures, progress in improving literacy cannot be properly monitored.’ (8) Also, in shifting the definition and the measurement from a dichotomy towards a continuum of skills, it will not be possible to present literacy and illiteracy rates as quoted presently when measuring progress of international goals and compiling indices, but rather have a more subtle understanding of the level of literacy for fine tuning the curriculum. (11)

A potential policy utility not mentioned in LAMP, but of much relevance to European policy, is its use for immigration policy. Europe is currently coping with a large immigrant population, among which there is a significant portion not able to communicate effectively in a European language, while some are not even literate in any language. Countries such as the Netherlands or the UK, have implemented language tests for new comers and demand a certain level before access is granted, but a general overview of the level of literacy among immigrants in Europe, with consistent levels would enable measurement of progress.

### **Quality and relevance of skills definitions**

LAMP adopted the IALS definition of literacy and numeracy. Literacy is defined as “the ability to understand and employ printed information in daily activities, at home, at work and in the community – to achieve one’s goals and to develop one’s knowledge and potential.” Numeracy is defined as “the knowledge and skills required to effectively manage the mathematical demands of diverse situations.” (3) LAMP also uses the IALS levels of literacy for the higher levels of literacy. This required a re-labelling the higher levels. For example if the components in which literacy is

measured create five groups/levels then IALS level 2 would become group/level 6 in LAMP, IALS level 3 would become level 7 etc. (7)

With regard to the components of the survey, the following selection criteria were used:

- Components must enable differentiation between individuals with different literacy skills or with different components skills.
- They must relate to the acquisition of literacy especially in developing countries context and provide information useful for designing policy interventions (e.g. curriculum development).
- They should measure an individual's ability to read for information purposes (e.g. not just fluent reading without understanding).
- They should be feasible to administer in a variety of languages and of cultural contexts.

According to these criteria, five proposed components were selected, as described above, while a number of proposed components were deemed less appropriate for the following reasons:

- Spelling: deals with a higher level of ability and is irrelevant in some languages (e.g. spelling mistakes are a rarity in shallow orthography languages).
- Memory (Digit Span test): does not discriminate well between people of different literacy levels, and may not be reliable if respondents are not motivated to try as hard as they can. Instead, memory can be assessed through ability to cope with longer sentences.
- Vocabulary (PPVT or similar): time-consuming test that is difficult to implement and to render culturally appropriate. In addition, it may not reveal much about literacy. Listening comprehension will serve to give an insight into mastery of vocabulary.
- Awareness of print: It has been shown by SIL International that awareness of print is highly correlated with formal education, and therefore not very useful for LAMP. Some information on this could be collected through the background questionnaire (asking questions about use and sources of information) rather than as part of the assessment.
- General cognitive deficiencies due to improper nutrition: measuring this would be complicated and very ambitious. If LAMP is successful then the World Health Organisation may be interested in a follow-up study that would investigate these deficiencies and their relationship to literacy. (7)

### **Validity of assessment methods**

Livesley notes that one of the major obstacles to understanding more about literacy is the discrepancy between nationally specific and cross-nationally comparable data. (1) "The tests are adapted to ensure that they are appropriate to people's circumstances. This is important in order to ensure the validity of the results given the aims of LAMP to assess functional literacy. Indeed, it is clear that issues of contexts, languages and cultures arise at all levels of proficiency. These might be more acute for the lowest levels, partly because people with low literacy skills may rely more on context when engaging in literacy activities, and partly because this area is still very much in development, with very few cross-linguistic experiences to learn from. In addition, the lower levels are more closely affected by the nature of each language. As an example, once phonetic awareness is acquired, other things being equal, a learner in a language where the relationship between sound and symbol is relatively stable (such as Italian) will be closer to achieving basic literacy than a learner in a language with irregular spelling (such as English). Thus, although it remains useful to build tests that are comparable in construct, it is important to recognise that results equivalence across languages and countries may not be achievable for all of the lower literacy levels." (10)

### **Reliability of assessment methods**

The LAMP project team claims that ‘most data on adult literacy are not sufficiently reliable to serve the needs of national and international users. Generally, they rely either on individuals’ self-declaration of their own literacy or on ”proxy” indicators such as educational levels. These are indirect measures, which have been shown not to reflect reality very accurately. Moreover, they are not always collected on a consistent basis, so can be difficult to compare, and there are many data gaps. More reliable measures require people’s literacy ability to be assessed directly, in surveys that can test their skills.’ They base their argument that ‘direct measurement has been shown to work’ on IALS. (8)

Another matter concerning reliability has to do with their ‘synthetic estimation’. Making use of the relationship between demographic and socio-economic characteristics however, their reliability depends highly on good national totals of demographics and other auxiliary information. One advantage of synthetic estimation is that they can estimate literacy levels in interim years when no assessment is conducted for a particular country. (8)

### **The quality of data collection**

Even though the LAMP survey starts out with pencil and paper, the project team will try to employ computers. The perceived advantages are as follows: it minimises variations in test administration (especially for timed tests) and scoring; it minimises the occurrence of errors of interviews-assessors that are not highly skilled in pen and paper literacy assessment; it will be easier to conduct the assessment in communities with various languages; and it will facilitate data capture. (7) The status of computer use is still unclear in the pilot phase.

## **3. Lessons learned**

Based on the experiences with LAMP, some lessons can be drawn for a European strategy on skills assessment:

- It should be considered to adopt the literacy scales of IALS, elaborated with the additional levels of LAMP. The lower levels are especially relevant in relation to immigration policy. More and more countries see language as the main focus for integration of new comers. But a significant segment of immigrants are even not literate in their own language. This specific group is currently assessed through self report before joining class, while there are many levels, each demanding different curriculum. Assessment of literacy levels could both enable a better match of the curriculum to specific needs, as well as monitor progress on literacy policy.
- It is a challenge to strike the right balance between measuring skills in a specific national context and enabling international comparison through universal items. The content of literacy surveys depends, naturally, on the language of the population and cultural context in which the language is used. In LAMP, it was agreed that in developing the lower level assessment, the aim should be to obtain comparability among languages in construct, recognising that it will not always be possible to ensure equivalence of the tests. (7) A European assessment will need to deal with a large number of different languages and cultures too. LAMP tries to find the right balance through a universal international core of common question and a set of separate country modules. The exact topic areas and questions are determined by the participating countries,

while the differences in socio-economic contexts are assessed through the background questionnaire. Based on the test specification and examples provided by the contractor, countries will develop items. These will be real life items. For example, a real newspaper article, or words that are found on sign posts. Another way to relate literacy performance to the national context is assessing the way in which the skills are acquired. LAMP distinguishes between formal, non-formal and informal acquisition of literacy.

- In this connection it must be taken into account that context dependency is more acute for the lower end of the literacy scale. Project Coordinator Livesly states that people with low literacy skills may rely more on context, and partly because this area is still very much in development, with very few cross-linguistic experiences to learn from. In addition, the lower levels are more closely affected by the nature of each language. As an example, once phonetic awareness is acquired, other things being equal, a learner in a language where the relationship between sound and symbol is relatively stable (such as Italian) will be closer to achieving basic literacy than a learner in a language with irregular spelling (such as English). Thus, although it remains useful to build tests that are comparable in construct, it is important to recognise that results equivalence across languages and countries may not be achievable for all of the lower literacy levels. (10)
- In a European survey, it should be considered to give an indication of levels of stratification within the survey. The right balance should be found: more levels result in a more precise measurement, but could demand a larger sample (smaller cells) and affect the testing time. LAMP differentiates between two levels of testing: one for the higher levels, based on IALS, and one on the lower levels.
- The LAMP-initiative underlines the significance of clear descriptions on the division of tasks and responsibilities in implementing skills assessment initiatives. If, for example, national states carry out the surveys, clear guidelines should be agreed upon. The organisation plan assigns a central position to the national minister of education and gives a description of the required roles and responsibilities in order to carry out the national project.

#### 4. References

- (1) World Bank on LAMP: [http://www.World Bank.org/wbi/B-SPAN/sub\\_reassessing\\_literacy.htm](http://www.World Bank.org/wbi/B-SPAN/sub_reassessing_literacy.htm)
- (2) Presentation Denise Livesley, project director: [http://www.WorldBank.org/wbi/B-SPAN/reassessing\\_literacy/reassessing\\_literacy\\_lievesley\\_slides.pdf](http://www.WorldBank.org/wbi/B-SPAN/reassessing_literacy/reassessing_literacy_lievesley_slides.pdf)
- (3) Presentation Scott Murray (IALS/ALL) feb 03: [http://www.World Bank.org/wbi/B-SPAN/reassessing\\_literacy/reassessing\\_literacy\\_murray\\_slides.pdf](http://www.World Bank.org/wbi/B-SPAN/reassessing_literacy/reassessing_literacy_murray_slides.pdf)
- (4) Presentation : Douglas Lynd, Senior Programme Coordinator UIS Education Statistics Programme: <http://www.World Bank.org/data/conference/6b.ppt>
- (5) UNESCO on LAMP: [http://www.uis.UNESCO.org/ev.php?ID=5243\\_201&ID2=DO\\_TOPIC](http://www.uis.UNESCO.org/ev.php?ID=5243_201&ID2=DO_TOPIC)
- (6) LAMP Progress note 28 May 2003:  
<http://www.uis.UNESCO.org/TEMPLATE/pdf/LAMP/HTML/Update030528.htm>
- (7) Summary of discussion: Developing an assessment module for basic literacy levels, Meeting Montreal, 16-17 April 2003 UNESCO Institute for Statistics (UIS):  
<http://www.uis.UNESCO.org/TEMPLATE/pdf/LAMP/HTML/Apr03MeetingNoteFinal.htm>
- (8) LAMP leaflet: <http://www.uis.UNESCO.org/TEMPLATE/pdf/LAMP/LAMPLeafletEng.pdf>
- (9) Telephone Interview with Denise Livesley, Director UNESCO Institute for Statistics, 15 January 2004

- (10) Lyon and Uphingh, The Literacy Assessment and Monitoring Programme (LAMP): Background and Approach (Draft for discussion purposes) November 2003
- (11) Terryn, Benedicte, The Literacy Assessment and Monitoring Programme (LAMP): A new Approach to Literacy Measurement (Draft for discussion purposes) September 2003
- (12) Terryn, Benedicte, The Literacy Assessment and Monitoring Programme (LAMP): Country Implementation (progress report) September 2003
- (13) Correspondence Benedicte Terryn, contact person LAMP

# The 1997 and 2001 UK Skills Surveys

## 1. Description

### Goals and ambitions

The background to the 1997 Skills Survey, undertaken with a nationally representative sample of respondents in England Wales and Scotland aged 20 to 60, was an academic research programme of the UK's Economic and Social Research Council, entitled "The Learning Society". The survey generated many findings of its own, and served as a background for a number of other qualitative research studies within the programme. The survey thus had multiple aims, but a central one was to measure the stock of skills being utilised in British workplaces. By collecting a rich set of supporting information about the jobs in which these skills were exercised, the objective was to analyse the various antecedents of different types of skills, and the extent to which economic performance measures (principally, wages) were associated with different skills.

The 2001 Skills Survey was designed, with same main objective, as a follow-up to the 1997 Skills Survey. It was financed by the Department for Education and Skills (then the Department for Education and Employment), following the interest shown by government in the findings of the earlier survey, and its usefulness as background to the thinking of the National Skills Task Force in 1999.

Taken together, the specific objectives of the two surveys were:

- To generate statistics for the levels of broad and detailed skill types actually in use in workplaces in Britain. A central principle was to use identical repeated questions in successive surveys, using high-quality representative samples of eligible respondents covering the whole of Britain, in order to generate a picture of change over time.
- To make methodological advances in the use of job analysis for measuring skills. Improvements were sought in assessing the extent of under-utilisation and over-utilisation of skills, and in several other areas, without compromising the intention of retaining comparability of the central skill measures with earlier surveys.
- To permit an analysis of the distribution of all skill types in use across British workplaces, and to provide a benchmark picture of the skills stock in 1997 and in 2001, suitable for comparison with each other and with future similar surveys. (Early 2001 was also the start date of a new institutional regime in training and further education in Britain, with the introduction of the Learning and Skills Councils.)
- To collect information from jobholders about characteristics of the organisations that they work for that may be relevant to the organisation's demand for and development of skills.
- To collect information about recent changes at the workplace that may be associated, directly or indirectly, with changes in the skill demands of jobs. These last two objectives were intended to facilitate analyses that develop understanding of the way skills change at workplaces.

### Competencies assessed

In addition to the conventional measures of occupation and educational qualifications, the two surveys measured utilised skills in two ways.

First, there were three indicators of the broad level of skills required in the job, measured in terms of the total training time required to do the job, the time spent learning on the job in order to become fully competent, and the qualification level required by employers for new recruits to the job. These three measures were designed to be identical to ones used in earlier surveys in 1986 and 1992, thereby facilitating an analysis of trends over a fifteen-year period.

Second, the surveys adapted the methods of occupational psychology to generate 38 items describing the generic activities involved in doing the job. The choice of items was informed by theories of skill and the practices of commercial psychology; but to reduce the multiple items to a smaller and more meaningful set of generic skills, statistical techniques were used to generate two measures of the importance and sophistication of computer use in jobs and ten other generic skill indicators from the responses on these items. The ten skills are: literacy, numeracy, technical know-how, high-level communication skills, planning skills, client communication skills, horizontal communication skills, problem-solving, checking skills and physical skills.

No attempt was made to measure motivations and beliefs. Therefore, what is being measured are, therefore, strictly speaking skills in use. They do not, and were not intended to, encompass the wider concept of competencies.

### **Applied Methods**

The basic method of measurement is the conducting of a nationally representative social survey, with multiple questions about the requirements and activities of respondents' jobs.

Underpinning this approach are three assumptions. First, it is assumed that measures of skills in use in jobs are a reasonable proxy for the skills of the jobholder. If an individual is using a computer for advanced programming, for example, it is assumed that he/she has the relevant skills, or would not have survived in the job. Nevertheless, discrepancies between jobholders' skills and job requirements are possible, however, and supplementary questions are asked to ascertain subjective views about skills mismatches. Some individuals may have an excess supply of some skills, and not be using them fully on the job; others may have insufficient skills for the job they are doing, and may survive despite the consequent poor performance. These mismatches are dynamic: they can appear and disappear as both jobs and people change.

The second assumption is that the individual is the best-informed person to report about the job he/she is doing. All jobs differ, even within quite narrowly categorised occupations, and normally one would expect the jobholder to know best.

The third assumption is that the individual reports these activities in an unbiased way. This assumption is arguable: individuals might talk up their jobs, to boost their self-esteem. But, it is held that they are less likely to do so when reporting their activities than reporting their competencies in the performance of these activities.

### **Results**

A substantive body of research findings has been generated by the two surveys. These fall into the following four areas:

- Descriptions of the distribution of skills across the population of employed people in Britain.

- Generic skills were found to differ across occupations and industries in expected ways; and the jobs held by women used on average somewhat different generic skills from those held by men (Felstead et al, 2002). A big difference is found in the use of generic and broad skills between full-time and part-time jobs, even after allowing for other job characteristics (Felstead and Gallie, 2004).
- Descriptions and analyses of how skills have changed in British workplaces.
- The analyses focus both on the changing utilisation, and the changing extent to which educational qualifications are being matched with the recruitment requirements of employers (e.g. Felstead et al, 2002; Dickerson and Green, 2004). Successive surveys show a continued rise in the utilisation of skills, as shown in the broad skills measures, between 1986 and 2001. Between 1997 and 2001 there was a rise in all types of generic skills, with one exception, namely physical skills, which showed declines in utilisation. These changes were not just the result of changing occupational structure: they were occurring within jobs. While the rise in the use of most generic skills was quite modest, as would be expected over a relatively short period, there was over this period a dramatic increase in the utilisation of computing skills, and on top of that quantitative increase a slight increase also in the average level of sophistication with which computers were used.
- Analyses of the antecedents of skills, with an emphasis on the respective roles of the education/training system and the workplace.
- High levels of broad skills (required qualifications, training and learning time) are associated with jobs involving computerised or automated equipment. Education is especially important for the generation of computing skills, but much less so for the generation of other generic skills. On the other hand, “new” human resource practices, such as appraisal systems, and workplace practices such as the use of continuous improvement circles, were found to be strongly related to high skills and skill acquisition.
- Analyses of the association between the utilisation of skills and economic outcomes, both in terms of wages and in terms of subjective well-being.
- A substantive wage premium is found to be associated with computer skills, and with high-level communication skills (such as the making of presentations, long report-writing), and with planning skills. Other generic skills had no wage premium attached to them in the labour market, and physical skills usage was in most occupations a signal for lower wages (Dickerson and Green, 2004).

## 2. Evaluation

### **Utility/usefulness for policy development or other purposes**

The surveys have been useful for academic research on the labour market -- including the description and understanding of changes in skills utilisation in the British economy, the sources of skill change in individuals' jobs, and the association of particular skills with good (or bad) labour market rewards. This academic research has fed into policy-related research, and served a few directed policy needs. The evidence that the surveys have produced about the growth in the utilisation of skills has been taken as support for its policies to improve education and training. Another function has been to help direct the government's thinking at the demand side as well as the supply side of the skills market. Another has been to add evidence in support of the government's policies on computer literacy.

But there are no specific policies that the surveys have been, or could have been, used to evaluate directly. The surveys have not been of much use in assessing policies for employability, one of the areas that governments find themselves most engaged in.

### **Quality and relevance of skills definitions**

Within their limits, the defined generic skills give a wide range of skills utilised in workplaces. The measured generic skills incorporate all those that have been alluded to in theoretical and case-study research, as being important for successful participating in modern workplaces.

The definitional limits are that the measures do not encompass measures of motivations and attitudes of respondents. At the time, this was a deliberate decision. Though recognising that attitudes and motivation are important for economic performance, the researchers took the decision that these were conceptually distinct from skills. Moreover, the development of the British Skills Survey approach pre-dated the DeSeCo project for the OECD, and took no account of its findings, either with respect to the broader relevance of the idea of competence, as compared to skill, or with respect to skills and competencies that are used in society generally (i.e. outside the workplace). The surveys have only loose measures of the extent to which jobs use occupation-specific technical skills. Intermediate technical skills relevant to particular jobs have been picked up only approximately through the role of require technical qualifications, and through some items in the job requirements part of the questionnaire. Occupation-specific technical skills may be very important in certain jobs.

Finally, the relevance of the skills measures is limited to the population of employed people. For unemployed and economically inactive people, job analysis is directly impossible by definition, though it would be possible to analyse the jobs of those who had recently become unemployed. For those who had not worked for a long time, assessing their job skills would require an imaginative and difficult adaptation of the skills survey approach, which would almost certainly be confined to a narrower range of skills than can be measured for jobholders.

### **Validity and Reliability of assessment methods**

The approach adopted had the big advantage of generating measures of a wide range of generic skills, for most of which there are no accepted tests that could be carried out for a representative sample of the population. Moreover, the measures come at the comparatively low cost of a good-quality survey, rather than requiring expensive testing procedures. The extra expense of skills tests derives partly from the developmental costs, which are to some extent non-recurring; but tests also cost more if one wants to ensure reasonable response rates. Even with extra resources for quality control, survey methods are likely to obtain higher response rates. The Skills Survey approach also has the advantage of measuring the skills as utilised in the workplace, rather than treating the individual's skills out of context. Research has shown that context is important for the acquisition and use of generic skills (e.g. Stasz et al, 1996).

Nevertheless, the job-requirements method of skills assessment raises a number of issues concerning the validity and reliability of the measures obtained. With regard to validity, as stated above the measures may be valid measures of the skills actually being utilised, but are only proxies for the skills possessed by working people. Validity tests have mainly comprised examinations of the distribution of broad and generic skills, to see how far they correlate as expected with conventional measures of skills based on occupation or educational attainment. These examinations

confirm that the measures are correlated in the expected way: for example, higher skilled occupations tend to have higher skills requirements. However, this is only a weak test of validity, because the conventional skills measures are themselves inadequate – indeed, that is the part of the reason for exercise. A full test of validity would involve an in-depth analysis of individuals' jobs, to see how far their responses on individual items correctly pick up the skills that such an in-depth analysis would reveal. A test like this would be prohibitively expensive, if extended beyond a small number of participants in a range of occupations. A small number of items were examined in this way in the piloting for the surveys, and a fraction of these were removed when found to be invalid or to have ambiguous meanings for respondents and researchers. The main broad skills questions and some of the generic questions were examined in this way, and were found to be valid measures of what they were aiming to capture.

With regard to reliability, there has been no assessment of whether and how far individuals might be biased in the reporting of their job activities (see assumptions above). There has, however, been an examination of inter-rater reliability between jobholders' and line-managers' assessments of jobs (Green and James, 2003). These show some differences appearing according to the match between the gender of the employee and the line-manager, with respect to certain skills; fairly good agreement with regard to most of the generic skills; but some disagreement with regard to contentious aspects of jobs such as the extent of task discretion available to jobholders. However, this study did not carry out a separate assessment of the skills used in jobs, and so made no judgement as to whether the jobholder or the line-manager made the most accurate judgement as to the activities involved in the job.

Overall, a fair judgement is that the skills measures do capture the range of generic skills utilised in jobs; they are not to be regarded as direct measures of the skills possessed by individuals, but can be seen as providing proxy measures through the assumption that on average skills possessed are matched to the requirements of the job. Some subjective items on skills mismatch can be used to make judgements about the correctness of that assumption.

### **The quality of data collection**

In both surveys, data collection was carried out using face-to-face interviews conducted in people's homes. One advantage of this method is that it enabled respondents to describe their work away from any potential fears of pressure or loss of confidentiality in relation to supervisors and other workers.

Clustered random sampling methods were used to select households from a database of addresses contained in the Postcode Address File. In selected households, one individual was selected at random among all those living there who were eligible for inclusion. Households were re-visited on multiple occasions, and at different times (evenings, weekends etc.) as necessary to ensure as high a response rate as possible. Usually, it was necessary to make an appointment to interview the eligible person at his/her convenience. As a result, the response rates were 67% in the 1997 Survey and 72% in the 2001 Skills Survey. A slightly higher non-response rate for men was allowed for in all analyses by the combining a gender weight (derived from Labour Force Survey estimates) with the sampling weights derived from the probability sampling procedures. These response rates are not perfect, but are higher than often achieved. The achieved samples, which were nationally representative, were 2467 in 1997, and 4470 in 2001.

These results were achieved by paying a lot of attention to interviewer training, and to insisting on persistence in returning to non-responding households. No payments were made to respondents. The two surveys could be said to have largely achieved the objectives set for them. Advances in the methodology of skills measurement were made; analyses of skills trends and of the roles of skills in the British labour market have been produced, and have been validated by peer review. The analyses have contributed to policy-making, though not through pushing any specific policies. There is currently interest in the Department for Education and Skills for a new survey that will update results to the mid-2000s. The time series of measures of the stocks of skills being used the British workplaces is likely to prove the most telling contribution to researchers' and policy-makers' understanding of modern workplaces and the changing skills that are being required. Their weakness is that no attempt has been made to measure the skills of unemployed people. This has meant that no analyses have been carried out of the issue of "employability", since no control group of non-employed people is available in the surveys to compare with those who are in jobs.

### **3. Lessons learned**

Some but certainly not all of the purposes of a European assessment might be served by something similar to these surveys. The essential principle of measuring skills in the context in which they are used is theoretically defensible; and the reasonable success of these surveys in their own terms mean that the principle can indeed be put into practice.

The surveys have been relatively cost-effective. Though good-quality surveys are expensive, they are likely to be cheaper than test-based methods. It is difficult to see how results can be achieved at lower cost; indeed, if there is concern to extend analyses to greater regional or even local disaggregation, it would be necessary to enlarge the sample considerably, with concomitant extra costs. The current size of approximately four thousand allows, however, several useful industry and occupation-based disaggregations, and is sufficiently large for most purposes. The surveys are also able to satisfy all reasonable ethical requirements; it is possible to guarantee confidentiality to respondents. Respondents did not have to feel threatened, as some may have done when asked to perform tests in their own homes.

The usefulness for policy purposes of this approach to measurement has arisen indirectly in the British case. The original survey was not designed with specific policy analyses at the forefront. Nevertheless, the two surveys together have opened the way to improved understanding by governments of the way that skills are used in modern economies, and how this is changing over time. This understanding is essential as a basis for informed policy-making. In an international assessment policy might have to be more emphasised. This would necessitate specific attempts to adapt the method to apply to recently unemployed persons, and perhaps to discriminate more finely among respondents at the lower end of the skills spectrum.

There arises, finally, the question as to whether the methods are adaptable for application in other countries. The surveys were not designed originally with any international project in mind, but the principles are applicable to any country. It would be possible to carry out a similar survey in other countries: the advantages and the disadvantages would not be any different from those encountered in Britain.

The more difficult but crucial issue concerns whether the results of such a survey can be made comparable across countries. The problems of international comparison are similar to those present with other methods of assessment. First there is the question of language and translation. These are resolvable through appropriate techniques. Second is the issue that the norms of assessment and reporting by individuals in different countries may vary. Therefore, it would be necessary to alter the response scales from those used in the Skills Surveys. Occupational psychologists use either “importance” scales or “frequency” scales of responses, and sometimes both. These two surveys used “importance” scales, but these are inappropriate for international comparison. “Frequency” scales could be instead devised in an attempt to ensure valid international comparisons, but these would need evaluation and adaptation in an international pilot. At any rate, this issue of comparability is the most important issue for resolution within the proposed strategy.

#### 4. References

Ashton, D., B. Davies, A. Felstead and F. Green (1999). *Work Skills In Britain*. Oxford, SKOPE, Oxford and Warwick Universities.

Felstead, A., D. Gallie and F. Green (2002). *Work Skills In Britain 1986-2001*. Nottingham, DfES Publications.

Green, F. and D. James (2003). "Assessing skills and autonomy: the job holder versus the line manager." *Human Resource Management Journal* 13(1): 63-77.

Stasz, C., K. Ramsey, R. Eden, E. Melamid and T. Kaganoff (1996). *Workplace Skills in Practice*. Santa Monica, RAND.

A list of all relevant publications that make use of the surveys can be obtained from the following web page: <http://www.kent.ac.uk/economics/staff/gfg/currres.html>

# The Skills for Life Survey

## 1. Description

### Goals and ambitions

The background to the survey was the government's recognition that there was a significant problem of lack of literacy or numeracy in a minority of the adult population in Britain. The main initial impetus for this was the International Adult Literacy Survey, in which Britain participated in 1995. The findings from this survey formed the basis of a report for the government (DfEE, 1999) (the "Moser report"), which called for a clear strategy from government to raise basic skills among adults. By basic skills was meant literacy, numeracy and ICT.

A number of programmes have been introduced since the Moser report to fill the need for basic skills improvement. The cornerstone of this strategy was the definition of a set of national standards for literacy and numeracy, categorising each into five broad levels. Preliminary work had also defined some basic standards in ICT. The Skills for Life Survey, commissioned by the Department of Education and Skills and carried out between June 2002 and May 2003, was designed to produce a profile of adults' basic skills in England, setting them in the broad levels defined by the national standards. It was also intended that the skills profile should be measured with sufficient precision at regional level.

### Competencies assessed

The survey was primarily aimed at measuring the basic skills, of literacy, numeracy and information and communications technology. It also aimed to measure proficiency in English language among those for whom this was not their first language. The main emphasis was on the lower skill levels. The relevant concept in each case is skill, rather than competence.

### Applied Methods

The method of investigation was in two parts. In the first part, literacy and numeracy were tested, alongside a background questionnaire that collected standard demographic data, and asked respondents to self-assess their skills and the impact of those skills on their lives. The size of the achieved sample was 8,730. The sample was obtained by a clustered random sampling procedure, with a target population of adults aged 16 to 65. Households were the initial basis for sampling, and then in each selected household one eligible member was selected at random for participation. Those who agreed to participate were paid £10. A response rate of 59% of eligible people was achieved.

In the second part, 4656 respondents took part in a follow-up interview, between two and six months afterwards, in which their basic ICT skills were assessed. These participants were paid an extra £5 for participation. The number of respondents to this part of the survey was lower mainly because a minority did not agree to take part or could not in the event be re-contacted. The ICT assessment comprised several multiple-choice questions on ICT awareness, followed by a practical test of ICT skills in a Windows environment. In each case, a routing procedure was followed to direct respondents quickly to relevant items. Those with no ICT awareness were, for example, not even asked to do the practical test; they were assigned the lowest ICT level automatically.

## **Results**

With respect to both literacy and numeracy, the results classify respondents according to the national standards at Entry Level (itself sub-divided into three levels), Level 1, and Level 2 or above. Because of the policy focus, the results are mainly devoted to analysing the proportions of adults who are at Level 1 and at the various entry levels.

The report finds that 16% were classified as below Level 1 in literacy, amounting to an estimated 5.2 million adults in England. This figure included 5% of the population who were below Entry Level 3.

With respect to numeracy, a greater proportion of the population was classified at the low levels, compared to those at low levels of literacy. Thus, nearly one in two (47%) were below Level 1 in numeracy. At the upper levels, only 18% achieved Level 2 or above in both the literacy and the numeracy tests.

The report placed 25% of the population at Entry level or below for their ICT awareness skills, 25% at Level 1, and 50% at Level 2 or above. However, on the practical test, 53% of the population was placed at Entry level or below. There were fairly close associations between the usage of computers at home and at work, and performance in the test.

The published report (Department for Education and Skills, 2003) goes on to detail the pattern of literacy, numeracy and ICT skills according to geo-demographic categories (occupation, gender, region, ethnicity, education etc.). It examines the correlation between the different skill domains. It also looks at the skills of those for whom English is a second language.

Unsurprisingly, self-assessed literacy and numeracy skills exceeded by a considerable margin those from the tests. Yet, when it came to ICT skills, respondents were quite close in their self-assessments.

## **2. Evaluation**

### **Utility/usefulness for policy development or other purposes**

As stated above, the objective of the survey was to provide a profile, so that the government could assess better the need for basic skills training, by comparing attainment with the national skills standards. The regional profiles are of particular relevance. The skills assessed are clearly particularly important for policy-makers to investigate, since basic skills are something for which there is a widespread consensus that the government has a responsibility to try to influence, both in the short and in the longer term, through its various education and training programmes.

The profile has been and is being used as support for targeting particular courses and groups. This is probably the most important policy function. A second important function within government as a whole is in the allocation of funds between departments. The estimates of the numbers of people below the second levels considered necessary for employability also provide important evidence-based support for the Department for Education and Skills' requests for treasury funding. Third, both the regional and the national skills profiles are being used in presentations to assist regions in developing their own skills strategies. In a few cases where the numbers permit it, the analyses have also been of use to Local Learning and Skills Councils, i.e. at the sub-regional level. Finally, there

is a more speculative possible future use of the survey. By identifying which detailed aspects of literacy or numeracy people are having problems with, the survey has provided background information that could be used for curriculum adjustment in future years.

The main usefulness of the self-assessment of skills is to assess the extent to which people perceive themselves as having low basic skills, since those who perceive no problem are less likely to be motivated into learning. The Department felt that this information was useful as an extra aide to designing policy. We judge, however, that self-assessment is not a useful route to consider for any international assessment.

A disadvantage of the methods used in this survey is that they do not permit direct comparison with earlier results from the International Adult Literacy Skills (IALS) assessments. If the tests had been designed to be identical – or at least some aspects of them – it would have been possible to track changes over time, and thereby begin to make an assessment of the overall impact of policies. The report discusses the difficulty of making comparisons, though it does then compare some of the patterns of skill across age groups, using a mapping of levels from IALS to the national standards. It cautions about the validity of directly comparing percentages in each level, but perhaps inevitably, given the published mapping, such comparisons were made for the benefit of the media. When a certain set of assumptions are made it can seem as if there has been a reduction in literacy problems between IALS and the Skills for Life Survey. As one critic has pointed out, in a report for the Basic Skills Agency (a quasi-independent UK government agency), a change in the assumptions easily leads to this conclusion being contradicted (Schicht, 2004).

Nevertheless, the decision not to use the same tests as in IALS highlights some of the issues facing the UK government. The policy requirement was to focus on the need for improvement at the lower end of the skills spectrum, and it was also found that the IALS tests did not permit an analysis of skills in relation to the national standards. For similar reasons the UK government did not participate in the OECD's Adult Lifelong Learning Survey (ALLS), but in addition it was felt that participation in that survey was too expensive, and that the pay-off was insufficient given the lack of participation by many other countries, especially the larger ones. One lesson is that any international assessment may have to set going a bandwagon of support: each country's support is increased by the number of other countries on board.

#### **Quality and relevance of skills definitions**

The different methods used in this survey would not have any particular advantage in a European-wide survey, unless other member countries were to adopt the same specific national standards as in England. Apart from the attempted conformity with those standards, there is no argument claimed that the methods used are in any other way superior to the methods used for assessing literacy in other countries.

#### **Validity and reliability of assessment methods**

The literacy and numeracy skills are tested using a series of multiple choice questions. Literacy skills comprised primarily reading skills, though a small number of writing skills were assessed (rudimentary spelling, punctuation and grammar). These methods are appropriate for matching to the English national skills standards. There is, however, some doubt as to the construct validity of the data. Schichte this approach has been used. This aspect of the survey is not worth considering in relation to any international assessment.

### **The quality of data collection**

The data was collected well, in that it was carried out by trained interviewers, and as prescribed by the survey client (the Department of Education and Skills). The chief problem lies with the achieved response rate. The first stage response rate, which was 59% is not all that good. It is less than is normally achieved on comparable government surveys that require no tests; it is, for example, much less than the response rate achieved in the Quarterly Labour Force Surveys. The original target was for a response rate of between 65% and 70%. Moreover, for the literacy and numeracy tests separately, the response rates were lower, at 53% and 54%; for both tests completed successfully, the rate was just 51%.

The low response rate was compounded in the second stage by further dropouts from the ICT tests. If one takes into account those respondents who refused to allow a further approach to them, and those who later refused to undertake the ICT assessment plus those respondents who could not later be traced for one reason or another, the response rate of eligible respondents was really only 31% (4656 interviews out of 14869 eligible), rather than the 70% of the issued addresses for the second phase. However, given the length of the first stage questionnaire, it is difficult to see how the ICT element could have been included at that stage.

These low response rates are unsurprising, given the nature of the survey, and the necessity of carrying out the tests. Although no analysis of the reasons for the low response rate are given, it was felt that the tests were likely to have put some respondents off; though apparently the response rate to the second stage was not affected by indices of social deprivation derived from the first stage data. A great deal of effort was put into maximising the response rate during the fieldwork stage. In retrospect, the judgement is that, with the tests and background survey taking more than 70 minutes on average, the response rate was probably as high as could be achieved. It looks like the prime problem was, however, the requirement to take the tests.

The largest single cause of failing to gain an interview at the first stage was the refusal by the selected person: roughly 17% of selected, eligible, persons refused. The effect of the payments does not seem to have rescued the survey from having low response rates, though of course they could have been even lower. The report shows how the questionnaire was designed to show what respondents could do, not what they could not do, and this was with the specific aim of encouraging respondents' continued participation. One positive aspect was that, once an interview and test had begun, very few respondents gave up.

While survey non-response rates are appropriately used to weight the findings for the purposes of generalisations to the population, analyses could be biased by the possibility that the propensity to respond is linked to variables of interest, not least the level of skill itself.

### **3. Lessons Learned**

The major advantage of this survey lay in the way it was tailored to fit the national standards that formed the backbone of English policy formation on basic skills at the time of the survey. If something equivalent were to apply at European level (that is, some European skills standards), the same consideration would hypothetically apply. In the absence of that, although the survey was successful in its own terms, there is nothing positive of major importance to be learned from this survey, relevant to a European strategy. The survey organisers took the view, however, that there

would be value in a European assessment of skills. Unlike the IALS, which is thought to have had a shock-tactic effect on British policy-making (leading to the greater emphasis on basic skills), any new European strategy is not expected to have such an effect. Yet, this is hard to predict: new skills might be highlighted as important and as deficient in any such comparison. Moreover, in any case a European assessment would still be of value in providing the quantitative background to the sharing of best institutional and policy practice from one nation to another.

This was quite an expensive assessment. The estimated cost per interview was in the region of £155 per interview, including all the survey costs and the analysis necessary for the published report. However, this includes neither the development costs, nor the costs of the ICT stage of the exercise. The unsurprising lesson for a possible European strategy is that testing methodologies, with all their advantages in terms of objectivity of standards, are nevertheless costly, and risk falling short of giving a nationally-representative description of skills owing to too low response rates.

The reliability of the results is partly undermined by the multiple-choice framework, which was dictated by a lack of time available to develop new instruments that matched the national skills standards. Multiple-choice questions were more easily developed on the basis of questions that had previously been validated in “key skills tests” applied in schools and colleges. Preferably, any tests to be used as part of the European strategy should avoid this approach if precision is required, especially at the bottom end of the skills spectrum.

It is not easy to see how the data collection could have been improved, yet maintain the same approach and length of survey. The ICT part of the survey was not, in our judgement, successful, because it had to be tacked on as a second stage some months later. Mainly as a result of that the response rate dipped to really quite low levels for a government-backed survey. The lesson would seem to be that it is very difficult to conduct tests in more than two skill domains, on top of a regular survey of background demographic data, with any one respondent. The Department for Education and Skills is currently reviewing its overall experience with the survey. Though the survey was not designed as part of a series, and though there are currently no plans for a repeat survey, it is possible that another one could take place within a few years. Meanwhile, there are definite plans to carry out a similar survey in Wales, and more tentative plans for a survey in Northern Ireland.

#### **4. References**

Interview with Gail Peachey of the Department for Education and Skills, 4/5/2004.

DfEE (1999). *A Fresh Start: Improving Literacy and Numeracy*. London, Department for Education and Employment (now DfES). See also: <http://www.lifelonglearning.co.uk/mosergroup/>

Department for Education and Skills (2003) *The Skills for Life Survey. A National Needs and Impact Survey of Literacy, Numeracy and ICT Skills*. London, DfES Research Report 490.

Schicht, T. (2004) *Critique of the Skills for Life Survey*. The Basic Skills Agency web site: <http://www.basic-skills.co.uk/site/print.php>

# The “Assessing Learning to Learn”-initiative

## 1. Description

### Goals and ambitions

The overall ambition of the Finnish learning-to-learn project is to measure learning-to-learn skills and use the results in the context of evaluations of school based teaching and education. ”Learning to learn”-skills is defined as the readiness and willingness to adapt to a novel task. It consists of a system of cognitive competencies and self- and context-related beliefs.

The Learning-to-Learn project was started in 1996 as a joint endeavour between the National Board of Education, the Centre for Educational Assessment at the University of Helsinki, and the City of Helsinki Education Department. Since then, a research group headed by professor Jarkko Hautamäki has developed and refined a comprehensive testing system, which has been implemented among Finnish 6<sup>th</sup> and 9<sup>th</sup> graders, and also in 2002 in the 17+ age group (vocational and general educational tracks, upper secondary school). The scales have been used also in a study among *folk high school* adult students involving a total of approximately 6 000 students. In total, more than 70 000 persons, mainly primary and lower secondary school students, have completed the tests in various versions since 1996.

The project presents itself as response to an overall policy on education in Finland in which increasing emphasis is given to the evaluation and assessment of educational outcomes. In this context, one of the key evaluation criteria has concerned the production of ”learning-to-learn” skills in education.

The latest versions of the test package have been translated into Swedish, English, Dutch and German. Translation into Swedish was done both with a view to carrying out tests among the Swedish-speaking minority in Finland, and with a view to testing the framework in a Swedish context (the test was carried out by the Swedish National Agency for Education among 500 Swedish 17+ aged pupils).

Translation into English was carried out in connection with the participation of representatives of the Learning-to-Learn initiative in two EU funded conferences, ”Learning to Learn as part of Cross Curriculum Competencies”, held in Finland in 1999, and has also been used in connection with the European Commission’s (DG Culture and Education) Standing Group on Indicators, the Expert Group on Basic Skills. However, the English language version has not been prepared with a view to actual testing.

Translation of some of the scales into Dutch was financed by the Dutch Government, who wished to test the usefulness of the test package in connection with the Dutch primary education system. Translation into German has been financed by the Centre for Educational Assessment because there is likely to be interest in the system among the German Länder for evaluating each area’s school system or using the scales for local developmental work with the schools.

## Competencies assessed

According to the methodological framework of the project, learning-to-learn competencies consist of both cognitive and affective competencies, of both knowledge of relevant facts and to the use of thinking and reasoning, and instances of emotional and affective experiences that happen during cognitive action.

Cognitive and affective competencies as well as social skills and study skills are assessed in the project's framework. The framework distinguishes between learning competencies, self-related beliefs and context related beliefs, cf. figure 1 below.



**Figure 1: The Conceptual Framework for the Learning-to-Learn Project**

The learning domain in turn comprises a set of variables including verbal-argumentational comprehension, quantitative-relational comprehension, logical reasoning, reflective abstraction, variables concerning the management of learning and affective self-regulation in connection with learning. All the components are not measured in the actual testing package, since some of them would require real-time observations or the use of expensive coding teams.

Self-related beliefs pertain to variables grouped into the categories learning motivation, action-control beliefs, “academic selves at school”, assignment/task acceptance, self-evaluation and future orientation. Finally, a number of variables are seen to measure relevant context-related beliefs: societal frames and perceived support for learning and studying.

During the six years since the pilot study, the test package has undergone considerable development. Even if the core of the instrument has remained the same, reflecting the theoretical construction of the concept of learning-to-learn, there have been changes both in response to the needs of adapting the instrument to serve students of widely differing ages, but also to reduce the original testing time of 6 to 8 hours. With the reduction of number of items into shorter scales (allowing the reliabilities to get somewhat lower, because no individual diagnosis is needed) pencil-and-paper form is now executable in 180 minutes of class time, supervised by a teacher. By using a computer-based form (pdf-based), testing time can be further cut to 90 minutes.

### **Applied Methods**

As for the assessment of cognitive competencies (the learning and reasoning domain in Figure 1 above), this part of the test requires more than half of the total testing time. Cognitive tasks are all in an easy-to-answer, easy-to-score multiple-choice format, even if the individual test differ in the why the task is set and in the type and number of choices. Three major forms are used:

- i) a standard two-to-five-choice format with only one correct answer
- ii) a true-false choice with, for some tasks, an additional question to test the pupil’s conviction of the correctness of his or her answer, e.g. whether the solution was concluded or guessed and how certain the pupil is of the answer,
- iii) a three-choice format for the assessment of the relative importance of different statements when reflected against a text that has been read.

As for the assessment of beliefs and attitudes, almost  $\frac{1}{4}$  of the total testing time is required for this part of the test. The different parts of the belief assessment are dispersed among and between competence-measuring tasks to avoid unwanted interaction between the different assessments. For example, the questionnaire for the assessment of learning motivation is presented to the pupils before any competence tasks are given.

All parts of the assessment of beliefs are carried out using self-report questionnaires. The format is a standard one offering statements on which the pupils are asked to take a stance in terms of the degree to which the statement reflects their opinions, their view of themselves, or their mode of action in different situations, for example. The response scale used is a seven point Likert scale with only the end points given a verbal description, where 1 equals (I feel that) “This is not at all true (of me)” and 7 equals (I feel that) “This is very true (of me)”.

As for background- or context related variables, there is a variety of options regarding the type of information to be collected. The actual variables to be used depend on the availability of different types of contextual information and on the different expectations concerning the use of the results at different levels of educational decision making. Gender, age and mother tongue could be considered the standard or minimum information to be used as classifying variables. The Finnish studies additionally included information on the educational level of the parents (both mother and father), and on the pupil’s school achievement during the year of the assessment or the previous year (GPA

and marks in some separate subjects relevant to the comparative analyses). Information about the pupils' possible enrolment in special education may also be significant. Contextual information concerning the school and its location (size, province and type of municipality) added to the data related to the individual pupil may contribute significantly to the comparative analysis and to later educational policy making.

Testing is organised at school level in normal classroom context. The assessment is to be performed in a normal classroom setting as part of an otherwise regular school day. The assessment is based throughout on paper-and-pencil or alternatively screen based task and questionnaires, with adjoining easy-to-read instructions, presented to the pupils in the form of booklets, in the paper version with separate fill-out sheets to allow for easy computerised scanning.

The assessment set includes detailed instructions for the supervising agent, who may be a teacher or a trained evaluator, but should not be the class teacher of the pupils in question. The assessment set includes detailed instructions for the administering of the testing in the form of a booklet for the supervising agent. The instruction booklet contains a description of the assessment procedure, the whole set of tests with adjoining separate instructions for the test supervisor, and approximate time frames for the separate parts of the test package. Apart from the instructions given in accordance with the Instruction Booklet, no additional help or instructions are to be given to the pupils during the assessment.

## **Results**

Analyses of the test results have followed two different paths: a variable based approach and a cluster-analysis approach. The results of the national studies have provided the norms against which the results of the individual school districts or schools have been compared, and can be compared when administering the scales for local schools. Analyses have also pointed to the significance of extra-scholastic factors, especially parents' education and the gender of the student, for learning-to-learn competencies.

The variable based approach has tried to identify the factors that contribute to Learning-to-Learn competencies, taking into consideration the various background variables (parents' education, gender, GPA and others). An overall result points to the conclusion that school-related factors explain about 10-15 per cent of the variance of pupils' learning-to-learn competencies.

It is characteristic of the Learning-to-Learn initiative that the various scales (i.e. strongly correlated questionnaire items) measuring the different learning-to-learn competencies are almost always treated as the dependent variable. Learning-to-learn competencies is the explanandum, the phenomenon whose variance is to be explained. As opposed, such competencies are not presently used to explain other later occurring phenomena, such as for instance career paths in adult life, incomes, or final education level obtained.

Many of the system's elements concerning cognitive competencies resemble ability and intelligence tests, and there is potentially considerable controversy or possibilities of misunderstanding connected to carrying out research which focuses on the relation between intelligence test scores among pupils and their subsequent success or lack of the same in secondary and tertiary education and working life. However, the project management is of the opinion that any learning-to-learn competence, problem solving or (meta)cognitive scales are bound to correlate with each other, with GPA and with any reliable intelligence or problems solving scale, potentially creating a space for

heated discussions of the roles of genetic, environmental, or destiny factors in scholastic achievement, working life success and economic outcomes.

## 2. Evaluation

### Utility/usefulness for policy development or other purposes

The test system of the Learning-to-learn initiative can be used for a number of different purposes. However, until now the initiative has primarily been used to provide information in relation to three themes:

- The reliability of Grade Point Averages in the school system.
- The "Educational Equity Balance"
- The competence and belief profiles of individual schools.

As for the first theme, the achievements of the school system at local, regional or national levels can be assessed on the basis of national achievement tests and schools' Grade Point Averages. The Learning-to-learn testing system provides a third possible way to assess school contributions to education and make possible comparisons between different schools. In this manner, the reliability of Grade Point Averages can also be tested. If there is a high correlation between the Learning-to-learn scores and GPAs it is evidence of a high reliability (provided that Learning-to-learn scores are valid measures of the various generic competencies assessed in the system). If GPAs differ between schools that have similar scores according to the Learning-to-learn test, it suggests that there are reliability problems with the GPAs.

The "Educational Equity Balance" is the term used for analyses seeking to explain variance in different Learning-to-Learn scores. The analytical ambition is to identify the factors that contribute to the various Learning-to-learn competencies. For example, in aggregate terms, findings so far suggest that about 10-15 per cent of the variance in competence scores can be explained by the school. The school factor is potentially relevant to policy makers. The more significant the school factor, the larger the significance of the activities that take place in school for learning-to-learn competencies. It may therefore be a political goal to increase the importance of the school factor. However, the level of the more general educational policy, there is little evidence so far that Learning-to-learn project in Finland has had a measurable national policy impact.

The impact is larger at more local levels. In particular in the City of Helsinki Educational Department, Learning-to-learn assessments have been used in the schools. Through comparisons of results between different schools of the city, school managers have been able to identify weak or strong aspects of their pupils' competencies, to compare these results with their self-assessment of the individual school, and potentially to adjust the contents of the education provided accordingly.

Time series data allow for different usages, for instance to identify trends in the development of particular competencies, for the total population of pupils in question or for sub-groups in terms of age, gender or other background variables.

### **Quality and relevance of skills definitions**

Skills, defined as competencies in connection with Learning-to-learn, have been theoretically defined in particular with reference to research literature on development psychology (Piaget 1985, 2001; Piaget & Garcia 1989), psychological competence models (Hirsh 1996; Nuthall 1999; Weinert 1999), and to previous work on educational assessment (Carroll 1993; Hautamäki A. & Hautamäki, J., 2001; Klauer 2000; Markman & Gentner 2001; Snow 1990, 1994; Spearitt 1996). Another source of inspiration for the identification and definition of relevant competencies are socio-culturally oriented approaches which analyse and make use of concepts such as "the learning society" and "the risk society" (Claxton 1988).

A second step in the definition of competencies has been empirical, as individual test items have been correlation tested for internal consistency. On this bases, more than 30 different belief scales have been identified, most of them consisting of no more than 3 items with an internal correlation of at least 0.7.

### **Validity of assessment methods**

Validity testing has primarily taken place as internal, structural consistency testing. As the learning-to-learn testing system has been developed over a number of years, the consistency testing has made it possible to reduce the number of items in relation to each scale, thereby also reducing the overall testing time required. Against this background, the level of coherence between theoretical and operational variables would seem to be rather high.

As for predictive validity, i.e., whether the operational variables can actually predict the outcomes which are expected at the theoretical level, this remains unclear. There is little data which can test predictive validity, since this test would require time series data for a long period. The first representative longitudinal data has been already collected, and the results are to be published in the future.

In connection with "learning-to-learn" competencies, the question of predictive validity would seem especially important. The initiative assesses a broad range of competencies, which are seen to be components in "the readiness and willingness to adapt to a novel tasks". This is appears to make the project particularly appealing in the context of a "life-long learning" and "knowledge society" agenda. If policy makers could be provided with precise information on their education systems' abilities to inject life-long-learning motivations and capacities into its citizens, systems could potentially be adjusted so as to increase the "life-long-learning" impetus of the educational system.

However, is the Learning-to-learn test module really measuring learning-to-learn motivations and capacities? Or is it for instance really measuring more general personal resources, be they cognitive, emotional or social? At least it seems entirely defensible to view categories such as for instance "self-concept", "self-evaluation", "future orientation" and "affective self-regulation" as elements in a broad concept of "personal resources" rather than just "willingness to adapt to novel tasks", not to mention some "lifelong learning" orientation. These personal resources may, of course, be very relevant for the motivations and capacities for further learning.

### **Reliability of assessment methods and the quality of data collection**

There is an elaborate set of guidelines in place for ensuring the reliability of the data collected, contained in the Instruction Booklet for persons supervising the implementation of the tests in

school classes. Standardised procedures are applied. Cheating does not seem to be possible. Risks of systematic biases as a consequence of data collection procedures seem low, and the reliability of the assessment methods appears high.

### **3. Lessons learned**

The Learning-to-learn project is directly relevant for a European adult skills assessment initiative to only a limited extent, and it points to certain limits and lessons as to the scope, contents and uses of a European skills assessment initiative.

#### **Issues of legitimacy**

The cognitive elements in the present test system accounts for more than half of the total testing time.<sup>23</sup> In some respects the competence scales resemble conventional intelligence tests - some critics of the Learning-to-learn project have, in private, considered the competence scales to be “old fashioned”, resting on an outdated or limited understanding of intelligence, or a too strong “Piagetian” framework. However, the system also includes scales for assessing reading comprehension, and understanding problems presented in mathematical notation, which resemble more school achievement oriented tasks, with some innovative elements. There are also scales for measuring the mastery of reflective abstraction and formal operational schemata, which have been shown to be teachable.

Nevertheless, the restrictions or cautiousness surrounding the uses of the Learning-to-learn initiative, specifically that Learning-to-learn competencies are mainly considered the dependent variable, points to the sensitivity and controversy potentially connected to large scale use of competence tests. This has been one of the major discussion topics with Swedish and Dutch field-tests. Thus, this approach is controversial in a school-based context where testing is a part of daily school life. How controversial would it be if a test that was widely seen as an intelligence test was part of an overall adult population skills assessment initiative?

The counter-argument in this connection would be that it is problematic to deliberately ignore factors which are known to have a strong positive correlation to school achievement and later learning-at-work. Abilities are important for “the readiness and willingness to adapt to novel tasks”, and if “learning-to-learn” is understood to consist of only beliefs, then something important may be lost which could help policy makers and educational reformers to understand the dynamics of modern education in the context of future needs of populations to engage themselves in more demanding re-and new-learning situations.

#### **The policy relevance of “learning to learn”-competencies**

Even if we accept that the Learning-to-learn initiative measures “the ability and willingness to adapt to novel tasks” and even capacities and motivations for lifelong learning, the analytical results of the initiative so far are sobering: Overall, only about 10-15 per cent of the variance in learning-to-learn scores can be explained by “the school factor”. The rest is explained by other factors, not least the parents’ educational background. Or to put it in other words: The differences between the

---

<sup>23</sup> However, there are tested modifications for sets with only 1-2 competence scales.

several thousand schools in Finland, their qualities and priorities, pedagogical principles etc. can only account for a very minor part of the differences in the pupils' motivations and abilities.

One implication appears to be that there are limits to the policy relevance of the overall set of competencies, scales and items contained in the Learning-to-learn initiative. It would seem that "self concepts", "self evaluations", "future orientation" and several of the other categories of the initiative are not primarily shaped in school, they are given shape elsewhere. Consequently, educational policies can only to a limited extent hope to affect them. Evidently, education policy is not irrelevant, but are there not other skills that are more directly affected by the formal education system and thus in the last instance by political priorities and decisions?

#### **Specific contributions in terms of design, items etc**

Could there be specific elements in the Learning-to-learn initiative that could contribute to a European adult skills assessment initiative? According to the manager of the project, Professor Jarrko Hautamäki, the "belief scales", the various sets of items which measure beliefs and motivations, could be adapted to other contexts, also for adult persons. They have the advantage of having been tested thoroughly, and internal consistency among the items of each scale is high, with a correlation of at least 0.7.

However, the reasoning tasks and other competence elements in the Learning-to-Learn test system are rather complex, and it is not easy to design suitable and economical versions of the competence scales for a household survey among adults. As for mathematics and literacy, the project manager himself assesses that the IALS/ALLS-methodologies are most likely more suitable approaches to an adult population.

## **4. References**

#### **Interviews conducted/persons contacted**

Professor Jarrko Hautamäki, Centre for Educational Assessment, University of Helsinki. Manager of the Learning-to-Learn project since 1996.

Irmeli Halinen, National Board of Education, Finland

Tuija Kirveskari, Director of Vocational Education, Helsinki City Educational Authority

Kirsti Mäensvisu, Chief Education Officer, Ph.D., City of Hämeenlinna

#### **Literature**

Carrol, J. (1993). *Human Cognitive Abilities*. Cambridge: Cambridge University Press.

Claxton, G. (1998). *Educational cultures for learning societies: a sociocultural perspective on the future of schooling*. Paper presented at the 4th ISCRAT Congress, Aarhus, Denmark, June, 1998.

Hautamäki, A. & Hautamäki, J. (2001). *Children's Self-Concept as Gendered and Contextual. Socio-Moral Self-Concepts Among 12-year-old Finnish Girls and Boys*. In Hedegaard, M. (Ed.),

Learning, Teaching and Knowledge Appropriation in Different Institutional Contexts. Aarhus: Aarhus University Press.

Hautamäki, J, et al. (2002). Assessing Learning-to-learn. A framework. Centre for Educational Assessment, Helsinki University in collaboration with the National Board of Education in Finland, Helsinki.

Hirsch, E.D. (1996). *The Schools We Need and Why We Don't Have Them*. New York: Doubleday.

Klauer, K.F. (2000). Das Huckepack-Theorem Asymmetrischen Strategietransfers. Ein Beitrag zur Trainings- und Transfertheorie. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie*, 32, 153-165.

Markman, A.B. & Gentner, D. (2001). Thinking. *Annual Review of Psychology*, 52, 223-47.

Nuthall, G. (1999). Learning how to learn: the evolution of students' minds through the social processes and culture of classroom. *International Journal of Educational Research*, 31, 139-257.

Piaget, J. (1985). *The Equilibration of Cognitive Structures: the Central Problem of Intellectual Development*. Chicago: Chicago University Press.

Piaget, J. (2001). *Studies in Reflecting Abstraction*. Hove: Psychology Press.

Piaget, J. & Garcia, R. (1989). *Psychogenesis and the History of Science*. New York: Columbia University Press.

Snow, R. (1990). New Approaches to Cognitive and Conative Assessment in Education. *International Journal of Educational Research*, 14, 455-473.

Snow, R. (1994). Abilities in Academic Tasks. In R.J. Sternberg & R.K. Wagner (Eds.), *Mind in Context. Interactionist Perspectives on Human Intelligence*. Cambridge: Cambridge University Press.

Spearrit, D. (1996) Carroll's model of cognitive abilities: educational implications. *International Journal of Educational Research*, 25, 2, 107-198.

Weinert, F.E. (1999). *Concepts of Competence*. DeSeCo paper.

# The Danish National Competence Account

## 1. Description

### Goals and ambitions

The Danish National Competence Account (NCA) was born out of the "knowledge society"/"lifelong learning" agenda. There was a certain euphoria in the late 1990s surrounding the concept of competencies and the idea of the knowledge society, and there was an idea that a range of key competencies were productive drivers in the knowledge economy. A Danish weekly, "Mandag Morgen" was particularly active in promoting the idea that a national competence account could be used as a tool to fine tune policies in the emerging knowledge/information society. The then Social Democratic government became convinced that this was a promising idea.

Against this background, the National Competence Account (NCA) officially aims to form a basis for identifying strengths and weaknesses in national competencies as a basis for policy initiatives, and for public debate concerning competencies in Denmark.

In the longer term, the account is intended act as a platform from which Denmark can compare its level of competence with that of other countries. The NCA has been developed as a part of a broader international effort, building directly on the skills definitions developed in the OECD's DeSeCo-initiative.<sup>24</sup> Certain modifications have, however, been carried out and considerable effort has been put into operationalising the theoretical concepts of the DeSeCo-framework.

The National Competence Account has been implemented as a 3-year project ending mid-2004, with the publication of main reports in September-October 2004. It is hoped that the competence account will form a template for competence accounts in subsequent years, but no decisions have been taken to this effect yet, and the present Danish government seems less interested in the project than the previous, focussing more on core competencies (literacy, numeracy, languages) than DeSeCo-s much more comprehensive list of key competencies.

The NCA is part of a broader project on the measurement of real competencies. The project is managed by a steering group consisting of nine representatives from three different ministries: Education, economy and employment.

### Competencies assessed

As mentioned, the National Competence Account builds on the DeSeCo definition of key competencies. Following this approach, it has been the starting point of the NCA at it is relevant to provide assessments of a broad spectrum of individual competencies, spanning from environmental and natural competencies, through physical competence, to social competence and learning competence.

Against this background, the initiative attempts to measure 10 key competencies: social competencies, literacy competencies, learning competencies, communicative competencies, self-management competencies, democratic competencies, ecological competencies, inter-cultural

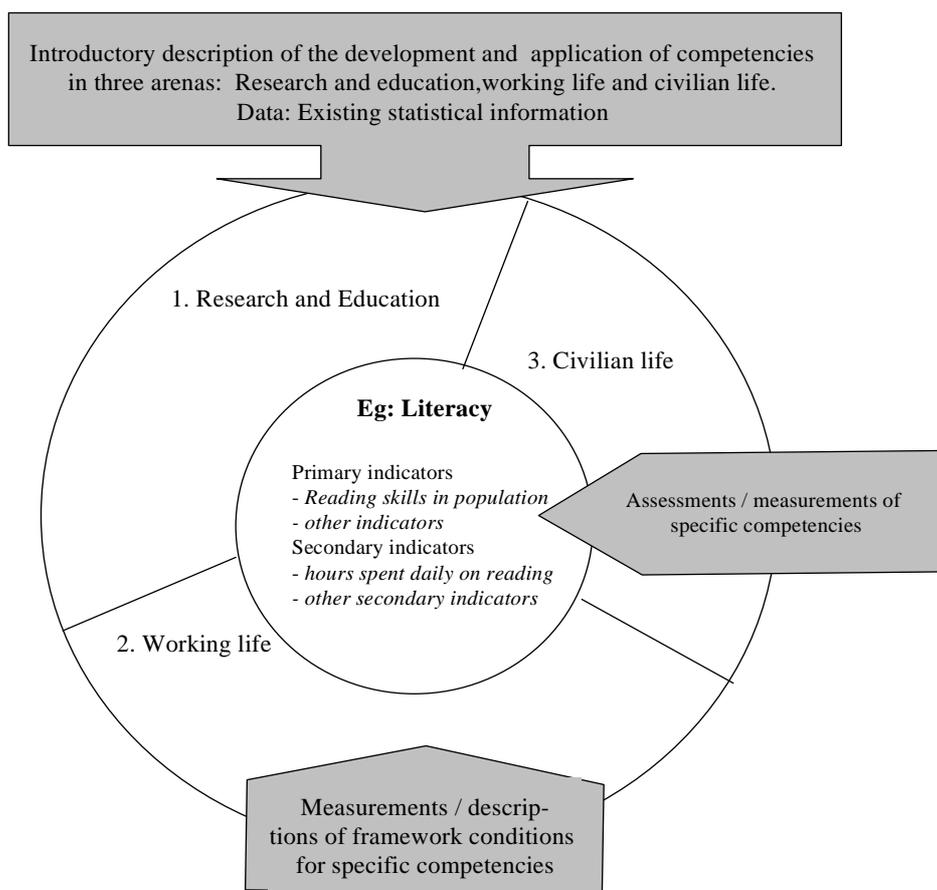
---

<sup>24</sup> [http://www.statistik.admin.ch/stat\\_ch/ber15/deseeco/](http://www.statistik.admin.ch/stat_ch/ber15/deseeco/)

competencies, physical- and health competencies and creative and innovative competencies. Compared to the original DeSeCo-framework, the Danish NCA substituted measuring creative and innovative competencies for the original DeSeCo-framework's "value" competence.

In the final National Competence Account, competencies will be expressed via specific quantitative and qualitative indicators. A number of indicators will be based on existing data, while others will need to be developed for the purpose, cf. Figure 1 below.

**Figure 1: Conceptual framework for the Danish National Competence Account**



**Applied Methods**

The Danish NCA thus makes use of both existing statistical information and survey information. Existing statistical information (e.g. input data from educational statistics) will be utilised to describe the framework conditions for developing and utilising competencies in society (the outer ring). Survey data will be utilised to develop indicators for individual competencies along the lines defined in the DeSeCo-framework (inner circle).

As for the survey data, the items of the questionnaire are intended to express in operational terms the theoretical DeSeCo definitions of competencies (with the above mentioned modifications). A group of Danish experts have been involved in this process of operationalisation. For each of the 10 competencies included in the survey, the expert group has developed proposals for definitions and

indicators, seeking to specify and operationalise the DeSeCo key competencies in a Danish context.<sup>25</sup> On this basis, the project staff has developed the specific questionnaire items.

A first comprehensive version of the questionnaire was pilot tested in mid-2003, where approximately 2000 telephone interviews were carried out. Subsequently, the number of items in the questionnaire was reduced significantly in order to decrease the time required to respond and to increase the response rate.

A large-scale survey was carried out in February 2004 with 5.500-6.000 responses from a random sample of the population aged 18 to 65 years. The response rate was about 75%.

The survey questions seek to capture relevant aspects of respondents'

- behaviour/actions
- motivation
- environment/surroundings

#### *Environment/surroundings*

As for environment/surroundings, the questionnaire captures both conventional socio-economic background variables such as among other things age, gender, the character of current employment, formal education, and income. A number of questions also seek to capture the context surrounding the use of specific types of skills, for instance how and in which situations knowledge obtained from learning activities are used in the daily work situation (in relation to learning competencies), or whether the closest superior in the job supports the respondent in the development of new ideas (in relation to creative/innovative competencies).

#### *Behaviour*

As for behaviour, a range of questions are asked in relation to training/education activities (learning competence), participation in development activities in the job (creative/innovative competence), the management and organisation of work in the job (self-management competence), reading and writing, in the mother tongue and in foreign languages, in connection with the job (literacy and communication competence) and the use of computers for different purposes (literacy and communication).

As for behaviour related to social competence, the items seek to describe the respondents' relations with colleagues, inside and outside the daily work place, as well as various types of social contacts outside working life (participation in voluntary adult education activities, board memberships, hobby activities and other leisure time activities, and involvement in "local community activities"). Behaviour related to "intercultural competence" includes professional and personal contact with persons residing abroad or originally coming from other countries, relations with ethnic non-Danes in the workplace, and stays abroad. Behaviour related to "democratic competence" in turn includes participation in meetings and other activities with a view to influencing general conditions in the workplace or as regards local or national political questions. Activities to inform oneself about society and politics through newspapers, radio or TV are also counted as expressions of behaviour related to democratic competence.

---

<sup>25</sup> <http://pub.uvm.dk/2002/nkr/dokumentationsrapport/>

Behaviour related to ecological competencies includes consumer behaviour (the extent to which the respondent purchases biodynamic foods, the degree to which the respondent acts in energy-conserving manners in different respects) as well as knowledge- and information seeking behaviour in relation to environmental or ecological questions.

As regards behaviour related to physical- and health competence, the questionnaire among other things seeks to capture the respondents' behaviour as regards reducing physical and mental stress in the workplace, nutrition behaviour, smoking and drinking behaviour and sports/fitness activities.

#### *Motivation*

The majority of the questionnaire items are behavioural in the above respects, asking the respondents' to describe their behaviour in a number of respects. However, a minor number of questions are directed towards the motivation of the respondents, e.g. whether respondents have been satisfied with the size of their continuing education activities during the past 12 months (learning competence), whether the respondent has considered to engage in formal education activities in fields where informal competencies have been acquired (learning), how important it is for the respondent to develop new ideas in the workplace (creative/innovative competence), or for which reasons the respondent makes use of re-cycling schemes.

#### *Self-reported competencies*

Finally, in a number of questions respondents are asked to report indirectly on their own competencies. This is so in particular as regards literacy and communication competence. In these fields, the respondents are asked to describe how often the respondent's job requires the use of information contained in for instance reports, articles, magazines or journals. Subsequently, the respondent is asked to assess the level of difficulty in reading the mentioned material.

The same methodology is applied in relation to writing skills. As for mathematical literacy (numeracy) and IT skills (both defined here as elements in literacy and communication competence), the respondents are asked to assess the level of skills in relation to the requirements of each respondents job (or latest job for unemployed).

### **Results**

The results of data collection are being analysed during the Spring and Summer 2004. Final reports are due September-October 2004. However, it seems clear that the management of the project has sought to scale down expectations to the results of the project significantly since it was originally launched. The original intentions of directly providing a knowledge base upon which policy decisions could be made in the field of competencies, have been replaced by a much more modest ambition.

Project management now views the NCA more as a heuristic device, as a source of inspiration and a tool to generate debate. The Competence Account is presented as an experimental learning exercise and a first step towards output measurement.

## **2. Evaluation**

### **Utility/usefulness for policy development or other purposes**

As mentioned, the fundamental ideal underlying the initiative was that a national competence account could be used as a tool to fine tune policies in the emerging knowledge/information society.

According to key staff at the project, this idea has, however, turned out to be quite unrealistic. Data will not allow the deduction of specific policy responses. At the time of our interview with the project staff, it was struggling to identify some "read thread" or predominant theme in the data, which would allow it to draw some interesting and relevant conclusions at all. Furthermore, there are many unresolved methodological concerns, cf. the sections below on the quality and relevance of skills definitions and the validity of assessment methods.

Finally, the project has until now been carried out in isolation in the Ministry of Education. No results have yet been produced that could potentially inform policy making, and there has thus been rather modest interest on behalf of other ministries.

Still, in the views of NCA project management there continues to be a need to gradually develop instruments which can contribute to the measurement of outcomes of the Danish investment in education and human resources. Denmark presently spends more than 8 per cent of GDP on education, and there is a need to assess and establish the outcomes of this investment, and to provide a basis for decisions on where to increase or decrease investment. The management of the NCA would like to see the initiative as one among several different first steps in this direction.

### **Quality and relevance of skills definitions**

The NCA rests upon the theoretical skills definitions of the DeSeCo-project. In the project management scepticism was, however, expressed as regards the status of the DeSeCo list of key competencies. In this view, there are no positive foundations for this particular list, for including these particular competencies and excluding others.

Rather, they were seen to reflect certain political priorities predominant in the late 1990s and early 2000s and reflecting a social democratic consensus across much of Europe that a comprehensive list of competencies were relevant as productive drivers in the knowledge economy.

The NCA-project itself appears to have carried out substantial work in relation to refining and operationalising the DeSeCo-competencies into indicators and further into questionnaire items. The expert contributions to this effect seem relevant in most respect, although a number of the contributions would seem to fail to go far enough in the direction of operationalising theoretical definitions into empirical indicators. This is so for instance in relation to communication competence and democratic competence. Perhaps not least as a consequence of this, the connection between the experts' development of indicators and the items contained in the questionnaire thus seems rather unclear for several of the competencies covered by the NCA.

### **Validity of assessment methods**

In this connection, questions regarding the validity of the assessment methods can be raised. As a general approach, the DeSeCo-competencies were operationalised into indicators which were then further operationalised into items in the questionnaire. Now that data has been collected, the

different items, that were defined as an operationalisation of an indicator, are being correlation-tested (Crombach's Alpha) in order to ensure that they measure the same underlying dimension. Some indicators fall apart in the correlation analysis. An attempt is being made to establish new indicators on the basis of correlation analysis.

As for conceptual or predictive validity, it is an open question whether many of the indicators and items that have been defined, really measure the competence/skills they are intended to measure. Take for instance some of the literacy items: Is it really a valid measure for reading skills that a person is asked to assess the level of difficulty of reading texts that are relevant in his or her job? The level of difficulty of these texts may, after all, vary quite significantly from job to job and from function to function, regardless of whether the type of text is the same, for instance "reports, manuals, articles or journals".

Or to take another example: Is the frequency of situations where the respondent has experienced that "his/her oral messages have been understood" by different groups of persons (colleagues, customers, really a valid measure of communicative competence? In this last instance, the question of the reference group is crucial: The assessment of oral communicative competence would seem to be a highly dependent on the context in which the assessment takes place (is a difficult or a complex message being communicated? If the message is understood or misunderstood, is it because of the communication skills of the person uttering the message or because of the skills of the persons understanding or misunderstanding the message?). However, in the questionnaire an implicit and highly questionable assumption is made that oral communication is not context dependent in these ways, and that "colleagues" or the other groups mentioned constitute identical groups with identical communication skills, regardless of context.

A third area in which validity would seem problematic is in connection with inter-cultural competencies. It could thus be argued that several items of the questionnaire refer to a theoretical concept of "tolerance" rather than intercultural competence.

In sum, there would seem to be a number of concerns regarding the conceptual validity of the NCAs survey methodology.

#### **Reliability of assessment methods**

Some items in the survey also fulfil a function as reliability checks. Thus, a high Crombach-Alpha suggest that people have answered consistently. No other reliability checks are carried out.

#### **The quality of data collection**

Data collection was carried out by the National Statistical Office, Danmarks Statistik. This institution has a good reputation for delivering quality data and for carrying out data collection in accordance with an elaborate framework of guidelines.

### **3. Lessons learned**

#### **The need for skills assessment**

According to the management of the NCA, an assessment of the return on investment in human capital is necessary. A European interim report on the Lisbon Strategy is being developed (Kok),

seeking to measure steps that are taken to achieve Lisbon. In this light, there is a need for a European skills assessment initiative.

### **The limits of the Danish approach: Ethical concerns**

However, the Danish strategy in connection with the NCA also directs attention to a number of ethical and political concerns in connection with adult skills assessment. Thus, there are clearly a number of questions on *limits to government* that should be explicitly addressed in a future European strategy.

Thus, along with the overall DeSeCo-framework, the NCA harbours great ambitions on behalf of governments. Not only are governments to provide citizens with necessary basic skills. They are to ensure "Key Competencies for a Successful Life and Well-Functioning Society" to quote the title of the main DeSeCo publication (Rychen and Salganik 2003).

This implies that there is some far-reaching public responsibility for the functioning of the population, far-reaching at least in terms of the different types of competencies which are of public concern. In the lenses of the National Competence Account, the concern is not only that the Danish population should be an internationally competitive workforce. It should also be a population who is environmentally conscious and democratically active, who leads a healthy life style and has abilities to communicate and cooperate with individuals from different cultural backgrounds.

This is, of course, a possible political position. But in the context of a future European skills assessment strategy, it seems necessary to highlight that this ambition is in many respects highly illiberal, potentially implying interference in individuals' personality and personal life, turning individual choices and preferences as regards life style habits and cultural activities into a government concern. This may be perceived as legitimate ambitions in a small, tightly-knit nation state where welfare state ambitions already imply government interference in many aspects of personal life. But is it a possible level of ambition in other EU Member States not to mention in the context of an EU wide adult skills assessment initiative?

### **Policy relevance**

A different point in this connection concerns the policy relevance or lack of the same the 10 key competencies covered by the Danish NCA. The project is currently working hard to identify relevant findings and develop relevant conclusions on the basis of the comprehensive data set. This seems indicative that it is difficult to point to implications of much of the collected information which invite specific policy responses.

This problem may partly be resolved if the NCA survey were to be repeated at a later point, allowing comparisons over time and the identification of trends in the population's behaviour in different respects. However, for some of the 10 competencies, for instance "self-management competence" and "social competence", it is questionable whether specific policy initiatives could even in principle address such competencies, within the formal education system or elsewhere. And at any rate, seeking to redress via public policy competencies which seem rather tightly connected to the personality of each individual would seem politically controversial in many contexts.

### **Methodological challenges**

A different set of "lessons learned" concern the methodological challenges confronted by the Danish NCA. The behaviour related assessment would seem to be relevant in connection with some skills, for instance communication (in particular IT) skills and literacy skills, and also in some connections in relation to democratic competencies and learning competencies. However, as for many of the other competencies, the experience of the Danish NCA would seem to point to the limits of a survey-based self-report methodology. There are clearly limits as to which and how many competencies one can assess via surveys in which individuals describe their activities.

## **4. References**

### **Interviews and persons contacted**

Stefan Hermann, project manager, the Danish National Competence Account  
Peter Høier, Head of Unit, Danish Ministry of Education

### **Literature**

Rychen, Dominique Simone and Laura Hersch Salganik, (eds.), Key Competencies for a Successful Life and a Well-Functioning Society. Göttingen: Hogrefe & Huber 2003.

Undervisningsministeriet, Nøglekompetencer – forskerbidrag til Det Nationale Kompetenceregnskab. Dokumentationsrapport fra Det Nationale Kompetenceregnskab, 2002.

Undervisningsministeriet, Projektintroduktion, Det Nationale Kompetenceregnskab, juni 2002.

Undervisningsministeriet, NKR projektgruppen. Notat om revideret tilgang til udvikling af det Nationale Kompetenceregnskab, januar 2002.

# Test of Workplace Essential Skills (TOWES)

## 1. Project description

### Goals and Ambition

TOWES (Test of Workplace Essential Skills) is a test of workplace essential skills that is being developed in Canada. Its purpose is to assist individuals, employers, unions, educators, training providers, and policy makers to assess literacy skills in Canadian workplaces. The TOWES project now has a bank of test items, which, collectively, form an assessment tool to measure essential skills in workplace settings.

Planning has been a collective effort, commencing in March 1997, involving Statistics Canada, HRDC, the National Literacy Secretariat (NLS) Bow Valley College, and SkillPlan (British Columbia Construction Industry Skills Improvement Council). In 1998, the project received funding from the NLS, and the TOWES project was jointly carried out by SkillPlan and Bow Valley College. As of April 2004, Bow Valley College became the sole owner of TOWES and distributes.

The test can be used in different ways and the developers feel that other uses will evolve as “TOWES takes its place in industry and education.” The developers envision the following following purposes ([www.towes.com](http://www.towes.com)):

- **“Entry-Level Assessment of Skills** - TOWES gives employers a way to set reasonable and valid entry-level job requirements. Many employers have been using years-in-school or other credentials as proxies for ability. This is unfair to individuals, some of them foreign-born, who have the skills but not the educational credentials in Canada.
- **School/Work Transition Programs** - A valid and reliable test of the essential skills needed for work is useful to high-school teachers and administrators. TOWES provides a basis for certifying students’ levels of competence, and a way to promote ‘employability skills’ to teachers and students alike
- **Educational Assessment** - Instructors, trainers, and literacy tutors can use the information from TOWES to place workers in their programs and to design programs to meet individual educational needs.
- **Work/Worker Adjustment** - Rapid changes in technology and work processes, together with plant closures and down-sizing, have created the need for effective adjustment programs. A wide-ranging test, keyed to essential skills descriptions for jobs listed in the National Occupational Classification gives workers and program developers a valid way to assess present skills and compare them to the requirements for a variety of possible jobs. Individual workers can use TOWES results to make decisions about future training.
- **Developing National Standards** - An essential skills assessment tool may be used to set standards and specify competence. It allows companies to assess workforce skills and make comparisons with other groups of workers”

### Competencies assessed

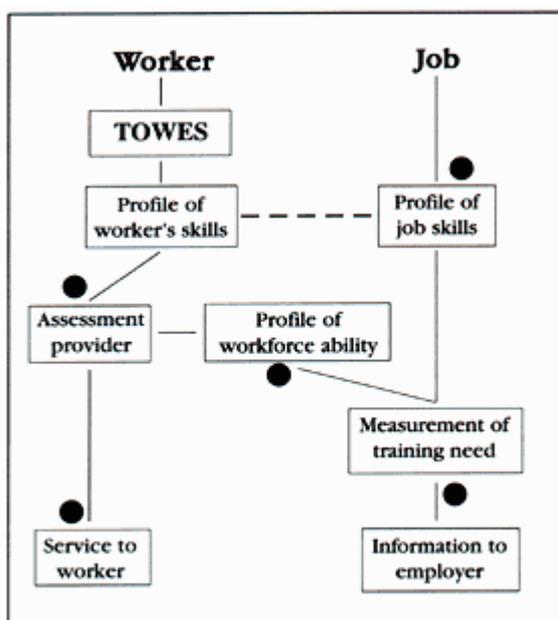
TOWES involves direct assessment of essential skill competencies in the areas of reading text, document use, and numeracy.

Reading text is defined as the ability to understand and use information contained in prose passages. Document use involves the skills and knowledge needed to understand and use information from documents such as tables, catalogs, maps and scale drawings. Numeracy is the ability to understand and use numerical information embedded in print.

TOWES test results are correlated to the 5-point scale used in the International Adult Literacy Survey (IALS) and by Human Resources Development Canada (HRDC) to determine the complexity or difficulty of tasks associated with specific occupations. Ratings of complexity are from 1 to 5, with 5 being most complex.

TOWES allows workers' skills to be tested in such a way that the results can be compared to the requirements of the job as described in an Essential Skills Profile. The profiles have been established through HRCD's Essential Skills Research Project. To date, profiles for 200 occupations are available (more in process). For each occupation, the profile provides a description and a list of the most important essential skills for that occupation. The essential skills include: reading, document use, writing, numeracy, oral communication, thinking skills, working with others, computer use, continuous learning, plus other information (e.g. physical aspects of the job). The skill profiles were based on over 3000 interviews with people in over 180 occupations. These jobs are also linked to the Canadian National Occupational Skill Standards.

The following figure illustrates the linkage between TOWES assessments and job requirements, and indicates different possible uses of the assessment.



### TOWES Methodology

TOWES employs a direct assessment paper and pencil test, using items, definitions, and scales similar to IALS. TOWES uses authentic documents – such as catalogues, order forms, labels, and schematics – as source material. Questions range in difficulty and mimic actual workplace tasks by having the test taker assume the role of a worker who needs to use information embedded in

documents. The bank of items may be used to create a variety of custom-designed tests for specific contexts.

In October 2000, the TOWES development team began implementing the TOWES-IALS Linking Study, a field test of over 2,600 test-takers at sites across Canada, to establish TOWES as a reliable and valid tool. Test items from the TOWES test bank were combined with test items from the International Adult Literacy Survey (IALS) in a variety of configurations to produce 15 different test booklets. These booklets were then administered to test-takers from a cross-section of the Canadian public aged 16 to 65 with a variety of educational backgrounds. Linking TOWES to IALS' international standard – test items from IALS were written by thousands of individuals in 20 countries – provided the development team with statistical information required to better interpret TOWES scores.

There is no target population as such. The test is available for use by assessment service providers such as community colleges, and, where appropriate, workplace literacy consultants. When test security can be assured and the purpose for administering the test is clear, it is also made available to Human Resource professionals in individual businesses. Guidelines are provided to administer tests under standard conditions.

The test-takers themselves receive a one-page individual report on their results. Program leaders can use this report to motivate upgrading and to guide realistic career planning. Individual reports can be customized to best suit the needs of the user group. Summary reports are also provided that give scores for a group of test takers.

### **Results**

The project is ongoing. Four national demonstration projects were initiated in 2000. In Calgary, TOWES was modified to see if it fit the assessment needs for entry-level workers in manufacturing. In the Northwest Territories special test items were developed to extend TOWES' ability to assess the skills of special populations such as Aboriginal workers. TOWES staff also ventured onto the long shore in Atlantic Canada. Here TOWES' role was to provide an entry-level standard for essential skills that employers, union, and workers could all agree was fair. On the prairies, where giant, concrete silos are replacing local, small-town elevators, TOWES was used to determine employee skill levels so that appropriate training and placement services could be offered during industry restructuring.

## **2. Project evaluation**

### **Utility/usefulness for policy development or other purposes**

In theory, the TOWES can be used by a number of different individuals/institutions for multiple purposes. For example, test scores may be used by workplace educators to provide service to workers in the form of individual educational counseling or direct instruction. Individual scores can be combined to generate a profile of a particular workforce and of job categories within the workforce. This statistical information can be given to management and used to set training policy and to answer questions about current and future training at a corporate level. However, we have not yet ascertained the full extent of TOWES activities that have taken place.

It seems that the assessment could be used in other countries, although it would require further development work in those countries to be used in the same manner (e.g., to link TOWES tests to in-country occupational standards).

The initiative is explicitly linked to IALS, as discussed above, so results are interpretable along the three main IALS dimensions. Bow Valley College supplies the test nationally through 33 college distributors. This is a commercial venture. Thus it is not clear if it is possible to use this assessment in the public domain.

#### **Quality and relevance of skills definitions**

The TOWES adopts IALS skill definitions and assessment method. The difference is that the reading, document and literacy tasks are specifically drawn from and tailored to different occupations and benchmarked to national occupational skill profiles.

#### **Validity of assessment methods**

A series of analyses were conducted to examine the internal consistency of TOWES items. Specifically, internal consistency, item-to-total correlations, and confirmatory factor analyses were carried out. Using Cronbach's alpha the resulting internal consistencies were high: 0.82 for Reading Text, 0.88 for Document Use, and 0.84 for Numeracy.

Twenty-seven of the 304 items (9%) showed lower than expected corrected item-to-total correlations with their respective subscales. The findings from these more fine-grained analyses will assist the TOWES team to determine which of the items from this data set should be removed or revised in the future.

Confirmatory Factor Analyses (CFA) showed that 9 of the 304 items had lower than expected factor loadings on their respective subscales. In addition, the CFAs revealed that the subscales are highly correlated with one another. Finally, a comparison of the proportion of individuals who made a correct response on the item with the items that had been dropped from an IRT analysis conducted earlier on this same data set was examined. It was concluded that about half of the items that had been dropped were very difficult. Some of the items that were dropped were the same ones that showed problems of internal consistency in the present sets of analyses

Overall, the internal consistencies reported in the analyses are respectable. The other analyses are also respectable, and available documentation indicates the data will be used to improve the tests.

#### **Reliability of assessment methods**

A series of analyses have been carried out to assess item discrimination and difficulty. These indicated that TOWES items have similar characteristics to IALS items. (Yamamoto and Kirsch, 2002).

Reliability studies were carried out by re-scoring a subset of tests (Yamamoto and Kirsch, 2002). Two hundred responses to each TOWES and IALS item were re-scored. The proportion of agreement was .96 and .97, respectively, indicating a high degree of reliability. Analyses of variance were used to check test-order effects. None were found.

Sampling is not really an issue, as the assessments are used for specific purposes, as discussed earlier. The data are not used to establish national norms, etc.

### **The quality of data collection**

TOWES includes a users agreement that outlines testing standards for users in education institutions and workplaces. Users are asked to read and sign the document, which discusses such issues as how to motivate the test takers and the conditions under which the test should be taken; how test scores can be used; how to report results to test takers, etc.(see [http://www.towes.com/pdfs/Users\\_Guide.pdf](http://www.towes.com/pdfs/Users_Guide.pdf)). However, as the test administration is left to the user groups it is difficult to know the extent to which these guidelines and standards are followed.

### **3. Lessons Learned**

The most interesting aspect of this initiative is the extension of IALS-type performance tests to assess skills as related to specific occupations. This linking between the assessments and the national occupational standards in Canada could potentially be of great interest as a research tool, for example in studies that attempt to understand the links between skills and productivity and characteristics of the firm.

It is difficult to draw lessons for a European strategy, however, as there is not much documentation available on this initiative, apart from a few academic studies on the properties of the assessment itself. The website does not provide information as to costs of the assessment, nor on the extent of its use in Canada.

### **4. References**

Kline, TJB. *Internal Consistency of the Test of Workplace Essential Skills (TOWES) Items Based on the Linking Study Data Set* (available at [www.towes.com](http://www.towes.com))

Yamamoto, K. and Kirsch, I. (March 2002). *Scaling TOWES and Linking to IALS*. (available at [www.towes.com](http://www.towes.com))

Introducing essential skills and TOWES: G series. (available at [www.towes.com](http://www.towes.com))

TOWES User's Guide (2001). BC Construction Industry Skills Improvement Council (Skillplan) and Bow Valley College.

<http://www.towes.co>

<http://www15.hrdc-drhc.gc.ca>