



The Role of Self-Assessment in Measuring Skills

REFLEX Working paper 2

March 2005

Jim Allen
Rolf van der Velden

Research Centre for Education and the Labour Market
Maastricht University
The Netherlands
e-mail: J.Allen@ROA.Unimaas.nl or R.vanderVelden@ROA.Unimaas.nl

The REFLEX project is funded by the EU 6th Framework Program (Contract No: CIT2-CT-2004-506-352) and several national funds. The project involves partners from sixteen countries (Austria, Belgium/Flanders, Czech Republic, Estonia, Finland, France, Germany, Italy, Japan, the Netherlands, Norway, Portugal, Spain, Sweden, Switzerland and the UK) and is coordinated by the Research Centre for Education and the Labour Market from Maastricht University. For more information, see: <http://www.reflexproject.org>.

Abstract

The last few decades have seen an increased awareness of human capital as one of the driving forces in economic development. This heightened interest in education and learning has been accompanied by a greater need to monitor and assess the stock of human capital. Since the 1990's several large-scale international surveys have been undertaken to measure skills. Most skills research focussed on aspects in the cognitive domain such as literacy and numeracy skills. Notwithstanding the importance of these skills for dealing with the complexities of today's world, they nevertheless represent only a fraction of the skills and competencies that are assumed to be 'key'. In a way we could say that skills researchers find themselves confronted by the limitations of classical methods of assessing skills, such as large scale testing. In this paper we have developed a plea for using self-assessments as a complementary tool to assess skills. We give an overview of different methods that are being used to assess the stock of skills and the skills required by employers. Finally we discuss the advantages as well as some of the problems arising in the use of self-assessment.

Introduction

The last few decades have seen an increased awareness of human capital as one of the driving forces of economic development. Grounded firmly in economic theory and empirical research about the individual and the social returns to education (Schultz, 1961; Becker, 1964; Psacharopoulos, 1973), different actors in society (policy makers, employers, students, employees) have realised the importance of investing in education and training as a way of improving the existing stock of skills. The fact that human capital is high on the policy agenda of national governments and international organisations can be noted in the titles of documents as "*A Nation at Risk: The Imperative for Educational Reform*" from the US National Commission on Excellence in Education (1983), the EC White Paper "*Teaching and Learning: Towards the Learning Society*" (European Commission, 1995), "*Skills for All: Proposals for a National Skills Agenda*", from the British National Skills Task Force (2000), or the World Bank's "*Lifelong Learning in the Global Knowledge Economy: Challenges for Developing Countries*" (World Bank, 2002).

A development accompanying this increased interest in education and learning was the need to monitor and assess the stock of human capital. During the 1980s the OECD started the large-scale project Indicators of Education Systems (INES), aimed to develop indicators of the input, process and output of education and training (OECD, 1994). The results of this project are published annually in the authoritative publication "Education at a Glance". Similar projects have been undertaken at national levels as well (e.g. the National Competence Account of Denmark).

What soon became clear is that education as such is only a poor indicator of the stock of human capital. Individuals with the same nominal level and type of education can differ markedly in their command of various skills. Countries that have more or less comparable levels of educational attainment can nevertheless differ substantially in the level of skills that are acquired in education. Processes of appreciation and depreciation of skills over the life course exacerbate this problem, often changing the stock of human capital completely and further loosening the link between educational qualifications, which are usually acquired at an early age, and the later stock of skills. The Programme for International Student Assessment (PISA) carried out in 41 countries under the auspices of the OECD, made quite clear that there is no one-to-one relation between a country's average level of educational attainment and its level of skills (OECD, 2004).¹

As the emphasis shifted from educational qualifications towards skill measurement, the question naturally arose what skills should be measured. Looking at what people need in order to perform even very basic things in life soon makes clear that the list of skills that can possibly be measured is practically unlimited. The sheer breadth and variety of skills that individuals draw on in performing at work and in life in general presents a major potential stumbling block for researchers attempting to take stock of the available human capital. The problem is even greater when we take into account that people not only make use of generic skills such as communication skills or learning abilities, but also a large number of highly specific skills pertaining to a particular task, situation or object.

¹ PISA comprises tests administered among 15-year-olds in the subject areas of math, science and problem solving. The disappointing results of the PISA survey in Germany gave rise to considerable unrest among policy makers and educationalists.

In order to introduce some order in understanding the diversity of human skills, many scholars have engaged in a quest for so-called core skills, sometimes called basic skills² or key competences. The term (*Schlüsselqualifikation*) was first introduced by Mertens (1974) to denote those skills that have a more permanent value in a time where specific skills may quickly be rendered obsolete and where workers need to be more flexible (for a description of the history of the concept of key qualifications, see Nijhof, 1998). Although this conceptualisation obviously does not in itself alter the complexity of the full range of human skills - which after all exists in the real world - it can to some extent be seen as a way of establishing priorities as to which particular skills should be measured. That said, just like the quest for the holy grail, the quest for key competences has proved to be a difficult undertaking. Numerous lists can be found in the literature that purport to more or less sum up the most important skills. To give some examples:

- The Secretary's Commission on the Achievement of Necessary Skills (SCANS, 1990) produced an influential report on the required skills needed to make a transition from a Fordist-type organisation to 'high performance' workplaces. Based on job analyses, literature reviews and consultation of experts, they identified five competence areas (resources, interpersonal skills, information, systems, and technology) and three foundational areas (basic skills, thinking skills, and personal qualities).
- The National Competence Account of Denmark (2002) distinguishes 10 key competencies: social competencies, literacy competencies, learning competencies, communicative competencies, self-management competencies, democratic competencies, ecological competencies, cultural competencies, health sport and physical competencies, and finally creative and innovative competencies.
- The Australian National Training Authority (2003) identifies the following types of skills: basic/fundamental skills, people related skills, personal skills and attributes, skills related to the business world, skills related to the community and conceptual/thinking skills.

Although most lists that can be found in the literature have much in common, they usually lack thorough theoretical foundations. A welcome exception to this is formed by the DeSeCo (Definition and Selection of Competencies) project. This project was initiated by the OECD to provide an overarching framework to international skills assessments, and the main results are contained in the report "Key Competencies for a Successful Life and a Well-functioning Society" (Rychen & Salganik, 2003a). Emphasising the need for competence assessment rather than a narrow focus on skills, competencies are defined in this project as: "the ability to successfully meet complex demands in a particular context through the mobilization of psychosocial prerequisites (including both cognitive and non-cognitive aspects)" (Rychen & Salganik, 2003b, p. 43). The basic difference with the earlier concepts of skills (however poorly they have been defined) is the holistic nature of the concept of competence. It refers not only to a range of cognitive and non-cognitive skills and other prerequisites that need to be in place in order to perform in a competent way, but it also refers to the notion of 'orchestration', the ability to use these constituent elements in a meaningful and deliberately arranged way. In that regard, the 'whole' that makes up a competence is more than just the 'sum of its parts'. Skills can therefore best be considered as one of the constituent elements of a competence.³

Based on theoretical reviews, consultations with experts as well as the identifications of important policy goals, the project identifies three categories of key competencies that require a reflective approach and critical stance (including meta-cognitive skills, see Rychen, 2003). These three categories are:

- interacting in socially heterogeneous groups (encompassing relating well to others, cooperating and managing and resolving conflict);
- acting autonomously (encompassing acting within the big picture or the larger context; forming and conducting life plans and personal projects, and defending and asserting one's rights, limits and needs) and
- using tools interactively (encompassing using language, symbols and text; using knowledge and information; and using technology).

The transversal feature cutting across these three categories of key competencies is reflectivity, the ability to make independent judgments and take responsibility associated with higher levels of mental

² The term basic skills is also used to denote the skills at a lower level or skills that are themselves a basis for developing other skills (such as basic reading and writing).

³ Given this definition, most assessments measure skills rather than competencies. We shall therefore mainly use the term skills in this paper and restrict the use of the term competencies to indicate conditions that meet the definition given above. The term skills is used as a shorthand for the individual components of competencies, such as knowledge, attitudes, motivations, etc.

complexity. It is this ability which makes the difference between just interacting with others and interacting in a *socially heterogeneous* group, the difference between just acting on one's own and acting *autonomously*, or the difference between just using tools and using tools *interactively*.

Although the theoretical framework provided by the DeSeCo project injects some welcome theoretical rigour into the discussion of skills measurement, it does not in itself directly give rise to clear recommendations as to the competencies to be measured. The best way to conceive of this overarching framework is that it indicates the main underlying competencies that render skills their significance. Many of the key skills that are identified in the literature fall under the heading of 'tools': numerical skills, literacy, computer skills are typical examples of tools. As outlined above, there are literally hundreds of these skills. The major problem encountered in skills assessments is therefore the need to focus on just some of these skills, as a result of limitations in time and resources. In practice therefore, the question what skills are measured is likely to be a result of practical considerations as well as theoretical notions as Weinert (2001) put it.

As a consequence of these practical considerations, attempts to measure skills on a large scale have deliberately focused on particular aspects of skills rather than attempting to measure the full range of skills. For example, the International Adult Literacy Survey (IALS) and its successor the Adult Literacy and Life Skills Survey (ALL) concentrated primarily on literacy and numeracy skills. Given the enormous importance of basic literacy and numeracy for a large range of challenges people are faced with in today's world, these and similar studies constitute hugely significant milestones in terms of coming to grips with the worldwide stock of human capital. That said, there are clearly other major areas of skills that merit attention. There is a clear need for a broadening of the scope of large-scale skill surveys, in order that they better reflect the range of competencies that are needed for, as DeSeCo put it, "a successful life and a well-functioning society". This may refer especially to many of the so-called 'soft skills', social and cultural competencies that are widely recognised as being very important.

At the same time, the available research methodology may form a potential barrier to such a broadening of scope. One of the reasons why existing studies focus so strongly on aspects like literacy and numeracy is that such skills are relatively well-defined and accessible to measurement under controlled conditions, while many components of key competencies that belong to the non-cognitive domain are conceptually more diffuse and more difficult to measure.⁴ Another obstacle in many existing skills assessments is the limitation in time that can be spent on testing. Even in the imaginary case of unlimited resources, there is a limit to what subjects in any assessment can endure. This would automatically imply that only a small fraction of such skills can be tested in a classical assessment.

Although we recognise the potential problems involved in assessing skills in the non-cognitive domain, we will argue in this paper that this does not have to mean that such skills should be overlooked in skills research. We will develop a plea for a judicious use of self-assessments, in addition to assessment and testing of individuals, as a way of developing indicators of the full range of competencies that may be needed in the complex world of today. We shall first reflect briefly on what aspects of skills we need to measure. Following that, we give an overview of different methods that are used to assess the stock of skills possessed. We then give an overview of methods used to assess skill requirements. Subsequently we provide the main arguments in favour of using self-assessments as a way of measuring both acquired and required skills. Finally we will discuss some methodological problems relating to the method of self-assessment, and propose some strategies that can be used to eliminate or at least reduce these problems.

What do we need to know about skills?

Before entering into a discussion of the different methods that may be used to measure skills, it is important to reflect briefly on what we would actually like to know about various skills. At first sight, this seems a strange, and perhaps even trivial, question. If we take numeracy and literacy skills as an example, existing tests are designed to gauge the degree of proficiency of different groups of subjects in dealing with numbers and mathematics on one hand and language and communication on the other. Surely an assessment of all skills should be directed at the same basic aim, that is at assessing

⁴ Although attempts have been made in these domains as well, e.g. cross-curricular competences (Peschar, 2001)

the degree to which individuals possess the skill in question. While it is evident that an assessment of skill levels should indeed be one of the primary targets of skills research, we argue that this is not enough. Particularly when venturing into the relatively uncharted waters of the measurement of 'soft skills', it makes good sense to supplement the measurement of the actual skills possessed by individuals with a measure of the extent to which these same skills are *required* in work or in daily life. From the point of view of education and training policy, the obvious question that needs to be asked when studying the results of skills research is when it can be said that there is *enough* of a given skill in the population. This is no trivial problem since regardless of the method used, the units of measurement used are likely to have been constructed by the researchers themselves, with little or no directly observable counterpart in the real world. Both required and acquired skills are significant in their own right, and it is important to monitor the development of skill requirements as well as the development in the stock of skills in the population. However, both aspects draw deeper significance in relation to each other, by providing a mutual frame of reference. In determining training needs for example, one can decide to invest in people whose skills have a low absolute level, but a more practical option is to focus on those groups whose skills fall short compared to what is required. Alternatively, a comparison of required and acquired skills levels may reveal instances where skills in the labour force are being underutilised. Given the fact that human capital is the driving force in determining productivity, any mismatches between actual and required competencies can be regarded as being less than optimal, both from the point of view of the individual employee and that of his or her employer (Sattinger, 1993; Hartog, 2000).

Even when required and acquired skills are in balance in the aggregate, major mismatches may occur at an individual level. It is therefore essential to measure both skills requirements *and* acquired skills at an individual level, using the same scale. Although there are a range of dimensions of skills and skill requirements that have received attention in the literature, most of these do not lend themselves to measuring both acquired and required skills. For example, dimensions such as importance pertain to skill requirements, but cannot easily be applied to acquired skills. Dimensions such as frequency of use are indirectly related to acquired skills as well as requirements, in the sense that one presumably cannot use skills one doesn't have, but this cannot form the basis for an independent measure of both acquired and required skills. Probably the only dimension that can serve this double purpose is skill level. One can measure the level of skill actually possessed by individuals using the same basic yardstick as is used to measure the level required to perform adequately in a given situation. Consequently, it is advisable that research into skills in a given population incorporates measures of the acquired and required level of different skills.

This is not to say that other dimensions are not relevant. Besides skill level there are a number of other dimensions that may convey useful information. Murray (2003) distinguishes several dimensions in skill use, of which criticality (or importance as it is usually called) and frequency are most often used. Importance may convey significant complementary information that allows us for example to assess the weight that has to be attached to certain skills shortages. Even small shortages may have a crucial impact if the required skill is regarded as important, while large shortages may be less alarming if it has been indicated that the skill in question is not particularly important. It is important to stress however, that importance cannot be considered a substitute for skill level.⁵ Some researchers commit this fallacy, assuming that if a given skill is regarded as important this implies that a high level is required. This need not be the case. For example, numeracy skills may be very important in a job as cashier, but the required level may be quite elementary.

Some researchers have preferred frequency of skill use as a measure of skills. One of the reasons of the popularity of the concept of frequency is that it offers far better possibilities to provide unambiguous anchors in the scale (like 'once a day', 'once a month' etc.). However, frequency is also no substitute for skill level. To give a simple example: Pilots are trained to carry out emergency landings. These require skills that they hopefully never have to use, but are nevertheless essential in becoming a competent pilot. In contrast to importance however, it is doubtful whether frequency of skill use conveys any useful additional information relating to the stock of human capital.⁶

⁵ An indication is the finding from Van Loo & Semeijn (2004) who show that use of skills and level of skills are better predictors for wages of higher education graduates than importance of skills.

⁶ This does not preclude that assessments of frequency may serve other useful purposes not directly relating to assessing the stock of human capital. In the context of occupational counselling for example, it is important to provide information on the frequency of certain tasks and related skills.

Methods used to assess the acquired level of skills

Table 1 gives an overview of the methods that are commonly used to assess skill levels in a given population.

Table 1. Methods to assess acquired level of skills

<i>Method</i>	<i>Level</i>
Proxy: <ul style="list-style-type: none"> • by education 	Aggregate of educational groups: level or field
Objective measures: <ul style="list-style-type: none"> • Assessment • Testing 	Individuals Individuals
Subjective measures <ul style="list-style-type: none"> • Supervisor rating • Individual self-assessment • Proxy by required skills 	Individuals Individuals Individuals

As outlined in the introduction, *level of education* or years of schooling have often been used as a *proxy* for the existing stock of skills in the labour force. In sociology as well as in economics, education is regarded as one of the most important and stable predictors of a range of outcomes, varying from socio-economic outcomes to political attitudes or health (Pallas, 2000). But educational credentials are not the same as skills and there is still a debate whether education actually causes these outcomes or not⁷. Moreover the effects of education may well underestimate the effects of skills. Murray (2003) for example indicates that even controlling for educational level, literacy skills have great additional explanatory value in explaining wage differences.

No systematic attempts have been made to analyse fields of study in terms of acquired specific skills (for an exception see Van de Werfhorst & Kraayvanger, 2001). This omission is odd, even more so in the light of the fact that occupations have been extensively assessed in terms of required skills. Since assessing acquired skills in a particular study program requires more or less the same methods as used in occupational analysis (see next section), one would expect this avenue to have been more thoroughly explored.

Assessment is usually carried out in specialised centres, where subjects are confronted with real life or simulated problems. Given the specific conditions, this method comes closest to our understanding of assessing competencies. It is context-bound, it involves solving complex real life problems, and it involves the mobilisation of cognitive as well as non-cognitive psychosocial prerequisites. This method is therefore often regarded as the 'gold standard' against which other measures should be judged (Ward et al., 2002). Where the methodology and practical considerations make it possible, the advantages of using this method are clear. It is however costly, which makes widespread application difficult. In addition, there may be a tendency for assessment methodology to lean more heavily towards aspects that are relatively easy to assess, and to neglect aspects that don't reveal themselves as clearly in concrete behaviour (e.g. Gray 1996; Arnold et al., 1985). As a result of the specificity of the assessment, comparability across the board is often very low. Finally, it is important to observe that even in those areas that do lend themselves well to this technique, measurement error (expressed for example in inconsistency among expert raters) can never be entirely eliminated (Harrington et al., 1997). So, although it is plausible that expert ratings provide better data than other methods, the difference in data quality may not be as great as sometimes assumed .

Testing is one of the most wide-spread methods for assessing skills. Examples include:

- The International Adult Literacy Survey (IALS) and its successor the Adult Literacy and Life Skills Survey (ALL). These are household based surveys focusing on the population of 16- to 64-year-olds. Both assess prose and document literacy and numeracy (see OECD, 2000). In

⁷ For example credentialists point out that the relation between education and wages is much stronger than between education and skills (see Collins, 1979 or Bills, 2003).

addition, ALL measures analytical reasoning and an array of questions regarding the actual use of skills and a self-evaluation of the adequacy of the respondent's skill levels.

- The International Association for the Evaluation of Educational Achievement (IEA) Trends in Mathematics and Science Study (TIMMS) and the OECD Programme of International Students Assessment (PISA). These are tests administered to students in primary (TIMMS: 4th grade) and secondary education (TIMMS: 8th grade and PISA: 15-year-olds). Both TIMMS and PISA measure mathematics and science, but PISA measures also reading literacy and some cross-curricular competencies (see Gonzales et al., 2004; OECD, 2004).
- The IEA Civic study carried out in 28 countries measures civic knowledge, skills and attitudes of 14-years-old students (Torney-Purta et al., 2001).

Testing is usually restricted to skills in the cognitive domain, although there is increasingly attention for non-cognitive skills as well (OECD, 1997; Peschar, 2001). The tests are individual based and results are comparable across the board. However even in the case of 'hard' skills like reading literacy, questions have been raised about the possible cultural bias (e.g. Emin, 2003). Moreover, tests may only be weakly related to the underlying competencies that we are interested in. For example in a study among surgeons Risucci et al. (1989) report only a moderate correlation between observation-based expert ratings and test scores.

In addition to these more or less 'objective' methods of skills assessment, there are several subjective methods. An example is *supervisor rating*, which aims to assess skills at an individual level and can include generic as well as specific skills. However, several problems preclude this method from becoming wide-spread. One of the reasons is that not everybody has a supervisor (think of managers, professionals, self-employed) or has a supervisor who is well-informed about actual performance (some workers have such high degree of work autonomy that the supervisor may not know much about the actual job content). Sample designs are also more complicated and would take at least two stages: one to identify workers and a second to identify their supervisors. As supervisors usually have more than one subordinate, a complex design is needed to ensure that the results can be generalised to the whole population of jobs. Furthermore, it is usually harder to get the cooperation of supervisors to participate in a general survey than to get the participation of workers, leading to substantial response problems, possibly biased towards workers who have good relations with their supervisors. An additional problem is that this method can only be applied to those who work.

Supervisor rating is an example of a more general cluster of methods to ask observers in an individual's direct environment, for example a colleague, a fellow student, a supervisor, a subordinate, or a client, to rate that individual's competencies. Several studies claim that peer assessment is more accurate than self-assessment (e.g. Bergee, 1997; Falchikov & Goldfinch, 2000). Ward et al. (2002) suggest that this may be due to the fact that individuals are capable of identifying good and bad performances, but are unwilling or unable to apply the same standard to their own performance.

Others question this assumption, pointing out that self-reports often provide more accurate information than information from observers (Mischel, 1968). Spenner (1990) concludes that self-reports offer relatively good prospects for skills measurement since there is no systematic evidence that people distort reporting of their job characteristics. In *individual self-assessment*, individuals are asked to rate their own level of skills in different domains. Examples include:

- The annual school-leaver surveys carried out by the Research Centre for Education and the Labour Market (ROA) in the Netherlands (ROA, 2004). These surveys are carried out approximately one year and a half after graduation and map the transition from school to work of school-leavers and graduates from secondary and higher education. A part of the mail questionnaire is focused on self-reports of the level of skills in different domains.
- The international CHEERS survey and its follow-up the REFLEX project (Teichler, in print; see also <http://www.uni-kassel.de/wz1/tseregs.htm> and <http://www.reflexproject.org/>). These surveys focus on the transition from higher education to work and are carried out in 11 European countries and Japan, respectively three (CHEERS) and five years (REFLEX) after graduation. The CHEERS survey concentrates on the skills level held at the moment of graduation, while the REFLEX project concentrates on the presently possessed skills level.

As will be elaborated in more detail below, this method has a number of clear advantages compared to alternatives. The advantages include the fact that such an approach is relatively cheap, easy to administer and flexible, making it well suited to large scale application in a range of situations. The main disadvantages of self-assessment revolve around the greater chance of measurement error.

Some researchers have therefore proposed the use of self-reported skill *requirements* in jobs as indicators of the actual skills of the holder of those jobs (see e.g. Green, 2004). The argument is briefly that if a job requires a certain skill, the job holder must also possess it to a certain extent. As mentioned above in the discussion of frequency of skill use, this claim is not without substance. Nonetheless, there are good reasons for caution in following this approach. Requirements and possession of skills are two different things, and there is clear evidence of both shortages and surpluses of skills, even among job incumbents with several years of tenure (Allen & Van der Velden, 2001). Especially the literature on over-education provides convincing evidence that employees may possess skills that are not optimally utilised in the work context. Equating possessed skills to required skills would fail to recognise this important fact. An additional objection is that skill requirements can by definition only be assessed among the employed population. The main reason why researchers have advocated this method is a conviction that self-reported skill requirements are less prone to response bias than self-assessments of own skills (Green, 2004). However, problems with response bias may equally hold for self-reports on required skills, and it is certainly not obvious that any advantage in this area would be strong enough to offset the disadvantages already stated. In our view response bias should be taken seriously, both with respect to possessed skills as with respect to required skills. We will return to this point later in more detail.

Methods used to assess required skills

The methods that are used to assess the required level of skills are to some extent the complement of the methods described above for acquired skills. Table 2 gives an overview of different methods used to assess skills requirements.

Table 2. Methods to assess skills requirements

<i>Method</i>	<i>Level</i>
Proxy: <ul style="list-style-type: none"> • by occupational analysis 	Aggregate of jobs: occupation
Objective measures: <ul style="list-style-type: none"> • Job analysis 	Individual jobs
Subjective measures <ul style="list-style-type: none"> • Employer survey • Supervisor rating • Worker's assessment 	Aggregate of jobs: sector or occupation Individual jobs Individual jobs

Analogous to education in the case of acquired skills, *occupational titles* are sometimes used as a *proxy* for required skills. However, in contrast to education, occupations have been subjected to extensive analysis, which permits a much greater level of detail in describing skills requirements. *Occupational analysis* is perhaps the most advanced method to assess skills requirements. Occupational analysis is usually carried out using a variety of instruments. Very often this kind of assessment starts with detailed analyses of 'typical' jobs within an occupation by job analysts, often combined with interviews with employers or supervisors on the present as well as the future skill requirements. This is sometimes followed by surveys among workers in a particular occupation asking them to rate the requirements in their job. The advantage of occupational analysis is that it gives very detailed information on specific skill requirements, but at the same time this makes it also more difficult to compare across different occupations. Nonetheless there are some good examples of systematic approaches covering hundreds of occupations. It is striking that these are all initiated by departments of labour or by central public employment offices. Just to give some examples⁸:

- The Occupational Information Network (O*NET) is a comprehensive database initiated by the US Department of Labor, comprising detailed information on worker attributes and job characteristics of hundreds of occupations (see <http://www.onetcenter.org>). It is the follow-up of the well-known Dictionary of Occupational Titles (DOT). The project started in 2001 with an annual data collection on 200 occupations. The goal is to replenish the database every five years. Information is based on surveys held among workers in these occupations on work

⁸ In other countries comparable initiatives are undertaken or being developed (e.g. Germany, the Netherlands).

tasks and job requirements and is complemented by a questionnaire focusing on abilities, which is completed by job analysts.

- COBRA (Competencies and Occupations Repertoire for the Labour Market) is a database from the Flemish Public Employment Office (VDAB). It consists of 550 detailed descriptions of occupations (see <http://vdab.be/cobra/info.shtml>). COBRA is based on the French ROME (Répertoire Opérationnel des Metiers et des Emplois). It includes descriptions of the job tasks, the required competencies, and the work environment. Data are mainly based on ratings by job analysts.

Job analysis is very much related to the former method but is used to evaluate individual jobs rather than occupations. It is the classical instrument for personnel managers to assess the requirements of individual jobs and to relate these to reward systems. It is usually carried out by experts (job analysts) who describe the different tasks in a job and relate these to specific skills requirements. Because of the specificity of the method, data are often not comparable across sectors or even across organisations within a sector. Occupational analysis can also be used to infer an individual's job requirements. The implication of course is that within-occupation variation is neglected as all workers within an occupational category will be assigned the same skill requirements. Moreover, some descriptions from occupational analysis may be outdated: in some sectors jobs may have changed completely since the last occupational analysis has been carried out.

In *employer surveys*, employers are asked about general job demands. This kind of survey concentrates on asking what employers think are the most relevant skills for the present workforce or what skills employers think will become most important in the future. The sample design usually comprises employers from all sectors, and the focus is on general rather than occupation-specific skills. Not surprisingly, the results have a high level of aggregation (that is they relate to all workers and not to some specific occupational group) and tend to be biased towards generic skills. Moreover there is a tendency to focus on skills shortages, while neglecting skills requirements that are well met by the workforce. The results are usually summed up in a list of skills that are most in demand. Examples include:

- The 'Employability Skills for British Columbia' project is a survey carried out among personnel managers from some 200 organisations in British Columbia (Debbing & Behrman, 1995). Respondents were asked to rate the importance of 187 different skills, that were derived from literature reviews on key skills.
- The 'Michigan Employability Skills' project is a survey among 2500 employers (O'Neil, Allred & Baker, 1992). It identifies 26 key skills (such as personality and team working skills), that were all considered important in meeting the demands of the employers.

An alternative to employer surveys that addresses some of its pitfalls is *supervisor rating*. Unlike employer surveys, it aims to assess required skills at the level of individual jobs and can include generic as well as specific skills. However, the problems indicated earlier with respect to acquired skills (complicated sample design; not every worker has a supervisor and not all supervisors have good knowledge about job content), preclude this method from becoming wide-spread.

The last method asks the worker to assess the skills requirements in his or her job. *Worker's assessment* is also sometimes used as part of occupational analysis (see above example of O*NET), but unlike occupational analysis, it provides up-to-date information about individual job requirements. Typical examples are:

- The British Skills Survey carried out in 1997 and 2001 among a representative sample of the British labour force by the ESRC Centre on Skills, Knowledge and Organisational Performance (SKOPE) at Oxford (Ashton et al., 1999; Felstead, Gallie and Green, 2002). The survey focuses on 36 job activities and related skills requirements.
- The above mentioned graduate surveys such as CHEERS, REFLEX and the school-leaver surveys carried out by ROA. Apart from assessing the possessed skills levels, these surveys also aim to assess to what extent these skills are required in the respondent's current job.

Worker's assessment of skill requirements shares most of the advantages and pitfalls of self-assessment of own skills. The rest of the paper will be devoted to a discussion of these points.

Advantages of using self-assessment

As outlined earlier, methods such as testing and assessment are well-suited to measuring skills in the cognitive domain, but are limited in their use for measuring other skills. Moreover, testing and assessment are time-consuming and therefore pose limits to the number of skills that can be assessed. Self-assessment may therefore provide an important complementary tool to testing.

Richter and Johnson (2001) list a number of clear advantages of using self-assessments in social research. Although their own research pertains to drug use, most of the advantages they mention apply equally to other kinds of 'hidden' personal information, including skills and competencies. The main advantages of self-assessments include the fact that they are relatively easy to administer to large samples, can be administered simultaneously in different locations, provide responses that are easily quantifiable and thus analyzable, are relatively inexpensive to produce and administer, and can be administered in any or all of a number of different ways, such as personal or telephone interviews, and questionnaire distributed by regular mail, email, or via the internet (see also Patrick & Sievert, 1994). An important advantage in the case of skills measurement is that self-assessments require less time than testing. Given the breadth of potentially relevant skills, the use of self-assessment offers the opportunity to dig into a wide array of skills that are thought to be relevant for well-functioning in work and in life.

In addition to these more or less practical advantages of self-assessment, there is, at least in theory, also a more substantive advantage of self-assessment as a method of data-collection, namely the fact that individuals have access to information about themselves that outside observers may not be aware of. Connally et al., (2002) point out that higher order competency are difficult to assess using direct observation. This implies that self-assessment need not only be regarded as a last resort when other methods are not feasible, but may have substantive advantages in its own right. That said it probably makes sense to not emphasize this advantage too strongly, since this self-knowledge is likely to be far from perfect, and more crucially, difficult to report in an objective way. This may result in problems relating both to the reliability and the validity of the results. We will discuss these problems as well as some of the solutions that have been offered in more detail in the next sections.

Problems arising in self-assessment

In a nutshell, the greatest disadvantage of self-assessment as a method of obtaining data is the greater chance of measurement error. In a meta-analysis of 44 self-assessment studies in higher education, Falchikov and Boud (1989) reported correlations between self-assessed and external measures of performance ranging from -0.05 to 0.82, with a mean correlation of 0.39. In a similar review of 18 self-assessment studies in the health professions, Gordon (1991) reported correlations ranging from 0.02 and 0.65. Although Ward et al. (2002) have cast doubt as to how much credence should be given to these correlations - most were between self-assessments and expert ratings, and the latter may themselves be flawed - it is clear that even in the most favourable case self-assessments paint a less than perfect picture. What are the main sources of error?

In principle, errors can be divided into those resulting from a more or less 'intentional' manipulation of answers by respondents, and unintentional discrepancies between the real and reported values. *Unintentional* measurement errors arise when the answers given by respondents in good faith do not correspond to the 'real' value on the variable in question. There are various reasons why unintentional measurement errors may occur. First of all, the *content of the question may be unclear or ambiguous*. This problem is likely to give rise to discrepancies between the concept as intended by the researchers and the concept as understood by respondents, as well as to discrepancies between the understanding of the concept by different groups of respondents (Ward et al. 2002). Dykema & Schaeffer (2000) have shown that complexity and clarity are strong predictors of measurement errors. Such errors seem particularly likely in the case of characteristics such as skills and competencies, which are by nature complex, abstract and difficult to delineate.

A second factor that can give rise to unintentional measurement errors is that of *limitations to respondents' comprehension or memory*. Even if the formulation of a question as such is completely clear, respondents can only report on what they understand, and what they can readily retrieve from memory. If confronted with questions that fall outside these limitations, respondents will be forced to

choose between skipping the question altogether or making a guess as to the answer. If the limitations are themselves differentially distributed between different groups of respondents, item non-response will be selective and constitute a form of measurement error. If graduates fail to comprehend the question fully but still offer an answer, the validity of the data will be compromised. An important point is that limitations to comprehension also apply to instructions or explanations given to help respondents understand the question. Very detailed or subtle instructions are themselves likely to be poorly understood. Problems with comprehension may be of particular importance when considering self-assessment of skills. It takes a certain level of (meta-)cognitive skills to be able to reflect about one's job, the requirements that are imposed and the possessed level. This may imply that self-assessment may be more difficult to administer among lower educated groups.

A third source of unintentional measurement error stems from the so-called *anchor problem*. This refers to ambiguity or lack of clarity of the measurement scale used. In contrast to variables such as working hours or income, there is no natural numerical scale on which to measure skills. This places a burden on researchers to provide a scale that is clearly understood in a uniform way by all categories of respondents. Ideally, all respondents should share the same understanding of what the extreme values and midpoint - the anchors in the scale - represent. Self-assessments of skills often use very general terms to indicate extreme values, such as 'very low' and 'very high'. Such scale values are not explicitly related to any objective characteristics in the real world. As a result of the ambiguous nature of the anchor points, different groups of respondents are likely to use their own frames of reference when answering the questions, so that the answers will not be comparable between groups (Ward et al., 2002). This can lead to systematic overestimation or underestimation of skills by different groups, whose reference groups have respectively a lower or higher level than the population at large. Implicitly, the extremes and midpoint on the scale might be assumed to correspond to the extremes and midpoint of the distribution of skills in the population as a whole. However, most respondents are unlikely to have a comprehensive overview of the total distribution of a particular skill in the population, particularly if that population is very broadly defined. In the absence of clear clues as to what 'very low' or 'very high' means, respondents will tend to use their own frame of reference of what is considered 'normal' or 'average'. In the case of skills, this is likely to be strongly biased by the respondent's own educational background or occupational affiliation. This implies that differences between occupational groups or fields of study are probably biased towards the mean, making it difficult if not impossible to assess the overall skill level or to compare different groups. Lack of clarity in the scale used may also give rise to the so-called halo-effect (Gray, 1996). This refers to the tendency of certain respondents to use only a small range of the scale (say 3 or 4 on a 5-point scale) for all questions.

In addition to these sources of unintentional errors by respondents, there are also various reasons why respondents might *intentionally* alter their true responses (Richter & Johnson, 2001). Many of these reasons fall under the general heading of what Orne (1962) calls "demand characteristics". This refers to any aspect of the research environment or the research instrument that communicates a "demand" for the respondent to behave in a particular way. One of the most commonly reported reasons is that of *social desirability* (Victorin, Haag-Gronlund, & Skerfving, 1998): respondents may alter their responses in order to appear more 'normal'. In the case of skills, some respondents may find it embarrassing to report very low or very high levels, for fear of appearing like 'dunces' or 'geeks'. Alternatively, respondents may have reasons to report more extreme values than apply to them in reality, for example out of *boastfulness* or *modesty*, or to deliberately mislead researchers. They may wish to appear consistent, unusual (Berg, 1967) or extreme. Despite reassurances about the confidentiality of the data, some respondents may fear that the information could be used against them. It is important to note that graduates may not always be fully aware of the fact that the answers they are giving are less than truthful. Even when they believe that they are answering honestly, individuals are often ignorant of their own motivations and internal states (Nisbett & Wilson, 1977). As a result, in practice it may at times be difficult to distinguish between intentional and unintentional measurement errors. This could imply that some remedies applied to reduce unintentional errors can also help reduce 'intentional' alteration.

Some solutions

There are several strategies that can be deployed to help researchers come to grips with measurement errors resulting from the use of self-assessments. The strategies can be divided into two broad categories. First of all, one can look for ways of improving the research instrument so as to

reduce or eliminate avoidable errors. Secondly, since some error is almost certainly unavoidable, an attempt can be made to gain an indication of the validity of the results, and thereby of the applicability of the data for various kinds of analyses. In some cases, the process of validation may provide a means of (partially) correcting for measurement errors, for example by recalibrating the data or developing appropriate control variables.

Addressing unintentional measurement error and increasing overall comprehension

There are a number of strategies that can be adopted to increase the overall reliability by reducing unintentional measurement error and increasing comprehension. Dykema & Schaeffer (2000) argue that complexity, clarity, and affective intensity are important determinants of measurement error. Although their approach cannot be directly applied to measurement of skills (their own research is about important events in respondents' lives rather than internal states such as skills), many of their arguments are relevant. The process of retrieval of information from long-term memory is affected by the nature of the stimuli used to trigger it. Retrieval is expected to be less accurate when the information is complex, indistinct from other information, and emotionally neutral. This suggests that measurement errors can be reduced by formulating items that are clear and unambiguous, that are clearly distinguishable from other items, and that elicit an emotional response from graduates. In the case of skills, the challenge is to formulate items that have a clear and uniform meaning to all graduates, to avoid items that are composites of several underlying dimensions, to choose items that are conceptually distinct from other skills, and to formulate the items in such a way as to tap into the feelings graduates have about their own (lack of) abilities. It is doubtful to what extent the latter suggestion can be implemented, but a minimum requirement is probably a formulation that is as active and - within the restrictions imposed by a general list - as concrete as possible. It seems advisable in any case to exercise a certain degree of caution and restraint in aiming for an emotional response, since a too emotionally charged formulation may constitute an unwanted "demand characteristic", and induce a deliberately altered response from some graduates.

Addressing the anchor problem

There are several ways to address the anchoring problem. *Ex ante expert anchoring* is probably the most widespread technique. It involves the a priori development by experts of an answer scale that has a clear and uniform meaning for all respondents. This comes down to providing explicit anchors for the evaluation criteria (Ward et al., 2002; Martin et al., 1998). The values assigned to the different levels of a rating scale convey information to the respondent regarding what is expected (Richter & Johnson (2001). Respondents will use such anchors as frames of reference for estimating their own responses (Schwarz, 1999). Ideally, the extreme points on the scale, as well as the mid-point, should correspond to something that all respondents know and assign the same meaning or interpretation to.

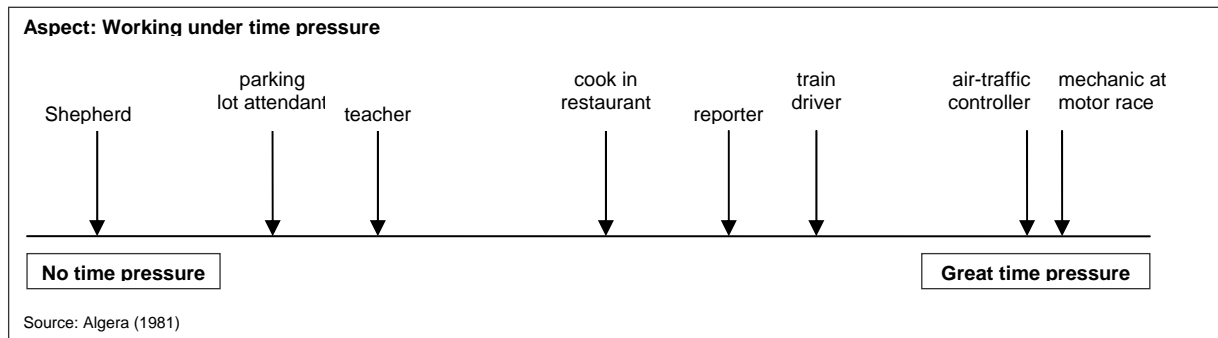
Two different forms of *ex ante expert anchoring* can be distinguished. One method uses occupational titles as anchors in the scale.⁹ ¹⁰ Algera (1981) applied this method to 24 different task characteristics and skill requirements (see Figure 1 for an example). On the basis of expert ratings, characteristic examples of occupations are located at appropriate points over the full range of the scale. Respondents are requested to position their own skill level with respect to the listed occupations. Although this method is attractive in theory, it is based on a number of assumptions which may not necessarily hold. First of all, the anchor occupations are assumed to be clear to all respondents. Finding occupations that fit this requirement may be easier said than done. For some skills it might prove difficult to find good anchors. Second, some inter-expert consistency has to be established before applying this method in a survey. This involves the usual set of methods to assess consistency between the rates of the different experts (e.g. inter-rating reliability tests). Third, the anchors must be clearly transitive: starting with the lowest level, each subsequent anchor in the scale must correspond to a more difficult level. Fourth, if the anchor coincides with the respondent's own occupation, he/she might fill in that anchor point even if their own level is very different. Fifth, a general assumption which

⁹ This method could also be applied using fields of education or other clearly recognisable social categories.

¹⁰ If occupations are used to provide the main anchors, one might also think of using experts to rescale the occupations afterwards: *ex post expert anchoring*. The assumption is that respondents will bias their group mean towards the over-all mean, leading to a decrease of the between-group variation (e.g. between occupations), compared to the within-group variation. Under the condition that the bias is only partial (i.e. there is still between-group variation left), the differences between groups can be rescaled using the rating of experts of some typical occupations at both ends of the distribution. Note however that the variation within groups may also be biased and that this way of anchoring does not change that problem.

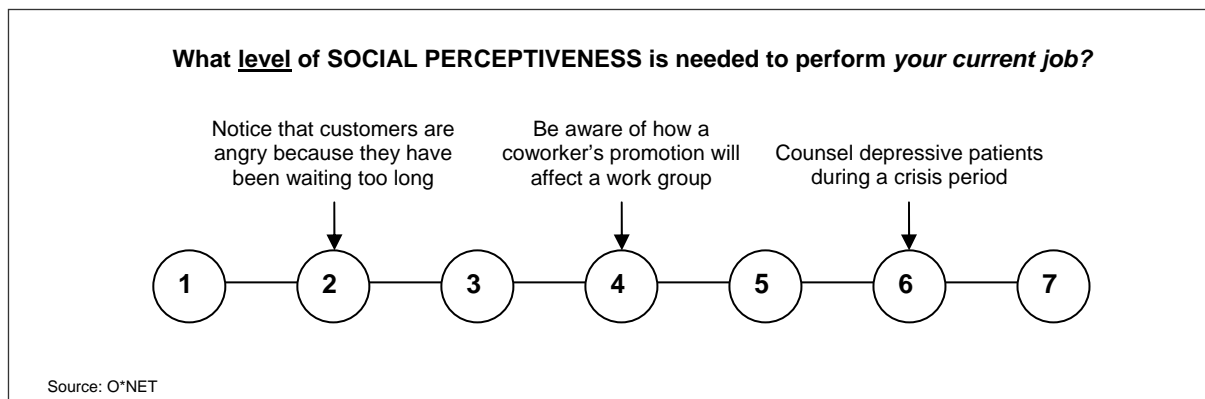
to some extent underlies all of the other assumptions is that an occupation can be regarded as a good proxy of given level on a particular skill. The actual assessment involves two steps: respondents are required to form an image of the skill level associated with each occupation, and then they are required to score their own level on that skill relative to these occupations. The use of this method may be less problematic if only three job titles are used, representing the extremes and the midpoint of the scale.

Figure 1. Example of ex ante anchoring using job titles



The second method of ex ante expert anchoring in effect skips the first of these two steps, using short descriptions of skill levels themselves as anchors. An example from the questionnaire used in the O*NET survey is provided in figure 2. In contrast to the former method, the anchors provide much clearer information about the concept that we are interested in. It should be remarked that this method is also subject to several of the objections that can be raised with respect to occupational anchors. It is still necessary that the anchor points are clear to respondents, that different experts agree on them, and that the answers are transitive. However, because the scale points directly describe skill levels rather than occupations, these assumptions are probably less problematic when using this method. If carefully and skilfully applied, this method can be expected to provide data of high quality.

Figure 2. Example of ex ante anchoring using descriptions of skill levels



Ex ante expert anchoring may not always be practicable. In particular in surveys where skills measurement is only part of a more comprehensive survey of study and/or work, it may be difficult to incorporate such an elaborate instrument without encountering a negative effect on response. In that case, another strategy is to use subjective anchors that are designed to elicit a more or less equivalent emotional response from all respondents (e.g. 'novice' and 'expert' instead of 'low level' and 'high level'). This is of course less precise than objective anchor points, but if carefully designed might at least have the effect of creating appropriate threshold levels for extreme scores. Ideally, the extremes should be formulated such that a small but significant minority of respondents feel that the description applies to them. If the threshold is too high, respondents will feel discouraged to use that end of the

scale, whereas if it is too low, there may be a glut of answers at that end of the scale. Ideally the scale should elicit a more or less normal distribution of answers.

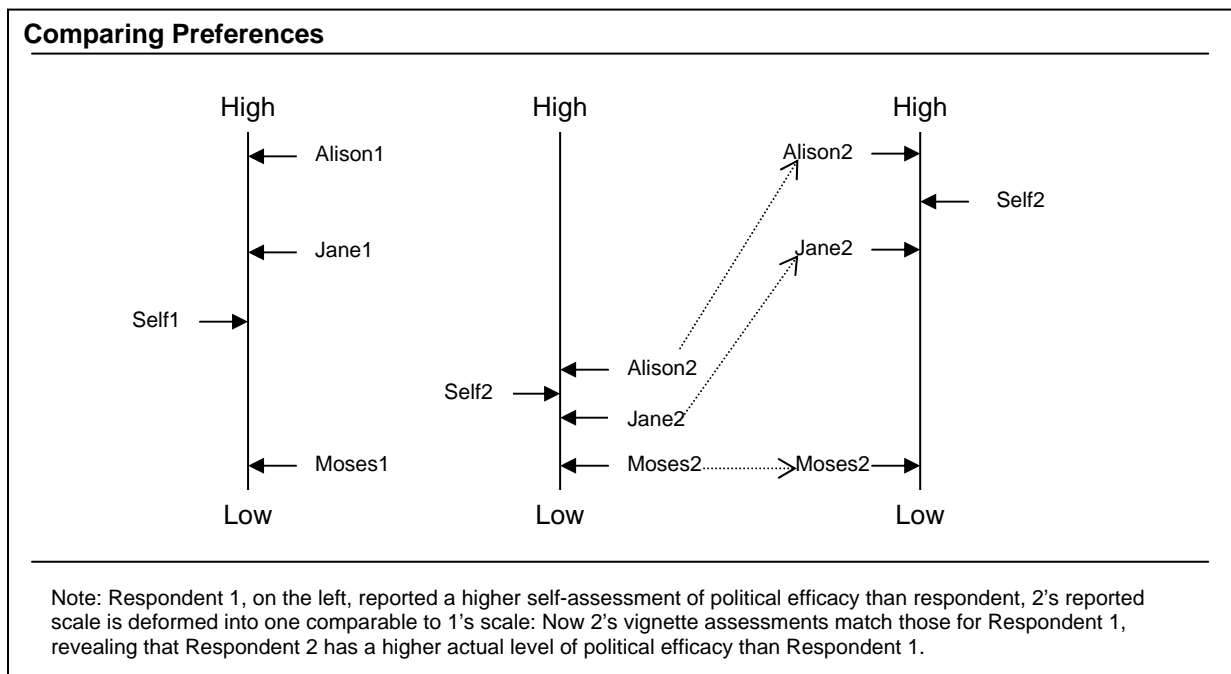
King et al. (2004) developed an alternative way of obtaining anchors. In their method it is the respondents themselves who provide the anchors: *inter subject anchoring*. The basic idea is that respondents are asked not only to rate themselves, but also to rate imaginary persons described on vignettes. As these vignettes are the same across all respondents, the ratings of these vignettes can be used to achieve. Although the method of anchoring by vignette was developed in the context of the World Health Survey and political science, the method can also be applied to the assessment of skills¹¹. An example is given in figure 3. Respondents are asked to assess their own political efficacy as well as the efficacy of each of the persons described on the vignettes¹². The basic idea is to recode the categorical self-assessment relative to the set of vignettes. In other words the vignette ratings are used to scale individual scores up or down, keeping the relative distances between the anchors the same across all individuals.¹³

Figure 3. Example of inter-subject anchoring using vignettes

1. “[Jane] lacks clean drinking water because the government is pursuing an industrial development plan. In the campaign for an upcoming election, an opposition party has promised to address the issue, but she feels it would be futile to vote for the opposition since the government is certain to win.”

2. “[Moses] lacks clean drinking water. He would like to change this, but he can’t vote, and feels that no one in the government cares about this issue. So he suffers in silence, hoping something will be done in the future.”

3. “[Alison] lacks clean drinking water. She and her neighbors are supporting an opposition candidate in the forthcoming elections that has promised to address the issue. It appears that so many people in her area feel the same way that the opposition candidate will defeat the incumbent representative.”



Source: King et al. (2004)

¹¹ Note that King et al. (2004) actually calls into question all assessments in sociological and political research using Likert items.

¹² Figure 3 shows only three of the five vignettes used by King et al.

¹³ King et al. (2004) provide an interesting and convincing result using this method. In a survey held for the WHO on the rating of political efficacy, the average political efficacy was (surprisingly) rated higher in China than in Mexico. After correcting with the vignettes the more plausible reverse rank order was obtained.

This method, although attractive, is extremely labour intensive, which may limit its usefulness in extensive skills surveys. In this respect, it is worth mentioning that it is possible to reduce the number of respondents that have to assess each vignette, by using parametric modelling. The model can best be thought of as an ordered probit model, in which the thresholds for the different categories are determined by a set of explanatory variables. These explanatory variables are used in turn to estimate the thresholds for those respondents for which only self-assessments are available. The same method can also be used post-hoc to recalibrate the scales, using a different sample. Two necessary conditions to be met is that there is overall *vignette equivalence* (i.e. there is only one underlying dimension determining the responses) and that respondents use the response categories in the same way for the self-assessment as for the assessment of the vignettes (*response consistency*). Specific software has been developed to assist researchers in applying this form of inter subject anchoring (see <http://gking.harvard.edu/vign/>).

A different way of dealing with the problem of anchor points is *anchoring by required level*. In that case the focus is on discrepancies rather than absolute skill levels. If both required and acquired skills are assessed using the same scale (skill level), the two aspects in effect form anchors for each other. Even if there is reason to doubt the comparability between groups of the two aspects separately, the discrepancies between acquired and required skills - that is skill shortages or surpluses - can be meaningfully compared between groups and even individuals. In terms of taking stock of available human capital, such information is of great value, since it includes an answer to the question 'is the current skill level enough?'

In some cases it may make sense to accept that there are no universal anchors, and to allow different (groups of) respondents to apply different anchors. For example, graduates could be asked to compare themselves with other graduates of their own (or similar) course(s) or workers could be asked to compare themselves with co-workers. Because the reference group is much clearer, and is more likely to fall within the limits of respondent's own field of comprehension or memory than the population as a whole, this is likely to be easier for respondents to answer. There are however some obvious drawbacks. Arguably the biggest drawback is that the data can no longer be used to give estimates of the overall skills level in the population. Such data can only be used in analyses that focus on determinants and effects of different skills levels within courses or occupations.

Some researchers have completely 'individualized' the frame of reference. Gruppen et al. (1997, 2000) and Fitzgerald et al. (2000) draw a distinction between the conception of self-assessment as an inter-individual process or as an intra-individual process. Most researchers adopt an inter-individual approach, but several researchers have moved more towards *intra-subject anchoring*. In this method the skills themselves act as each others' anchors. A number of researchers (e.g. Regehr et al., 1996; Ward et al., 2002) have developed such an approach, whereby respondents are asked to rank their own skills, first indicating a 'best', 'worst' and 'average' skill, and then scoring other skills with respect to these anchor points. Although such an approach would allow for a comparison of relative strengths and weaknesses across different (groups of) respondents, a drawback is that it in effect treats all respondents as having the same overall absolute level of skill.

Addressing intentional response style behaviour

In order to reduce intentional manipulation of answers by respondents, it is important to be aware of any "demand characteristics", i.e. characteristics associated with the instrument or the environment in which the research is conducted that may constitute a reason for respondents to alter their true answers. A basic point is that respondents should be assured that the result will be treated with complete confidentiality, thereby removing any objective motivation of respondents to alter answers for strategic reasons. But even if respondents are confident that their anonymity is guaranteed in the analysis and reporting of the results, some may still feel an emotional pressure to answer in a particular way. As outlined in the Expected Value Theory of Response Behaviour, it is the respondent's aim to achieve positive and avoid negative reactions from others (Stocké, 2004). In the case of skills, most respondents would probably regard it as more desirable to claim to be good at something than to be less competent.

To correct for social desirability, various measures of social desirability have been developed (e.g. Crowne & Marlowe, 1964), which can, at least in theory, be incorporated into the questionnaire and used to correct the data afterwards, using factor analysis (Morf & Jackson, 1972; Paulhus, 1982) or covariance analysis (Norman, 1967). However, most authors agree that it is nearly impossible to eliminate the effect entirely (Richter & Johnson, 2001), and many tools for assessing social desirability are inconsistent with one another (Strohshal, Linehan & Chiles, 1984).

Various methods have been developed to pressure or trick respondents into answering more honestly, such as the "bogus pipeline" technique (Jones & Sigall, 1971) or the randomized response method (Greenberg, et al., 1969), but such techniques usually require subjects to complete the self-assessment in the presence of an investigator. However, most research on the problem of social desirability indicates that the presence of an investigator actually aggravates this kind of response bias (Krysan, 1998).

It is therefore important that respondents are given the feeling that any 'legal' answer (i.e. any answer falling within the specified range) can also be regarded as 'normal', and to remove as far as possible any "demand characteristic" that may influence the way they feel they should answer the questions. A clear recommendation is therefore that self-assessments should be carried without the presence of an interviewer.

Since many respondents are not fully aware that they are manipulating their answers, a clear and unambiguous formulation of the questions may help, by reducing the scope respondents have to unknowingly 'bend the truth'. Similarly, as Richter & Johnson (2001) point out, the use of scales with clear anchor points can convey information to the respondent regarding what is deemed a normal or average response, an example of a case where demand characteristics can be used to reduce rather than increase measurement error. This is likely to work best when objective anchor points are used, but even judiciously chosen subjective 'anchors' (such as 'novice' and 'expert') may reduce unconscious manipulation of answers.

There is probably little that can be done to prevent respondents from answering questions in an unduly 'boastful' (or 'modest') manner if they deliberately decide to answer falsely. Boasting and modesty refer to the tendency of certain people to systematically over- or underestimate their capacities (and probably their job requirements as well). Since there is no natural scale against which to measure skills, 'boastfulness' and 'modesty' are relative rather than absolute concepts. If all graduates are 'boastful' to an equal degree, the data can still be regarded as painting a reliable picture of the distribution of skills among the population. However, this is not a very likely proposition and boasting will probably differ among groups (e.g. males and females) leading to systematic biases in the estimation of skills. A good way to address boastfulness is to have independent objective measures of skills alongside the self-assessment. For example one might have objective test results on literacy skills as well as a subjective self-assessment of the same skill. Under the assumption that boasting (or modesty) will affect self-assessments of other skills (for which no objective test results are available) in more or less a similar way, the difference in ranking position of the respondents on the two variables can be used either to correct the other self-assessments or to statistically control for boasting in multivariate analyses.

Other approaches towards differences in response style look at tendencies to exhibit extreme response behaviour. This may relate to tendencies to avoid extreme categories in a scale, or the opposite, to use systematically lower or upper ends of the scale. These approaches use structural equation modelling (Billiet & McClendon, 2000) or a latent class approach (Moors, 2004), to detect response style factors. The approach requires at least two different sets of indicators referring to two different latent constructs. Both sets must be balanced, that is must contain both negative and positive worded items. Because of these characteristics, this approach seems less applicable for skills assessments, where usually items are formulated in one direction. Piquero et al. (2002) advocate Rasch modelling to detect response style differences. They analysed the validity of self-reports on delinquency. In particular they show that certain items in the Self-Reported Delinquency Scale show considerable differences in item difficulty between groups (Differential Item Functioning or DIF) when analysed in a Rasch model. In classical test theory these items would not have been detected as eliciting different responses from different groups (Piquero et al., 2002).

The importance of validation and testing

In any form of research it is important to take appropriate measures to ensure the validity and reliability of the data. Given the possible problems involved in self-assessments, rigorous validation and testing procedures may be considered even more important than usual. Put simply, validity addresses the question of whether we measure what we believe we are measuring (Baker, 1988). Most standard methodology handbooks distinguish several forms of validity. The most basic form is content validity. This amounts to a critical examination of the measure of a concept in the light of its intended meaning. This involves among other things asking whether the empirical indicators fully represent the domain of meaning of the underlying concept (Bohrnstedt, 1983). We already saw that discrepancies between the meaning of a given question as intended by the researchers and the meaning as understood by graduates is a potential major source of measurement error. Validity may however already be compromised if the operational definition of skills (the questions in the questionnaire) differs from the theoretical definition (what we really mean by skills). Obviously, considerations of content validity are particularly important role during the stage of developing the research instrument, because it is then still possible to adjust the instrument to improve its validity.

Content validity is usually regarded as part of a wider concept of construct validity (Cronbach & Meehl, 1955), The respondent's self-reports are intended to indicate the actual scores on an underlying set of skills or competencies. Probably the most powerful form of construct validation of self-assessments involves checking the self-assessments against more objective measures of the variables they are intended to indicate. As self-assessments are often used as an alternative for testing, these objective test results will not always be available for the same group of respondents. Even in that case however, it is useful to compare the results with other data sets in which test results for the same kind of skills are available and see if comparable patterns of association with other variables (such as educational attainment) exist.

A straightforward and transparent form of validity is predictive validity (de Groot, 1981). This form is applicable when a variable is intended as an indicator or predictor of another (possibly latent) criterion variable. It is based on forming and testing hypotheses about the concepts that are being measured (Baker, 1988). The hypotheses normally take the form of predictions as to the kinds of other variables the measures are likely to vary with. For example, in the case of skills it could be hypothesized that certain skills are required more in certain occupations than in other occupations. Alternatively, it might be predicted that certain skills will be good predictors of different labour market outcomes. To the extent that the hypothesized relationships with other variables are found, the measures could be regarded as being valid measures. A problem with using this kind of approach to validation might arise when such relationships are actually what one is trying to establish in a project using skills assessments. It would be unsound methodologically to use the same substantive empirical results that one would like to publish as part of the outcomes of the project as proof that the variables one uses to derive those results are valid.

Predictive validity can also be used to correct the scaling of a variable. Optimal scaling refers to a kind of advanced regression analysis in which the data are recalibrated for different groups of respondents in order to give the best possible prediction of a particular criterion variable (see e.g. Ganzeboom et al., 1992). There are several objections to using this method to recalibrate in the case of skills, the most important of which is the lack of a sufficiently valid criterion variable. If we were to use for example income as a criterion, the skills would be recalibrated so as to give an optimal prediction of income across all categories of respondents. However, this ignores the fact that in many cases, income differences may have nothing at all to do with skills, but may be due to differences in institutional arrangements, particular patterns of lifetime earnings profiles, etc. A further objection is that it would be difficult if not impossible to recalibrate both required and actual skills in such a way as to retain meaningful differences (i.e. discrepancies) between the two.

Conclusions

The last few decades have seen an increased awareness of human capital as one of the driving forces in economic development. Different actors in society (policy makers, employers, students, employees) have realised the importance of investing in education and training as a way of improving the existing stock of skills. This heightened interest in education and learning has been accompanied

by a greater need to monitor and assess the stock of human capital. In the absence of better indicators, many have relied on education as a proxy for human capital. However education is a rather poor indicator. Individuals with the same nominal level and type of education can differ markedly in their command of various skills, and this problem is further exacerbated by processes of appreciation and depreciation of skills over the life course.

Since the 1990's several large-scale international surveys have been undertaken to measure skills. These skills assessments focussed on areas like prose and document literacy, mathematics, science, problem solving as well as some cross-curricular competencies (OECD, 1997; 2000; 2004; Torney-Purta et al., 2001). Notwithstanding the importance of these skills for dealing with the complexities of today's world, they nevertheless represent only a fraction of the skills and competencies that are assumed to be 'key' (Rychen & Salganic, 2003a). In a way we could say that skills researchers find themselves confronted by the limitations of classical methods of assessing skills, such as large scale testing. First of all, not all skills lend themselves easily for testing. This applies especially for skills in the non-cognitive domain that are less-well defined and more difficult to measure in tests. Secondly, even if all skills could be tested, there is a physical limit to what respondents can endure in such an assessment. If 90 minutes is taken to be a maximum for survey time (Murray, 2003), even a small number of tests would already consume a large part of this.

This could potentially impose an undesirable limit to the breadth of skills that are measured. In this paper we have developed a plea for using self-assessments as a complementary tool to assess skills. Self-assessments are relatively easy and cheap to administer to large populations and the method can be used to measure a wide array of skills. As respondents are probably the best informants about their own skills, self-assessment can reveal information that cannot directly be tested or observed by outsiders. That being said, the method has its drawbacks, and serious questions have been raised about the reliability and validity of the responses.

In the paper we have discussed a number of problems that may arise when using self-assessments. What lessons can be learned from this?

1. If possible use a combination of different methods.
2. Assess both the level of possessed *and* required skills.
3. Remove any characteristics that may elicit responses that are socially desirable or manipulated in other ways.
4. Provide clear anchors in the scale, by giving short descriptions that make clear what level is indicated.
5. If this is not possible look at other forms of anchoring, for example anchoring by vignette or anchoring by required level.
6. Avoid items that are composites of several underlying dimensions.
7. Make items as concrete and active as possible.
8. Make wording of questions and answer categories so that any 'legal' response looks normal.
9. As measurement errors are unavoidable, it is important to plan in advance on ways of checking for, and if possible correcting errors.
10. Finally, one needs to be aware at all times when analysing and reporting on the data what the limitations of the data are.

We are still left with some unresolved issues that were not discussed in this paper. An important one relates to the difference between measuring skills and measuring competencies. Given the more 'holistic' concept of competencies, an important implication of this is that competence can only be assessed indirectly (Oates, 2003). Assessment of competencies should be based on the performance of individuals in dealing with a complex demand in a variety of settings. Most traditional ways of testing can only give an approximation of this ideal. Even more advanced ways of assessing skills are usually performed in situations that are artificial at best. Self-assessments do not provide a way of out for this problem, as by definition they are restricted to provide only indirect measures of competencies.

A second issue relates to the kind of skills that are being measured. Given the difficulties in comparing specific skills across different settings, these specific skills tend to be undervalued in most assessments involving broader cross-sections of society. Although understandable, this could potentially result in an underestimation of the importance of these skills and therefore to an imbalanced view of the total stock of human capital. There is probably no easy solution to this problem. Some attempts have been made

to assess domain-specific skills in a context neutral way (see Allen, Ramaekers & Van der Velden, forthcoming). But we have to keep in mind that even such attempts are likely to underestimate the importance of specific skills.¹⁴

Although self-assessment as a way of measuring 'hidden' characteristics of individuals such as skills has its drawbacks, the method is popular and widely used. This popularity reflects in particular the convenience of this method as a way of quickly obtaining a large amount of usable data. However, the popularity is also testimony to the fact that the measures obtained, although never perfect, can shed real light on the capacities of the groups that one is interested in. The potential value of even imperfect knowledge of these skills is great. It goes without saying that such research should be undertaken in full knowledge and awareness of its limitations and of the pitfalls of reading more into the results than the methodology justifies. No method of measuring skills is without its flaws, but given that sufficient care is taken in both the collection and the use of the data, the benefits of using self-assessments should almost certainly outweigh the disadvantages.

Literature

Algera, J.A. (1981), *Kenmerken van Werk*, Lisse: Swetz & Zeitlinger.

Allen, J., G. Ramaekers & R. van der Velden (forthcoming), Measuring Competencies of Higher Education Graduates, *New Directions for Institutional Research*.

Allen, J. & R. van der Velden (2001), Educational Mismatches Versus Skill Mismatches: Effects on Wages, Job-related Training, and On-the-job Search, *Oxford Economic Papers*, 3, 434-452.

Arnold, L., T.L. Willoughby & E.V. Calkins (1985), Self-evaluation in undergraduate medical education: A longitudinal perspective. *Journal of Medical Education*, 60, 21-28.

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

Australian National Training Authority (2003), *Defining Generic Skills at a Glance*, Adelaide: NCVET Ltd.

Baker, T.L. (1988), *Doing Social Research*, New York, etc.: McGraw-Hill International Editions.

Becker, G. S. (1964), *Human Capital. A Theoretical and Empirical Analysis, with Special Reference to Education*. New York: NBER.

Berg, I.A. (ed.) (1967), *Response Set in Personality Assessment*. Chicago: Aldine.

Bergee, M.J. (1997), Relationships among Faculty, Peer and Self-evaluations of Applied Performances. *Journal of Research in Music Education*, 45, 601-612.

Billiet, J. & M.J. McClendon (2000), Modeling Acquiescence in Measurement Models for Two Balanced Sets of Items, *Structural Equation Modeling*, 7, 608-628.

Bills, D. (2003), Credentials, Signals and Screens: Explaining the Relationship between Schooling and Job Assignment, *Review of Educational Research*, 73: 441-70.

Bohrnstedt, G.W. (1983), Measurement. In P.H. Rossi, J.D. Wright & A.B. Anderson (eds.), *Handbook of Survey Research*, New York: Academic.

Collins, Randall (1979), *The Credential Society: An Historical Sociology of Education and Stratification*. New York: Academic Press.

¹⁴ To make things even more complicated, we may wonder to what extent it is even possible to assess generic skills in a non-context bound way. Some cognitive psychologists doubt whether generic skills even exist (Perkins & Salomon, 1989).

Connally, J., K. Jorgensen, S. Gillis & P. Griffin (2002), *An Integrated Approach to the Assessment of Higher Order Competencies*. Paper presented at the Australian Association for Research in Education Annual Conference, Brisbane, Australia, December 2002.

Cronbach, L.J. & P.E. Meehl (1955), Construct Validity in Psychological Tests. *Psychological Bulletin*, 52, 281-301.

Crowne D. P. & D. Marlowe (1964), *The Approval Motive*. John Wiley, New York.

Debbling, G. & B. Behrman (1995), *Employability Skills for British Columbia*, Victoria British Columbia: Ministry of Advanced Education, Training and Technology.

Dykema, J. & N.C. Schaeffer (2000), *Development of a Personal Event Schema*. CDE Working Paper 99-27, Madison: Center for Demography and Ecology, University of Wisconsin-Madison.

Emin, J.-C. (2003), Proposal for a Strategy to Assess Adults' Competencies, In: D.S. Rychen, L.H. Salganic & M.E. McLaughlin (eds.), *Selected Contributions to the 2nd DeSeCo Symposium*, Neuchâtel, Swiss Federal Statistical Office.

European Commission (1995), *Teaching and Learning. Towards the Learning Society*, White Paper on Education and Training, Brussels: EC.

Falchikov, N., & D. Boud (1989), Student Self-assessment in Higher Education: A Meta-analysis. *Review of Educational Research*, 59: 395-430.

Falchikov, N. & J. Goldfinch (2000), Student Peer Assessment in Higher Education: a Meta-analysis Comparing Peer and Teacher Marks. *Review of Educational Research*, 70, 287-322.

Felstead, A., D. Gallie & F. Green (2002), *Work Skills in Britain 1986-2001*, Oxford: SKOPE, Oxford and Warwick Universities.

Fitzgerald, J.T., L.D. Gruppen, & C.B. White (2000), The Influence of Task Formats on the Accuracy of Medical Students' Self-assessments. *Academic Medicine*, 75, 737-741.

Ganzeboom, Harry B.G.; De Graaf, Paul; Treiman, Donald J.; (with De Leeuw, Jan) (1992), A Standard International Socio-Economic Index of Occupational Status, *Social Science Research*, 21, 1, 1-56.

Gonzales, P. J.C. Guzman, L. Partelow, E. Pahlke, L. Jocelyn, D. Kastberg & T. Williams, (2004), *Highlights From the Trends in International Mathematics and Science Study: TIMSS 2003*. Washington DC: NCES.

Gordon, M.J. (1991), A Review of the Validity and Accuracy of Self-assessments in Health Professions Training. *Academic medicine* 66: 762-769.

Gray, J.D. (1996), Global Rating Scales in Residency Education. *Academic Medicine*, 71 (Supplement): S55-S63.

Green, F. (2004), *First Thoughts on Methodological Issues in an International Assessment of Adult Skills*, Expert paper prepared for the first PIAAC IEG meeting, Paris, 26-27 April, OECD.

Greenberg, B. C., A.L. Abdula, W.L. Simmons, & D.G. Horvitz (1969), The Unrelated Question in Randomized Response Model, Theoretical Framework. *Journal of the American Statistical Association*, 64, 520-539

Groot, A.D. de (1981), *Methodologie: Grondslagen van Onderzoek in de Gedragwetenschappen*. The Hague: Mouton.

Gruppen, L.D., J. Garcia, C.M. Grum, J.T. Fitzgerald, C.A. White & L. Dicken, (1997), Medical Students' Self-assessment Accuracy in Communication Skills. *Academic Medicine*, 72(10 Supplement 1): S57-S59.

Gruppen, L.D., C. White, J.T. Fitzgerald, C.M. Grum, & J.O. Wooliscraft (2000), Medical Students' Self-assessments and their Allocation of Learning Time. *Academic Medicine*, 75, 374-379.

Harrington, J.P., J.J. Murnaghan, & G. Regehr (1997), Applying a Relative Ranking Model to the Self-assessment of Extended Performances. *Advances in Health Sciences Education*, 2, 17-25.

Hartog, J. (2000), Over-Education and Earnings: Where are We, Where Should We Go?, *Economics of Education Review*, 19, 131-147.

Jones, E. E., & H. Sigall (1971), The Bogus Pipeline: A New Paradigm for Measuring Affect and Attitude. *Psychological Bulletin* , 76, 349-364.

King, G, C.J.L. Murray, J.A. Salomon & A. Tandon (2004), Enhancing the Validity and Cross-Cultural Comparability of Measurement in Survey Research, *American Political Science Review*, 98, 1, 191-207.

King, G & J. Wand (2004), Comparing Incomparable Survey Responses: New Tools for Anchoring Vignettes, Mimeo, Cambridge MA: Harvard University

Krysan, M. (1998), Privacy and the Expression of White Racial Attitudes: A Comparison Across Three Contexts, *Public Opinion Quarterly*, 62, 506-544.

Loo, J. van & J. Semeijn (2004), Defining and Measuring Competences: An Application to Graduate Surveys, *Quality and Quantity*, 38, 3, 331-349.

Mertens, D. (1974), Schlüsselqualifikationen: Thesen zur Schulung für eine moderne Gesellschaft, *Mitteilungen aus der Arbeitsmarkt- und Berufsforschung*, 7, 1, 36-43.

Mischel, W. (1968), *Personality and Assessment*. New York: Wiley.

Moors, G. (2004), Facts and Artefacts in the Comparison of Attitudes Among Ethnic Minorities. A Multigroup Latent Class Structure Model with Adjustment for Response Style Behaviour, *European Sociological Review*, 20, 4, 303-320.

Morf, M. E., & D.N. Jackson, (1972), An Analysis of Two Response Styles: True Responding and Item Endorsement. *Educational and Psychological Measurement* , 32, 329-353.

Murray, T.S. (2003), Reflections on International Competence Assessments, In: D.S. Rychen & L.H. Salganic (eds.) *Key Competencies for a Successful Life and a Well-functioning Society*, Göttingen: Hogrefe & Huber, pp. 135-159.

National Commission on Excellence in Education (1983), *A Nation at Risk: The Imperative for Educational Reform*, Washington, DC: U.S. Government Printing Office.

National Competence Account (2002), *On the Track of Danish Competences*, Copenhagen: Danish Ministry of Education. Retrieved from: <http://www.nkr.dk/db/filarkiv/4162/NKRfolderengelsk.pdf>

National Skills Task Force (2000), *Skills for All: Proposals for a National Skills Agenda*, Suffolk: Department for Education and Employment.

Nijhof, W.J. (1998), Qualifying for the Future, In: W.J. Nijhof & J.N. Streumer, *Key Qualifications in Work and Education*, Dordrecht: Kluwer Academic Publishers, 19-38.

Nisbett, R., & T. Wilson (1977), Telling More Than We Can Know: Verbal Reports on Mental Processes. *Psychological Review*, 84, 231-259.

Norman, W.T. (1967), *2800 Personality Trait Descriptors: Normative Operating Characteristics for a University Population*. Ann Arbor: University of Michigan, Department of Psychological Sciences.

Oates, T. (2003), Key Skills/Key Competencies: Avoiding the Pitfalls of Current Initiatives. In: D.S. Rychen, L.H. Salganic & M.E. McLaughlin (eds.), *Selected Contributions to the 2nd DeSeCo Symposium*, Neuchâtel, Swiss Federal Statistical Office.

O'Neil, H.F., K. Alfred & E.L. Baker (1992), *Measurement of Workforce Readiness Competencies: Design of Prototype Measures*, Los Angeles: University of California, CRESST

Organisation for Economic Co-operation and Development (1994), *Making Education Count: Developing and Using International Indicators*, Paris: OECD.

Organisation for Economic Co-operation and Development (1997), *Prepared for Life?*, Paris: OECD.

Organisation for Economic Co-operation and Development (2000), *Literacy in the Information Age: Final Report of the International Adult Literacy Survey*, Paris: OECD and Ottawa: Statistics Canada.

Organisation for Economic Co-operation and Development (2004), *Learning for Tomorrow's World: First Results from PISA 2003*, Paris: OECD.

Orne, M. T. (1962), On the Social Psychology of the Psychological Experiment: With Particular Reference to Demand Characteristics and their Implications. *American Psychologist*, 17, 776–783.

Pallas, A.M. 2000. The Effects of Schooling on Individual Lives. In: M.T. Hallinan (ed.), *Handbook of the Sociology of Education*, New York: Kluwer Academic/Plenum Publishers, 499-525.

Patrick, T.B. & M.C. Sievert (1994), Electronic Communication on the Internet, *Missouri Medicine*, 91, 1, 25-26.

Paulhus, D. L. (1984), Two Component Models of Socially Desirable Responding. *Journal of Personality and Social Psychology*, 46, 598-609.

Perkins, D.N. & G. Salomon (1989), Are Cognitive Skills Context Bound?, *Educational Researcher*, 18, 1, 16-25.

Peschar, J.L. (2001), Cross Curricular Competencies: Developments in a New Area of Educational Outcome Indicators, In: E. Owen (ed.), *International Education Outcome Indicators*, Paris: OECD and Washington DC: NCES.

Piquero, A.R., R. MacIntosh & M. Hickman, (2002), The Validity of a Self-Reported Delinquency Scale. Comparisons Across Gender, Age, Race, and Place of Residence, *Sociological Methods and Research*, 30, 4, 492-529.

Psacharopoulos, G. (1973), *Returns to Education: An International Comparison*, Amsterdam: Elsevier.

Regehr, G., B. Hodges, R. Tiberius, & J. Lofchy (1996), Measuring Self-assessment Skills : An Innovative Relative Ranking Model. *Academic Medicine* 71(10 Supplement: S52-S4).

Researchcentrum voor Onderwijs en Arbeidsmarkt (2004). *Schoolverlaters tussen Onderwijs en Arbeidsmarkt 2003*. ROA-R-2004/3A. Maastricht: ROA

Richter, L, & P.B. Johnson (2001), Current Methods of Assessing Substance Use: A Review of Strengths, Problems, and Developments. *Journal of Drug Issues*, 31, 4, 809-832.

Risucci, D.A., A.J. Tortolani, & R.J. Ward (1989), Ratings of Surgical Residents by Self, Supervisors, and Peers. *Surgery, Gynaecology & Obstetrics*, 169, 519-526.

Rychen, D.S. & L.H. Salganic (eds.) (2003a), *Key Competencies for a Successful Life and a Well-functioning Society*, Göttingen: Hogrefe & Huber.

- Rychen, D.S. & L.H. Salganic (2003b), A Holistic Model of Competence, In: D.S. Rychen & L.H. Salganic (eds.) *Key Competencies for a Successful Life and a Well-functioning Society*, Göttingen: Hogrefe & Huber, pp. 41-62.
- Rychen, D.S. (2003), Key Competencies: Meeting Important Challenges in Life, In: D.S. Rychen. & L.H. Salganic (eds.) *Key Competencies for a Successful Life and a Well-functioning Society*, Göttingen: Hogrefe & Huber, pp. 63-107
- Sattinger, M. (1993), Assignment Models of the Distribution of Earnings, *Journal of Economic Literature*, 31, 851-880.
- Schultz, T.W. (1961), Investment in Human Capital, *American Economic Review*, 51, 1, 1-17.
- Schwarz, N. (1999), Self-reports: How the Questions Shape the Answers. *American Psychologist*, 54, 93-105.
- Secretary's Commission on the Achievement of Necessary Skills (1990), *Identifying and Describing the Skills Required by Work*, Washington: Pelavin Associates.
- Spenner, K.L. (1990), Skill: Meaning, Methods and Measures, *Work and Occupations*, 17, 4, 399-421.
- Stocké, V. (2004), Determinants and Consequences of survey Respondents' Social Desirability Beliefs about Racial Attitudes, Sonderforschungsbereich 504, Mannheim: Universität Mannheim.
- Strohshal, K., M.M. Linehan, & J.A. Chiles (1984), Will the Real Social Desirability Please Stand Up? Hopelessness, Depression, Social Desirability and the Prediction of Suicidal Behavior. *Journal of Consulting and Clinical Psychology*, 52. 449-457.
- Teichler, U. (ed.) (in print), *Higher Education and Graduate Employment in Europe*, Dordrecht: Kluwer Academic Publishers.
- Toolsema, B. (2003), *Werken met Competenties: Naar een Instrument voor de Identificatie van Competenties*, Thesis, Enschede: Universiteit Twente.
- Torney-Purta, J., R. Lehmann, H. Oswald & W. Schulz (2001), *Citizenship and Education in Twenty-Eight Countries: Civic knowledge and Engagement at Age Fourteen*, Amsterdam: International Association for the Evaluation of Educational Achievement.
- Victorin K., M. Haag Grönlund & S. Skerfving (1998), Methods for Health Risk Assessment of Chemicals - Are they Relevant for Alcohol? *Alcoholism Clinical and Experimental Research*, 22 (7) suppl, 270S-276S.
- Ward, M., L. Gruppen, & G. Regehr (2002), Measuring Self-assessment: Current State of the Art. *Advances in Health Sciences Education*, 7, 63-80.
- Weinert, F.E. (2001), Concept of Competence: A Conceptual Clarification, In: D.S. Rychen & L.H. Salganic (eds.) *Defining and Selecting Key Competencies*, Göttingen: Hogrefe & Huber, pp. 45-65.
- Werforst, H.G. van de & Kraaykamp, G. (2001), Four Field-Related Educational Resources and Their Impact on Labor, Consumption, and Sociopolitical Orientation, *Sociology of Education*, vol. 74, 4, pp. 296-317.
- World Bank (2002), *Lifelong Learning in the Global Knowledge Economy: Challenges for Developing Countries*. Washington D.C.: World Bank Group.