Educational Research Review 7 (2012) 1–18

Contents lists available at SciVerse ScienceDirect

Educational Research Review

journal homepage: www.elsevier.com/locate/EDUREV

Review

Self-rated competences questionnaires from a design perspective

Edith Braun [a,*], Alan Woodley [b], John T.E. Richardson [b], Bernhard Leidner [c]

[a] HIS Institut für Hochschulforschung, Goseriede 9, 30159 Hannover, Germany
[b] Institute of Educational Technology, The Open University, Milton Keynes MK7 6AA, United Kingdom
[c] Department of Psychology, University of California, One Shields Avenue, Davis, CA 95616, USA

ARTICLE INFO

Article history:
Received 22 May 2010
Revised 15 November 2011
Accepted 16 November 2011
Available online 25 November 2011

Keywords:
Self rated-competences
Higher education
Questionnaire design

ABSTRACT

This paper provides a theoretical review of self-rated competences questionnaires. This topic is influenced by the ongoing world-wide reform of higher education, which has led to a focus on the learner outcomes of higher education. Consequently, questionnaires on self-rated competences have increasingly been employed. However, self-ratings are often criticised for their lack of validity. Our intention is to outline some principles of good questionnaire design and to use these principles to contrast questionnaires on self-rated competences. We begin with an overview of research about questionnaire design. Then we introduce seven questionnaires and portray them in terms of their design characteristics. A comparison reveals some significant points: biographical data need to be handled more carefully, and there is an overuse of vague and abstract language. On the positive side, all of the questionnaires that were examined provide reliable sub-scales covering important facets of competences.

Contents

* Corresponding author. Present address: HIS Hochschul-Informations-System GmbH, Goseriede 9, 30159 Hannover, Germany. Tel.: +49 511 1220 477; fax: +49 511 1220 431.
E-mail addresses: braun@his.de (E. Braun), A.Woodley@open.ac.uk (A. Woodley), J.T.E.Richardson@open.ac.uk (J.T.E. Richardson), bleidner@ucdavis.edu (B. Leidner).

1747-938X/$ - see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.edurev.2011.11.005

1. Introduction

The aim of this paper is to introduce a number of scientific questionnaires that attempt to measure self-rated competences in higher education students. Many researchers and quality agencies are interested in the assessment of competences—ideally in an efficient way. Therefore, self-rating questionnaires are very common because they are relatively cheap and easy to administer. This paper will describe seven available questionnaires and provide an overview of their strengths and limitations. In particular, this article will point out the importance of taking the test-development stage into consideration.

The assessment of competences within higher education becomes more and more important. The main impetus is a reform that has already begun, the so-called Bologna Process. This reform focuses, among other things, on educational outcomes in terms of the competences students have gained while attending a single course or a whole study programme in different European countries (Bologna Working Group on Qualifications Frameworks, 2005; European Association for Quality Assurance in Higher Education, 2005).

Higher education reform goes beyond European countries. In a national report written for the US government, Adelman (2008) writes: "Parts of the Bologna Process have already been imitated in Latin America, North Africa, and Australia. The core features of the Bologna Process have sufficient momentum to become the dominant global higher education model within the next two decades" (p. V). He points out that higher education systems are being reconstructed worldwide. One shared goal is to prepare alumni to be flexible and competent participants in a global knowledge-based economy (Organisation for Economic Co-operation and Development [OECD], 2009).

A *competence* is a concept that means a certain level of expertise and capability. A particular competence is generally conceived as involving knowledge, skills, attitudes and predispositions (Weinert, 2001). It can be acquired, developed, or lost.

It can also be understood as a complex arrangement of a person's attributes that are called into play in a variety of situations. In fact, the term *competence* "carries the dual meaning that there is a track record of such achievement (competent performance) and also that the individual has the capability to perform well in the future. It refers to good adaptation and not necessarily to superb achievement" (Masten & Coatsworth, 1998, p. 206).

At the end of a course or programme, competences encompass far more than expert knowledge about a given field of study. They include competences in other key areas such as the occupational, social, and personal domains. The OECD has initiated an international and interdisciplinary research programme called Definition and Selection of Competencies (DeSeCo). The aim of DeSeCo is to define key competences that are "the psychosocial prerequisites for a successful life and a well-functioning society" (Rychen & Salganik, 2003, p. 22) within a shared framework of democratic values and the need for sustainable development. The breadth of this vision is indicated by the three top-level general categories, shown below with examples that have been adopted (adapted from OECD, 2005, pp. 10–14):

Use tools interactively

- Use language, symbols and texts interactively.
- Use knowledge and information interactively.
- Use technology interactively.

Interact in heterogeneous groups

- Relate well to others.
- Co-operate, work in teams.
- Manage and resolve conflicts.

Act autonomously

- Act within the big picture.
- Form and conduct life plans and personal projects.
- Defend and assert rights, interests, limits and needs.

The progress, or the lack thereof, by students towards such competences needs to be measured. The Programme for International Student Assessment (PISA) began by comparing students' knowledge and skills in the domains of reading, mathematics, science and problem solving. PISA intends to assess "knowledge and skills that are essential for full participation in society", and "the domains of reading, mathematical and scientific literacy are covered not merely in terms of mastery of the school curriculum, but in terms of important knowledge and skills needed in adult life" (see http://www.oecd.org/department/0,3355,en_2649_35845621_1_1_1_1_1,00.html; for further information, see PISA, 2005). Nevertheless, even despite this broadening out, the areas under investigation (reading, mathematics, and sciences) are closely related to school subjects, and therefore standardised tests either are available or can certainly be developed.

However, the assessment of student performance in higher education in selected school subjects took place with the understanding that students' success in life depends on a much wider range of competences. The DeSeCo project provides a framework to guide the longer-term extension of assessment into new competency domains and into higher education. While the project notes how certain competences related to cognitive abilities can be assessed in traditional ways, it is acknowledged that the measurement of attitudes and dispositions is also required.

The assessment of these complex types of competence demands expertise in test development. This has led many researchers into the areas of psychometric testing and self-assessment by students, the topic of this article. Therefore, test quality criteria will be described in the following section to recall guidelines of questionnaires' development.

2. Test quality criteria

2.1. Reliability

One fundamental requirement of a psychometric instrument is reliability. This refers to the need for a questionnaire to yield consistent results if used repeatedly under the same conditions with the same participants and therefore to be relatively unaffected by errors of measurement. The most common index is Cronbach's (1951) coefficient alpha. This statistic measures the internal consistency of an instrument as an estimate of its reliability, by comparing the variance of the total score with the variances of the scores on the constituent items.

This measure is generally regarded as a useful indicator of the reliability of a test instrument. However, as Fan and Thompson (2001) pointed out, confidence intervals or other estimates of measurement error should be provided as with any reported statistics. They also noted that a low value of coefficient alpha could mean that the instrument is an unreliable measure of a trait or construct, or it could mean that the instrument is a reliable measure of two or more different traits or constructs. Factor analyses are used to address the second possibility, which is a question of validity.

2.2. Validity

The 1999 Standards for Educational and Psychological Testing define validity as the "degree to which evidence and theory support the interpretation of test scores entailed by proposed uses of tests. . . .The process of validation involves accumulating evidence to provide a sound scientific basis for the proposed score interpretations" (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 1999, p. 9). This mirrors other definitions put forward by leading theorists (e.g., Cronbach & Meehl, 1955; Kane, 2001; Messick, 1989) who, for the last 60 years, have defined validity as a matter of the degree to which a test measures what it is supposed to measure. To assess this, researchers have traditionally relied mostly on correlational approaches. Kane (2001), among others (e.g., Richardson, 2004), divides these approaches largely into three categories: (a) content-based, (b) criterion-based, or (c) construct-based validity—or rather validation, which is the *process* to assess the test *property* of validity.

2.2.1. Content-based validity

Questions or whole measurement instruments can be judged in terms of their content (e.g., item wording). Content or face validity is commonly assessed by judges or raters (e.g., Angoff, 1988), whose judgments can then be intercorrelated to create quantitative approaches of content-based validity (e.g., inter-rater agreement, intraclass coefficient). Such quantitative approaches to content-based validity are seen as superior to qualitative approaches (e.g., expert ratings, where disagreements are resolved by discussion). At the same time, however, they are limited by their reliance on intersubjective agreements, which cannot reach the objectivity that scientists like to claim (cf. Guion, 1977; Messick, 1989).

2.2.2. Criterion-based validity

A more objective approach, in the sense that it does not rely on intersubjectivity, is involved in the concept of criterion (or predictive) validity. Here, the goal is to identify one or more criteria that the test in question should be able to predict, and then to assess the extent to which the test actually predicts the (concurrent) criterion. Again, this is commonly done by correlating test and criterion scores (e.g., Cronbach & Gleser, 1965). The limitation of criterion-based validity is the issue of how far the criterion is valid itself, and therefore of how far it is appropriate to be used in the validation of the test of interest. The suggested solution to this problem—evaluating the validity of the criterion based on content or face validity (e.g., Ebel, 1961)—emphasises the problem rather than solving it.

2.2.3. Construct-based validity

The validation method that most researchers regard as the method of choice is that of so-called construct validity (Cronbach & Meehl, 1955). This approach to validation is more theory-driven than the others. The goal here is to identify other concepts that the concept measured by the test of interest, ideally from a theoretical perspective, *should* be related to, and concepts that the concept of interest *should not* be related to. Once more, this is commonly examined empirically by means of correlations between the test scores and the constructs to which it is supposedly related (*convergent validity*), which should be high, and correlations between the test scores and constructs to which it is supposedly not related (*discriminant validity*), which should be close to zero. In practice, particularly in the social sciences, construct validity is often evaluated by what Cronbach (1988) called a "weak programme of construct validity" (pp. 12–13). In contrast to the "strong programme of construct validity", the validation process is exploratory rather than confirmatory in nature, somewhat betraying the theory-driven pretence with which construct validity had originally been conceptualised. Simple correlation coefficients and exploratory factor analyses are employed in an ad hoc or post hoc fashion, whereas in the "strong programme" correlations, regression, and confirmatory factor analyses are used in a theory-driven fashion.

2.2.4. Recent developments

Evident in the extensive literature on validity and validation briefly reviewed above is a lack of clarity and theoretical rationale, as well as confusion between *validity* and *validation*. This creates the risk that researchers pick and choose what is in their best interest to demonstrate high validity—particularly since validation studies are mostly carried out by the developers of the instrument to be validated themselves (Kane, 2001). This problematic situation has recently been addressed by Borsboom, Mellenbergh, and van Heerden (2004), who clarified the definition of validity and the process of validation. According to Borsboom et al., a "test is valid for measuring an attribute if (a) the attribute exists and (b) variations in the attribute causally produce variation in the measurement outcomes" (p. 1061). This definition diverges from the definition that we gave at the beginning of Section 2.2 in that it turns the question of *the extent to which* a test is valid into the question of *whether or not* a test is valid. This goes back to the roots of the validity literature, which started out with precisely this "either/or" definition (Cattell, 1946; Kelley, 1927).

A test should be constructed based on theoretical notions of what happens between the attribute or concept to be measured and the measurement scores obtained, rather than based on purely empirical considerations (see Borsboom et al., 2004, pp. 1067–1068). Due to the ambiguity of exploratory techniques and correlations in particular, theory must take precedence over empiricism. In our view, only when the data support an a priori hypothesis derived from theory can we be sufficiently sure that they provide evidence for the validity of a test. It is important to note that, from this perspective on validity, reliability becomes all the more important. Once there is a strong theoretical foundation for two different tests

and validation has provided some evidence for the validity of both, the main psychometric property that can distinguish between the quality of these tests is their reliability.

2.2.5. Deciding between validity and validation approaches

Having briefly reviewed different kinds of validity and approaches of validation, it is important for the purposes of this review paper to weigh these and evaluate their merit in terms of establishing validity of an instrument. It is our opinion that, from a philosophy of science perspective, the definition of validity by Borsboom and colleagues (2004) is most appropriate. From this, then, it follows that theoretical considerations should precede purely empirical considerations. With respect to validation, a "strong" programme of construct-based validity is thus the validation approach of choice. While a "weak" programme of construct-based validity, or content- or criterion-based validity, with its exploratory and correlative techniques, can lend preliminary evidence for the validity of an instrument, only the "strong" programme with its confirmatory techniques can fully establish the validity of an instrument. Therefore, validity studies employing confirmatory techniques such as confirmatory factor analysis (CFA) will be weighed more than validity studies employing exploratory or correlative techniques such as exploratory factor analysis (EFA). Furthermore, in line with Borsboom et al. (2004), evidence for the validity of different instruments being equal, the reliability of the instruments will be crucial to distinguish between the (psychometric) quality of these instruments.

2.2.6. Validity of self-ratings

Most academic course-evaluation instruments rely on self-reports by students, particularly those instruments that are most widely used. There has been extensive debate over whether students are able to judge their acquisition of competences or the quality of an academic course (e.g., Greenwald & Gillmore, 1998; Roche & Marsh, 1998). Marsh and Roche (1997) argued that an instrument's construct validity should be demonstrated using several indicators of successful learning. Some evidence for this has been found in research on self-ratings of personality characteristics and performance. Gosling, John, Craik, and Robins (1998) and Spain, Eaton, and Funder (2000) showed that, although the accuracy of self-ratings depended on the person and context, self-ratings of personality characteristics or behavioural dispositions were correlated with alternative measurements of the same constructs. Lucas and Baird (2006) concluded that errors in self-report measures did not severely limit their validity.

As later sections will show, the types of validity that we have described have been employed in research on self-rated competences. Before introducing the questionnaires themselves, we consider some aspects of good questionnaire design, which is particularly important in light of the more recent literature on validity (e.g., Borsboom et al., 2004).

3. Research on questionnaire design

There has been a great deal of well-conducted research on how questionnaires should be designed to minimise extraneous influences on the self-rating process. In particular, Schwarz (1999) and Lucas and Baird (2006), among others (Crocker & Algina, 1986; Krosnick, 1999; Lord & Novick, 1968), have systematically manipulated the design of questionnaires in order to investigate the effects of certain design features. Based on their results, they have provided guidelines on how to design a scientific self-rating questionnaire.

In what follows, certain basic aspects of good questionnaire design will be described. We will focus on the *context* in which a question appears, the way a question is *expressed*, and the kinds of *possible answers*. When presenting the questionnaires that are the focus of this paper, we will report whether or not the questionnaires' authors have empirically investigated the psychometric properties that we introduced earlier. More complex aspects, such as how to create a controlled setting or how to administer a test, will not be considered here (see AERA, APA, & NCME, 1999; Krosnick, 1999).

Filling out a questionnaire is a complex and subtle process (Richardson, 2000, pp. 110, 185). The responses given to standardised questionnaires are communicative and collaborative acts; they are based upon the same principles of everyday communication as responses to an interview (Strack & Schwarz, 1992). The context seems to have a strong influence. When deciding how to interpret items in questionnaires, participants will make use of the immediate context: namely, the content of neighbouring items. If the context is changed, then their responses may well be different (Strack & Schwarz, 1992). In the absence of any additional guidance, respondents will endeavour to make sense of the items in questionnaires (in the case of competences, in terms of their own conceptions of learning and knowledge). Therefore, the context in which a question appears, the way that a question is expressed, and the range of possible answers are very important. Some of these topics may appear trivial, but nevertheless they are regularly overlooked.

3.1. Research intention

Generally speaking, one person in conversation with another will pay attention to the characteristics of the other person. They will then try to provide answers that are of interest to the other person. In the context of questionnaires, this characteristic is the participants' hypotheses about the survey designer's research intentions.

Norenzayan and Schwarz (1999) presented respondents with newspaper accounts of mass murders and asked them to explain why the mass murder occurred. In one condition, the questionnaire was printed on the letterhead of an alleged "Institute for Personality Research", whereas in the other condition it was printed on the letterhead of an "Institute for Social Research". As expected, the respondents' explanations showed more attention to personality variables or to social-

contextual variables, depending on whether they thought the researcher was a personality psychologist or a social scientist. Apparently, they took the researcher's affiliation into account in determining the kind of information that would be most informative, given the researcher's likely epistemic interest. Consequently, no research intention or hypothesis should be communicated in a survey (Schwarz, 1999).

3.2. Order of demographic information question

Questionnaire developers should also consider the order in which questions appear. For example, if one first mentions a well-known person (e.g., Barack Obama) who belongs to a specific group (the US Democratic Party) and then asks for the attitude of the respondents to this group, the answers given will be different from those given if the order had been reversed (Schwarz, 1999). This becomes especially important if demographic information is requested. Asking for ethnicity or gender might activate the respondent's affiliation to a particular social group (Steele & Aronson, 1995). Questions that are presented afterwards will be answered more in line with the way in which the social group is expected to answer (Spencer, Steele, & Quinn, 1999). Therefore, only essential biographic questions should be included, and they should appear at the end of the questionnaire.

3.3. Social desirability

Social desirability refers to survey respondents' tendency to give answers that serve the goal of impression management: to produce a positive image of themselves for others. To avoid social desirability issues, Schwarz (1999) and Lucas and Baird (2006) advise that the anonymity of the survey needs to be stressed and that questions should not be too personal. Joinson's (2001) research demonstrated that self-disclosure was greater in computer-mediated interactions than in face-to-face interviews. Moreover, his study suggested that assured or assumed anonymity was the key element in people's willingness to provide personal information.

3.4. Question wording

It is easier for people to respond concerning characteristics that are central to their self and that have been evoked recently (Schwarz, 1999). Asking about a particular situation rather than a general one will prime valid memories (Hannover, 2000; Noftle & Fleeson, 2010). Therefore, questions should be concerned with specific behaviour in an authentic situation. The period of time in which the behaviour occurs should be short and clearly defined. If you want to measure whether or not a person goes to church, for instance, it is better to ask "Have you been to Mass in the last two weeks?" than "Do you attend Mass regularly?"

Questions should be worded precisely and only contain terms that are familiar to the survey population. Vague terms, by which we mean those that are ambiguous, unclear, or abstract, should be avoided. In addition, there should be no "double-barrelled items" (i.e., items that involve more than one statement or question), because respondents may not know how to answer if only one characteristic applies to them while the other does not (Kalton, Collins, & Brook, 1978; Schuman & Presser, 1981; Schwarz, 1999).

3.5. The time period concerned

If people are asked to rate their competences retrospectively (e.g., students rating their competence level at the beginning of the semester), they will compare their previous competences with changes in competences since then. If alumni who graduated 5 years ago are asked to rate their competences at the end of their study programme, they will do so with the knowledge of the level of competences required in their actual vocational setting. Imagine two people who both graduated with the same level of competences from higher education some time ago. One obtained a challenging managerial position, whereas the other is in a routine mundane job. It is very likely that the person with the higher task demands will rate their level of competences at graduation as lower, since they have increased their knowledge and skills more than the other person (Schwarz, 1999). We return to this issue in the following section when considering the role of implicit theories in how people judge personal change.

3.6. Response alternatives

Different frequency scales convey different meanings to respondents. Schwarz's (1999) hypothesis, which his research findings supported, was:

> Suppose that respondents are asked how frequently they felt "really irritated" recently. To provide an informative answer, respondents have to determine what the researcher means with "really irritated". Does this term refer to major or to minor annoyances? To identify the intended meaning of the question, they may consult the response alternatives provided by the researcher. If the response alternatives present low-frequency categories, for example, ranging from "less than once a year" to "more than once a month", respondents may conclude that the researcher has relatively rare events

in mind. Hence, the question cannot refer to minor irritations that are likely to occur more often, so the researcher is probably interested in more severe episodes of irritation. (p. 95)

Frequency scales are more vulnerable to such interpretations than are scales of agreement. Consequently a questionnaire designer should consider using fewer frequency items (how often) and more agreement items (I fully dis-/agree).

Similar processes are involved when retrospective estimations are asked for. People who are asked about behaviour in the last week remember unimportant and important events equally. People who are asked about behaviour in the last year recall more important events (Schwarz, 1999). Richardson (2000) raises another issue on this topic: "Social psychologists have shown that people sometimes denigrate their past capabilities in order to fit their own implicit theories about personal change, and this can certainly occur when students are asked to assess the value of recent educational experiences" (p. 42).

Conway and Ross (1984) selected students who wanted to take a course on improving their study skills and asked them to evaluate their own study skills. They randomly assigned them either to a group that took the course or to a waiting list. Afterwards, they asked both groups to evaluate their study skills at present and in retrospect, at the outset of the study. After the course, the two groups gave similar ratings of their own study skills, and so there was no evidence that the course had had any effect whatsoever. But the group who had taken the course produced lower retrospective ratings of their study skills before the course. In other words, they felt their study skills had improved as a result of taking the course, but the only way that they could demonstrate this was to denigrate their previous study skills. Thus, theories of personal change have to be kept in mind when retrospective questions are used (see also Ross, 1989).

3.7. Numerical response scales

Lucas and Baird (2006) suggest using only positive numerical values within response scales. If scales incorporate both negative and positive values, respondents are less likely to choose response values at the lower end of the scale. Negative values seem to lead respondents to a certain interpretation.

For example, if a questionnaire asks about competences, and the response scale ranges from "–2" to "+2", people interpret the endpoints of the response scale as "very incompetent" and "very competent", respectively—a bipolar construct is assumed. On the other hand, if the response scale ranges from "+1" to "+5", people interpret the scale endpoints as "less competent" and "very competent", respectively. While people will avoid estimating someone as incompetent, they are willing to evaluate someone as less competent. Respondents will use the full range of the scale if only positive values are offered.

Printing numbers on a response scale helps to generate an approximately normal distribution of responses (Schwarz, 1999). However, to avoid any struggle with numbers, including positive or negative values, other researchers (Krosnick, 1999; Richardson, 2004) recommend using verbal descriptors.

3.8. Summary of research on questionnaire design

As Strack and Schwarz (1992) demonstrated, responses to questionnaires are communicative and collaborative acts. In the absence of explicit feedback, respondents will use cues that allow them to make pragmatic inferences about the intended meaning of the questions and potential answers. In order to generate valid and reliable answers from the respondents, both the questions and the response alternatives should be clear, unambiguous, and easily cognitively accessible to the respondent. The features described above are intended to facilitate this process.

4. The seven questionnaires

In this section we describe seven questionnaires aimed at capturing students' self-rated competences in higher education. (The questionnaires are listed in the Appendix together with relevant websites.) We will introduce publicly available instruments, which we found via literature research or through our attendance at scientific conferences. All questionnaires are influential for one of the following reasons: they are administered nationally (CEQ[1], NSSE), or they are pioneers within a certain context (HEsaCom: innovative orientation on competences; PEDI: unique focus on long-distance learning), or they have a wide application (CIRP, CSEQ, SIR II). They are discussed in alphabetical order. In each case, we begin with a short overview of the aims of the questionnaire and a description of the sample(s) used to verify it. It is then evaluated against the design features outlined in Section 3 and against the kinds of reliability and validity listed in Section 2. Finally, we provide an overall evaluation of all seven instruments, together with a summary of their characteristics against these criteria.

4.1. The College Student Experiences Questionnaire (CSEQ)

The CSEQ has a long history. It was first published in the US in 1958 by Pace and Stern, and is now in its fourth edition (Gonyea, Kish, Kuh, Muthiah, & Thomas, 2003). The CSEQ asks students about their college activities, the institution's environment, and biographical information. It was founded on the theoretical basis of student engagement: "The more effort

[1] At this stage we use abbreviations for reasons of easy-reading. All questionnaires will be introduced with full name later on.

students expend in using the resources and opportunities an institution provides for their learning and development, the more they benefit" (Gonyea et al., 2003, p. 4).

Of most interest to us are the 25 items in the *Estimate of Gains* section. Here, students are asked to consider their college experience thus far and to estimate the extent to which they have gained or made progress in several areas. Their answers can be "Very much", "Quite a bit", "Some" or "Very little". The research group obtained responses from over 100,000 students and carried out an exploratory factor analysis on their responses. The following scales emerged:

- *Personal/Social Development* (coefficient alpha = .83)
  (1) Developing your own values and ethical standards.
  (2) Understanding yourself, your abilities, interests, and personality.
  (3) Developing the ability to get along with different kinds of people.
  (4) Developing the ability to function as a member of a team.
  (5) Learning to adapt to change (new technologies, different jobs or personal circumstances, etc.).
  (6) Developing good health habits and physical fitness.
- *Science and Technology* (coefficient alpha = .87)
  (1) Understanding the nature of science and experimentation.
  (2) Understanding new developments in science and technology.
  (3) Becoming aware of the consequences (benefits, hazards, dangers) of new applications of science and technology.
  (4) Analysing quantitative problems (understanding probabilities, proportions, etc.).
- *General Education* (coefficient alpha = .81)
  (1) Developing an understanding and enjoyment of art, music, and drama.
  (2) Broadening your acquaintance with and enjoyment of literature.
  (3) Seeing the importance of history for understanding the present as well as the past.
  (4) Gaining knowledge about other parts of the world and other people (Asia, Africa, South America, etc.).
  (5) Becoming aware of different philosophies, cultures, and ways of life.
  (6) Gaining a broad general education about different fields of knowledge.
- *Vocational Preparation* (coefficient alpha = .78)
  (1) Acquiring knowledge and skills applicable to a specific job or type of work (vocational preparation).
  (2) Acquiring background and specialisation for further education in a professional, scientific, or scholarly field.
  (3) Gaining a range of information that may be relevant to a career.
- *Intellectual Skills* (coefficient alpha = .82)
  (1) Writing clearly and effectively.
  (2) Presenting ideas and information effectively when speaking to others.
  (3) Using computers and other information technologies.
  (4) Thinking analytically and logically.
  (5) Putting ideas together, seeing relationships, similarities, and differences between ideas.
  (6) Learning on your own, pursuing ideas, and finding information you need.

4.1.1. Demographics

The respondents' anonymity is assured, and so biases of social expectancy should be minimised. However, the whole of the first page of the CSEQ is devoted to personal information; respondents are even asked to report their parents' education.

4.1.2. Wording

The wording of most items in the CSEQ is quite lengthy. The authors appear to have expanded the items in a quest for clarity, but the result is that almost all items contain two or more questions. There is no use of specific situations. The period of time is rather general, stretching from entry to college "up to now" and could cover a long period. This questionnaire does not reveal any research intention, and so the respondents are not influenced by any expectations. Some items are vague (e.g., "Gaining a range of information that may be relevant to a career"), and others are rather abstract (e.g., "Learning to adapt to change"). Some include several options for optimising (e.g., "becoming aware of different philosophies, cultures, and ways of life"). We conclude that the wording of the CSEQ could be easily improved.

4.1.3. Response alternatives

There are no numerical scales for the competence items. All of the possible answers are verbally labelled and there are only positive values, varying from "Very little" to "very much", which is as recommended by Lucas and Baird (2006).

4.1.4. Psychometric properties

The content validity of the CSEQ is derived from theoretical research on the engagement of students. The CSEQ was the first scientific questionnaire to evaluate the student experience of higher education. It has endured, and it has influenced those that have followed. The value of the CSEQ to institutions has been shown by its extensive and long term use.

4.2. The Cooperative Institutional Research Program (CIRP)

The CIRP is a long-standing research group in the United States that is interested in the general personal development of students. The group surveys students at the beginning and again towards the end of their study programmes. The survey questionnaire covers a variety of topics, such as family origins, private and study activities, values and beliefs, and, of interest in the present context, self-rated competences (see http://www.heri.ucla.edu/abtcirp.php).

In a block of 18 items students are asked to rate their current level of competence in comparison to an average person of their age. Here, the five points on the response scale are labelled "Highest 10%", "Above average", "Average", "Below average" and "Lowest 10%". In another block of 19 items students are ask to compare their current skill levels with when they first entered the college, using a different 5-point scale: "Much stronger", "Stronger", "No change", "Weaker" and "Much weaker".

The groupings of items rated against the average person were:

- *Social Self-Concept* (coefficient alpha = .71)
  (1) Leadership ability.
  (2) Public speaking ability.
  (3) Self-confidence (social).
- *Emotional and Interpersonal Self-Concept* (coefficient alpha = .67)
  (1) Self-understanding.
  (2) Understanding of others.
  (3) Emotional health.
  (4) Cooperativeness.
- *Self-Assessed Academic Motivation* (coefficient alpha = .56)
  (1) Self-confidence (intellectual).
  (2) Drive to achieve.
  (3) Writing ability.
- *Respect for Diverse Perspectives* (coefficient alpha = .85)
  (1) Tolerance of others with different beliefs.
  (2) Ability to work cooperatively with diverse people.
  (3) Openness to having my own views challenged.
  (4) Ability to discuss and negotiate controversial issues.
  (5) Ability to see the world from someone else's perspective.

Those items that measured gains in competences were grouped as follows:

- *Self-Assessed Cognitive Development* (coefficient alpha = .77)
  (1) Critical thinking skills.
  (2) Analytical and problem-solving skills.
  (3) Ability to work as part of a team.
  (4) General knowledge.
  (5) Ability to conduct research.
  (6) Knowledge of a particular field or discipline.
  (7) Knowledge of people from different races/cultures.
- *Informed Citizenship* (coefficient alpha = .80).
  (1) Understanding of national issues.
  (2) Understanding of global issues.
  (3) Understanding of the problems facing your community.

4.2.1. Demographics

At the end of the survey, the questionnaire asks about private information such as political views, ethnicity, native language, and gender. Also names and e-mail addresses are required at the very beginning of the questionnaire. Even fairly insensitive people might feel threatened by the lack of anonymity and consequently give less valid information.

4.2.2. Wording

Recalling the test criteria from Section 3, none of the items is clearly and precisely formulated (e.g., "cooperativeness"), and abstract terms are often used ("global issues", "critical thinking", etc.). The wording of competence items tends to be open to individual interpretation; for instance, "Leadership ability", "Self-confidence (social)", "Emotional health", "Understanding of national issues". The items themselves are single words or phrases and not whole sentences. Consequently, formulation in the first-person singular is not possible. On a positive note, there are no double-barrelled items, but all the items lack reference to a specific behaviour in an authentic situation. The questionnaire does not declare any research agenda.

4.2.3. Response alternatives

Students are asked to compare their competences with the average person of their age, but it is not clear how they could be aware of the national distribution of such competences. They are also asked to compare their competences over time. However, Schwarz (1999) suggests that one should use response scales expressing level of agreement. The two lists of competences are different but overlap to some extent. No logic is discernable regarding which items are repeated. Potential answers are labelled with words, which helps students to interpret the response scale.

4.2.4. Psychometric properties

The items used are wide-ranging and cover many different areas. However, the items and the scales seem to suffer from a lack of theory. Using exploratory factor analysis on data from a sample of 31,500 students, Liu, Sharkness, and Pryor (2008) showed that certain items tended to be interrelated and could be grouped together in constructs that could be logically labelled. Given the exploratory nature of the factor analysis, this can only be seen as preliminary indication for construct validity, and a more theory-driven approach would be desirable to strengthen this claim. Despite its psychometric limitations, the fact that the CIRP questionnaire is widely employed suggests that it is generally found to be useful.

4.3. The Course Experience Questionnaire (CEQ)

The CEQ (Ramsden, 1991; Wilson, Lizzio, & Ramsden, 1997) was devised in Australia, where it is distributed annually to all new university graduates as part of a wider survey into their current employment situation. It has also been adapted in several countries for the evaluation of both courses and programmes in higher education. In the original version of the questionnaire, the CEQ consisted of five scales on the student experience of higher education that were devised according to classical test theory as well as item-response theory. Subsequently, additional scales were developed, including one called "Generic Skills" to obtain self-ratings of the competences acquired in higher education. This scale is the focus of our interest, and it consists of six items:

(1) The course developed my problem-solving skills.
(2) The course sharpened my analytic skills.
(3) The course helped me develop my ability to work as a team member.
(4) As a result of my course, I feel confident about tackling unfamiliar problems.
(5) The course improved my skills in written communication.
(6) My course helped me to develop the ability to plan my own work.

4.3.1. Demographics

In the Australian survey questionnaire, some fields are designated for "office use only". This might put off some sensitive students, because a lack of transparency of any kind might arouse suspicions. An additional criticism is that biographical information is requested at the beginning of the questionnaire and takes up a whole page. Also, anonymity is not guaranteed because the name and address of the respondent is requested.

4.3.2. Wording

The competence items are clearly worded and there are no double-barrelled items. However, abstract concepts such as "problem-solving" and "analytical skills" are used. If the survey is administered some time after graduation, there may be memory problems. The CEQ does not state any research intention. Students are just asked, "Please tell us about your course experience," and so the questions could not be described as leading the respondent. Questions include specific behaviour in an authentic situation, namely the course or programme. From a design perspective, it is interesting that the original CEQ contained 15 positively scored items and 15 negatively scored items, apparently to control for acquiescent response bias. Even so, all of the items in the generic skills scale are positively worded and positively scored.

4.3.3. Response alternatives

A response scale from "Strongly disagree" to "Strongly agree" is used, in line with our earlier recommendations. The answers are not given numerical values, and only the two scale endpoints are labelled, so neither positive values nor negative values are used.

4.3.4. Psychometric properties

Martin (1996) analysed the responses given to the CEQ by 127 students and reported a coefficient alpha of .81 for the generic skills scale. This implies a good level of consistency.

Investigations of criterion validity (see Section 2) have also reported empirical evidence for theoretically expected relationships (Wilson et al., 1997). Self-rated *generic skills* correlated with course grades ($r = .23$). The authors also assumed that students who were interested in acquiring a deep understanding of the topic of the course ("deep learning approach") would gain more generic skills than students who were not interested in a deep understanding. Conversely, students who were motivated to reach the goal of the course without spending much time on learning ("surface learning approach") would gain fewer generic skills than those who showed less evidence of a surface learning approach. In accordance with these

assumptions, the authors found a correlation coefficient of .37 between generic skills and deep learning approach and a correlation coefficient of –.20 between generic skills and surface learning approach. Different versions of the CEQ contain varying numbers of items, but it would seem that the scale is stable and robust for scientific purposes.

4.4. Evaluation in Higher Education: Self-Assessed Competences (HEsaCom)

The HEsaCom (Braun, 2007; Braun & Leidner, 2009) asks students about gains in competences as a result of a module that they studied. During the development of the HEsaCom, the developers attempted to follow the principles of good questionnaire design outlined earlier (Sections 2 and 3). For example, during development, the items were arranged randomly on the questionnaire to avoid possible item order effects. The questionnaire is published in an academic journal and is therefore available to everyone.

Institutions can also purchase the questionnaire from the company Electric Paper Services (see http://www.electricpaper.biz/information/newsarchiv/singlenews/article/290/hesacom-eval.html) along with technical support. This questionnaire has been used regularly at the Freie Universität Berlin and several other institutions since 2003. Over 25,000 students have completed the HEsaCom. The survey can be administered on paper or online.

In total there are 35 items that can be theoretically and empirically grouped into six areas of competences based on theoretical concepts found in the higher education research literature. Three of the six areas begin with a "filter item": students only answer the subsequent questions if they agree with the filter item. The empirical factor structure of the data is consistent with the theoretical model of six factors, as demonstrated by confirmatory factor analysis:

- *Knowledge Processing* (coefficient alpha = .94)
  (1) As a result of this course, I can remember most of the important terms and facts from this course.
  (2) As a result of this course, I can give an overview of the course.
  (3) The course has helped me improve my analysis of complex issues in this subject area.
  (4) This course has helped me improve my handling of typical problems in this subject area.
  (5) This course has helped me both to see the connections and to notice inconsistencies in this subject area.
  (6) This course has helped me judge the quality of academic articles in this subject area.
- *Systematic Competence* (coefficient alpha = .92)
  (1) This course has helped me to acquire information more efficiently.
  (2) This course has helped me organise my work.
  (3) This course has helped me improve the way I work.
- *Presentational Competence* (coefficient alpha = .93)

  FILTER: I gave an oral presentation as part of this course.

  (1) After presenting in this course I feel I can engage better with the audience.
  (2) After presenting in this course I can structure my talks better.
- *Communication Competence* (coefficient alpha = .95)

  FILTER: I regularly spoke in this course.

  (1) This course has helped me express my opinion.
  (2) This course has helped me to ask for clarification when I have difficulty understanding.
  (3) This course has helped me speak in a way that others can understand.
  (4) This course has helped me speak more precisely.
  (5) This course has helped me to improve the way I moderate discussions.
- *Cooperation Competence* (coefficient alpha = .91)

  FILTER: I've worked with other students in a work group for more than two weeks during this course.

  (1) My participation in the group work made it easier for me to help delegate tasks.
  (2) My participation in the work group made it easier for me to know when to hold back from contributing.
  (3) My participation in the work group made it easier for me to stand up for constructive team spirit.
  (4) My participation in the work group helped me take personal responsibility for my share of the work.
  (5) I feel identified with our work group's accomplishment.
- *Personal Competence* (coefficient alpha = .92)
  (1) I have grown more interested in the subject matter as the course has progressed.
  (2) The course encouraged me to continue my studies.
  (3) The course has increased my joy of carrying out assigned tasks.

(4) I feel more inspired by the topics studied in this course than at the beginning.
(5) The course has inspired me to study the subject further in my own time.

4.4.1. Demographics

No demographic or private information is requested, so anonymity is guaranteed.

4.4.2. Wording

Recalling the guidelines on test design, no research intention is mentioned. Each item is a single statement and is to be answered with regard to the student's own behaviour in one specific course. The questionnaire contains a good many items and is relatively long for six competence areas. Some items are not very specific and seem to be rather abstract: for example "My participation in the work group helped me take personal responsibility for my share of the work", "This course has helped me express my opinion". Students are asked about gains in competences "as a result of this course. . .", but some gains will not be only due to the course itself, if at all. Students may find it hard to make such distinctions.

4.4.3. Response alternatives

The points on the response scale range from strong agreement to strong disagreement. In the online surveys, the points are labelled only with the positive values 1–5.

4.4.4. Psychometric properties

The areas in the HEsaCom were developed on a theoretical basis from a consideration of competence dimensions. The theoretical structure has been supported by the confirmatory factor analyses and is a strong indication of the instrument's construct validity. There have been a number of validity studies that demonstrated the HEsaCom's robustness. In one investigation of criterion validity, independent observers rated the learning environment of a module. The idea was that productive discussion involving most of the students should lead to higher gains in the communication competence of students. It turned out that the empirical correlations between the observers' ratings and the six competences estimated in the HEsaCom were indeed significant (multi-level regression: γ_{01} = .12–.52: Braun & Hannover, 2011).

Additionally, the approaches to teaching (Prosser & Trigwell, 2006) of the lecturers have been surveyed. In courses where the lecturers adopt a more student-centred approach to teaching, the students report higher gains in all six areas of competence (Braun & Hannover, 2008).

4.5. The National Survey of Student Engagement (NSSE)

The NSSE was developed by a team sponsored by the Pew Charitable Trusts in the United States and managed by the Indiana University Center for Postsecondary Research (Kuh, 2009; Kuh, Cruce, Shoup, Kinzie, & Gonyea, 2008). First-year and final-year undergraduate students are asked about five aspects of their experience: their participation in educationally purposeful activities; the requirements of their institution; their perceptions of the academic environment; personal and demographic characteristics; and their educational and personal growth since their admission to university. In 2010, responses were obtained from more than 362,000 students at nearly 600 participating institutions. Aggregated data are published in annual reports. Institutional data are returned to the participating institutions, which are encouraged to publish key indicators on the website of the newspaper *USA Today* and elsewhere.

Some questions in the NSSE are based on those in the CSEQ. The theoretical background of both instruments is research on "student experiences and engagement". The term *engagement* is used to refer to "the quality of effort students themselves devote to educationally purposeful activities that contributes directly to desired outcomes" (Hu & Kuh, 2002, p. 555). Kuh (2003) emphasised that the adoption of the NSSE was itself an intervention promoting engagement: by filling out the questionnaire, students can communicate their perspective and will be involved.

The section concerning self-rated competences consists of 16 items. Students are asked: "To what extent has your experience at this institution contributed to your knowledge, skills and personal development in the following areas?" The items concern, for instance, "Acquiring a broad general education", and responses are made on a four point scale: "Very much", "Quite a bit", "Some", "Very little". The three scales of self-reported competences in the NSSE and their reliability are as follows:

- *Gains in General Education* (coefficient alpha = .85)
  (1) Writing clearly and effectively.
  (2) Speaking clearly and effectively.
  (3) Acquiring a broad general education.
  (4) Thinking critically and analytically.
- *Gains in Practical Competence* (coefficient alpha = .83)
  (1) Acquiring job or work-related knowledge and skills.
  (2) Working effectively with others.
  (3) Using computing and information technology.
  (4) Analysing quantitative problems.
  (5) Solving complex real-world problems.

- *Gains in Personal and Social Development* (coefficient alpha = .88)
  (1) Developing a personal code of values and ethics.
  (2) Understanding yourself.
  (3) Understanding people of other racial and ethnic backgrounds.
  (4) Voting in local, state, or national elections.
  (5) Learning effectively on your own.
  (6) Contributing to the welfare of your community.
  (7) Developing a deepened sense of spirituality.

4.5.1. Demographics

Demographic information is required on the last page of the instrument but anonymity is assured. Hence, the NSSE is not leading the respondents in any directions that might promote socially desirable responses.

4.5.2. Wording

Some of the items could be more precise, as in the case of "Acquiring a broad general education", "Solving complex real-world problems" and "Contributing to the welfare of your community". Also, a few items include more than one question, as can be seen with "Speaking clearly and effectively", "Thinking critically and analytically". Specific behaviours in authentic situations are not used. The questionnaire does not declare any research intention.

4.5.3. Response alternatives

On the positive side, the response scales are all expressions of agreement or disagreement, and all of the alternatives are labelled with words.

4.5.4. Psychometric properties

The validity of the questionnaire has been extensively researched in impressive scientific studies. In the NSSE manual (Kuh, 2003), a weak but significant correlation between the scores on the scale "Gains in general education" and Grade Point Average ($r = .16$) is reported, which can be interpreted as evidence of criterion validity. Furthermore, in regression analyses, Pike (2006) found that institutional characteristics taken from the Integrated Postsecondary Education Data System, such as percent of full-time students, and learning environment (e.g., "academic challenges", "cooperative learning methods") could explain up to 80% of the scale variance in gains in general education, which is a good indicator of the criterion validity of the NSSE.

Competence categories have been emerged using exploratory factor analysis, providing a preliminary indicator of construct validity, and the questionnaire has been widely used. Although the wording of the items could be improved by following more closely the guidelines for item development above, the theoretical background is undoubtedly strong. The NSSE has recently been adapted to yield the Australasian Survey of Student Engagement (Coates & Edwards, 2009) and the Beginning College Survey of Student Engagement (Cole & Gonyea, 2010).

4.6. The Personal and Educational Development Inventory (PEDI)

The PEDI was developed at the UK Open University and has been described in an article in an academic journal (Lawless & Richardson, 2004). The aim of the questionnaire is to monitor students' experiences for both curriculum development and quality assurance purposes. It was devised for use in distance education but can also be used on campus-based programmes.

The PEDI contains 26 items which are grouped into four empirical dimensions derived by exploratory factor analyses. The items contain short expressions, like "Critical analysis" or "Presentation skills". Respondents were asked to rate the extent to which their studies had enabled them to develop in each item along a 4-point scale from "Not at all" to "A great deal". The initial sample consisted of 3118 recent graduates from the Open University.

- *Cognitive Skills* (coefficient alpha = .83)
  (1) Critical analysis.
  (2) Evaluation skills.
  (3) Ability to apply knowledge.
  (4) Acquire specialist knowledge.
  (5) An understanding of new concepts.
  (6) Research skills.
  (7) Problem-solving skills.
  (8) Writing skills.
  (9) Reflective skills.
  (10) Desire to go on learning.
- *Mathematical Skills* (coefficient alpha = .82)
  (1) Ability to use numerical data.
  (2) Ability to analyse numerical data.

(3) Information management.
(4) Computer literacy.

- *Social Skills* (coefficient alpha = .83)
  (1) Leadership skills.
  (2) Interpersonal skills.
  (3) Ability to work in teams.
  (4) Entrepreneurial skills.
  (5) Presentation skills.
- *Self-Organisation* (coefficient alpha = .85)
  (1) Self-discipline.
  (2) Self-reliance.
  (3) Time management.
  (4) Ability to prioritise tasks.
  (5) Independent learning.
  (6) Self-confidence.
  (7) Awareness of own strengths and weaknesses.

4.6.1. Demographics

Demographic information is required on the last page of the instrument.

4.6.2. Wording

Regarding the item phrasing, the items are mostly abstract (''Self-confidence'', ''Leadership skills''). The items themselves are single words or phrases and not whole sentences. On a positive note, there are no double-barrelled items.

4.6.3. Response alternatives

All possible answers are labelled with words. The respondents indicate how much they developed each skill.

4.6.4. Psychometric properties

The items were based on an existing list of competences and not on any particular theory. Exploratory factor analysis was used to identify discrete scales whose reliability is supported by the values of coefficient alpha (Lawless & Richardson, 2004).

Various aspects of validity are tested in empirical research. As theoretically expected, graduates from diverse faculties at the UK Open University rated the extension of personal development differently, which supports the PEDI's discriminant validity: students of Natural Science, Technology and Mathematical Sciences reported the highest scores in mathematical skills, whereas those taking Literature, Humanities (Art History) and Law obtained the lowest scores in mathematical skills. Furthermore, graduates who had achieved a better class of degree produced higher scores on cognitive skills, which supports the PEDI's criterion validity. Finally, there is a substantial overlap in variance between graduates' scores on the PEDI and their scores on the CEQ, providing evidence of convergent validity. In particular, graduates who reported higher levels of personal development also reported more positive perceptions of the course materials and the tutorial support (Lawless & Richardson, 2004).

The PEDI was subsequently used by Edmunds and Richardson (2009) in surveys of students in 15 departments of bioscience, business studies and sociology across the United Kingdom (five departments in each discipline). They found that students' scores on the PEDI were related to their conceptions of learning and their approaches to studying. They also found the ''computer literacy'' loaded on the ''Social Skills'' factor, not on the ''Mathematical Skills'' factor. This suggested that, by the mid 2000s, British students regarded computers as devices for communication (by email and text-messaging) rather than as devices for computation.

The PEDI is a relatively short instrument intended to provide an insight into important dimensions of students' development. The research on the validity of the PEDI is promising, and it deserves to be evaluated in further scientific investigations.

4.7. The Student Instructional Report (SIR)

The first version of the SIR, which evaluates academic courses, was developed by Centra (1979). Currently, the revised version (SIR II) is sold by the Educational Testing Service (ETS) and is available for online or paper administration. The sample size for standardisation is large and involves tens of thousands of students. SIR II includes a section called ''Course Outcomes''. This contains five items, and the response alternatives are ''Much more than most courses'', ''More than most courses'', ''About the same as others'', ''Less than most courses'' and ''Much less than most courses''.

- *Course Outcome* (coefficient alpha = .96)
  (1) My learning increased in this course.
  (2) I made progress toward achieving course objectives.
  (3) My interest in the subject area has increased.
  (4) This course helped me to think independently about the subject matter.
  (5) This course actively involved me in what I was learning.

4.7.1. Demographics

The SIR II assesses personal and demographic information at the end of the survey. This section of the questionnaire routinely covers the student's gender, the student's year in college (e.g., freshman, sophomore), language proficiency, the student's Grade Point Average, and whether the class is compulsory or optional. The questionnaire itself contains no introductory explanation. The ETS recommends that instructors assure students of full anonymity before handing out the questionnaire forms.

4.7.2. Wording

The wording of the items is in accord with accepted standards of questionnaire design (Section 3). The five items are generally well expressed with no abstract words and no double-barrelled items, and the focus is on a particular course. However, certain items such as "This course helped me to think independently about the subject matter" and "I made progress toward achieving course objectives" could be more closely linked to actual behaviour.

4.7.3. Response alternatives

All possible answers have verbal labels. Furthermore, the items are formulated as agreements.

4.7.4. Psychometric properties

An exploratory factor analysis of responses to the whole questionnaire given by students taking 1200 courses revealed eight factors (Centra, 1998). However, the questionnaire appears to lack a clear theoretical underpinning, and so it is not clear whether these factors were in line with the designer's intentions. To our and ETS's knowledge there is one study that specifically assessed the validity of the SIR II. It is published as a report available on the ETS website (http://www.ets.org/sir_ii/about/research). This study found student ratings of instruction to be predictive of student perceptions of learning, indicating criterion-based validity of the SIR II (Centra & Gaubatz, 2005). In addition, norms based on large numbers of students are available, and these allow comparative evaluations to be undertaken.

5. Conclusions

Our aim in this article was to provide an overview of existing questionnaires on self-rated competences. We have introduced a number of test-design criteria to evaluate the questionnaires, and our conclusions are summarised in Table 1. On balance, we feel that each questionnaire has specific strengths and weaknesses. The interested researcher or practitioner may be able to choose one questionnaire that is appropriate to their specific purpose, or they might wish to employ our criteria as guidelines for designing their own questionnaire.

Questionnaires asking for students' self-reports on their competences are now in common use, reflecting the new focus on the outcomes of higher education. This is probably a worldwide phenomenon, but we have focused on seven questionnaires from Europe, the United States, and Australia, because of the longer history of such surveys in these countries and their availability in the English language. The United States was the first in the field, but, as Adelman (2008) points out, even this country might well be able to learn from the European Bologna Reform.

Comparing the seven questionnaires using our criteria on instrument development reveals some noteworthy findings. First, there is evidence of the validity of self-reported competences:

- Evidence of criterion validity has been shown in the case of the CEQ, the HEsaCom, the NSSE, the PEDI and the SIR.
- The CSEQ has demonstrated a level of content validity.
- The HEsaCom has shown construct validity.
- The PEDI has shown convergent validity.

The correlation coefficients between self-rated competences and grades are somewhat low. However, even if the absolute correlations are low, they are of similar magnitude to those found in other research. For instance, Mabe and West (1982) found in a meta-analysis a mean correlation of .29 between self-evaluations of ability and other measures of performance. Meyer et al. (2001, p. 133) noted with regard to self-ratings and other measurements "how challenging it is to consistently achieve uncorrected univariate correlations that are much above .30". The findings of Gosling et al. (1998, p. 1346) were similar (a mean correlation between observer ratings and self-ratings of behavioural acts of .19).

Nevertheless, the arguments of Borsboom et al. (2004) that were mentioned earlier might be helpful in the context of self-ratings in general. Performance on a test can be strongly influenced by other variables, and therefore the advantage and significance of a test is strongly dependent on the context. Yet the validity of a specific test is not threatened by possible context variables. According to Borsboom et al., a test is valid if it shows empirical results that are consistent with one's hypotheses (for example, plausible differences between subgroups of students), independent of potential biases or systematic errors.

Personal information is frequently requested in these questionnaires. This might be attributed to the need to control for sample bias or to provide breakdowns of the results. However, some questionnaires dispense with this information, so it might not be necessary in all cases. The possibility should be considered that by asking these questions there are unwanted

Table 1
Using test-design criteria to compare seven questionnaires on self-rated competences.

Criterion	CSEQ	CIRP	CEQ	HEsaCom	NSSE	PEDI	SIR II
Social desirability							
Demographic and private information should be surveyed as little as possible.	–	–	++	++	+	+	+
Demographic information should be requested at the end of the questionnaire.	–	++	–	++	++	++	++
Anonymity should be assured.	++	–	–	++	++	++	++
Wording of questions							
Vague terms should be avoided.	+	–	+	++	–	–	++
Retrospective estimation should be avoided.	+	+	+	++	–	++	++
Double-barrelled questions should be avoided.	–	–	++	++	+	++	++
Questions should concern specific behaviour in an authentic situation.	–	–	++	++	–	–	++
Response alternatives							
Use agreement instead of frequency scales.	++	+	++	++	+	++	+
Number the response alternatives.	Verbal labels are used	+	–	+	Verbal labels used	–	Verbal labels used
Use only positive numbers to label the responses.	Verbal labels are used	Verbal labels used	–	+	Verbal labels used	Verbal labels used	Verbal labels used
Psychometric properties							
Reliability (coefficient alpha)	Better than .80	Varies from .60 to better than .80	Better than .80	Better than .90	Better than .80	Better than .80	Better than .90
Content-based/theory-based validity	Based on student engagement theory	–	Based on student learning research	Theoretical hypotheses confirmed	Based on student engagement theory	Theoretical hypotheses confirmed	Theoretical hypotheses confirmed
Criterion-based validity	–	–	++	++	++	++	++
Construct-based validity	–	EFA	EFA	CFA	EFA	EFA	EFA

Note: ++, criterion met; +, criterion partly met; –, criterion not met. CFA, confirmatory factor analysis; EFA, exploratory factor analysis.

effects on the self-rating of competences. Even if this information is essential, it would be easy to request it at the end of the questionnaire rather than at the beginning.

The use of abstract or vague expressions in the questionnaires is very common. The chances of respondents varying in their interpretation of concepts such as "Emotional health" and "Leadership skills" are high. However, when the authors attempt to elaborate on such items this often results in double-barrelled items (e.g., "Developing an understanding and enjoyment of art, music and drama"). *Avoid abstract expressions* may seem like a truism but there seems to be a lot of room for improvement here.

The popularity of questionnaires on self-reported competences is understandable because they appear to offer a simple way to tap into the outcomes of higher education. We hope that our listing of some sound design features for questionnaires serves to highlight the fact that the design of a questionnaire containing self-rated competences needs to be undertaken as rigorously as the development of an objective test. A lot of time and research is needed at the development stage if subsequent surveys are to yield benefits. However well designed the questionnaire is, it seems unlikely to us that one can assess something as complex as competences using single items. Rather, several theoretically-based questions should be used to cover the different facets of a given construct, as is the case in the HEsaCom and the NSSE.

There has been a fair amount of research into the different types of validity of self-assessment in general. Several findings of theoretically expected correlations and variations among subgroups have been reported in the literature. Lucas and Baird (2006) conclude that: "Although errors surely do occur, they often do not severely limit the validity of measures... .Self-reports often demonstrate impressive accuracy, predictability, and utility in important research settings" (p. 41). However, alongside this research is another body of findings that demonstrates the possibly distorting effects of certain types of questionnaire design. These effects need to be taken into account when developing an instrument but also when interpreting the results. As Schwarz (1999) says: "The problem is not the context dependency of human judgment but researchers' hope that this context dependency may—miraculously—not apply to their own study" (p. 103).

We are convinced that existing questionnaires and those that will be developed can be much improved simply by taking the features of good questionnaire design that we have outlined into consideration. Nevertheless, the questionnaires that we have examined have revealed some interesting results (for example, differences among subgroups of students) and have produced empirical findings that are consistent with hypotheses. What we would like to see more of in the published literature is evidence of *predictive validity*. For example, do those who rate themselves highly on "leadership ability" actually go on eventually to be good leaders?

Appendix A. Questionnaires

College Student Experiences Questionnaire (CSEQ): http://cseq.iub.edu/.

Cooperative Institutional Research Program (CIRP): http://www.heri.ucla.edu/abtcirp.php.

Course Experience Questionnaire (CEQ): http://start.graduatecareers.com.au/AGSoverview/CEQ/index.htm.

Evaluation in Higher Education, Self-Assessed Competences (HEsaCom): http://www.ewi-psy.fu-berlin.de/einrichtungen/arbeitsbereiche/ewi-psy/forschung/fb_lehrevaluation2/index.html.

National Survey of Student Engagement (NSSE): http://nsse.iub.edu/.

Personal and Educational Development Inventory (PEDI): http://dx.doi.org/doi:10.1080/03075070410001682628.

Student Instructional Report II (SIR II): http://www.ets.org/portal/site/ets/menuitem.1488512ecfd5b8849a77b13bc3921509/?vgnextoid=ff79af5e44df4010VgnVCM10000022f95190RCRD&vgnextchannel=39f1be3a864f4010VgnVCM10000022f95190RCRD.

References

Adelman, C. (2008). *The Bologna club: What US higher education can learn from a decade of European reconstruction?* Washington, DC: Institute for Higher Education Policy [Retrieved from <http://www.ihep.org/assets/files/TheBolognaClub.pdf>].

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.

Angoff, W. H. (1988). Validity: An evolving concept. In H. Wainer & H. Braun (Eds.), *Test validity* (pp. 9–13). Hillsdale, NJ: Lawrence Erlbaum.

Bologna Working Group on Qualifications Frameworks. (2005). *A framework for qualifications of the European Higher Education Area*. Copenhagen: Ministry of Science, Technology and Innovation [Retrieved from <http://www.bologna-bergen2005.no/Docs/00-Main_doc/050218_QF_EHEA.pdf>].

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2004). The concept of validity. *Psychological Review, 111*, 1061–1071. doi:10.1037/0033-295X.111.4.1061.

Braun, E. (2007). *Das Berliner Evaluationsinstrument für selbsteingeschätzte studentische Kompetenzen (BEvaKomp)* [The Berlin evaluation tool for students' self-assessed competences (BEvaKomp)]. Göttingen, Germany: Vandenhoeck & Ruprecht unipress.

Braun, E., & Hannover, B. (2008). Zum Zusammenhang zwischen Lehr-Orientierung und Lehrgestaltung von Hochschuldozierenden und subjektivem Kompetenzzuwachs bei Studierenden [On the relationship between teacher orientation and teaching skills in university lecturers and subjective increase in competences in students] [Special Issue No. 9]. In M. A. Meyer, M. Prenzel, & S. Hellekamps (Eds.), *Perspektiven der Didaktik [Perspectives on Teaching] (Zeitschrift für Erziehungswissenschaft* (pp. 277–291). Wiesbaden: VS Verlag für Sozialwissenschaften.

Braun, E., & Hannover, B. (2011). Gelegenheiten zum Kompetenzerwerb in der universitären Lehre: Zusammenhänge zwischen den Einschätzungen Studierender und unabhängigen Beobachtungen relevanter Merkmale universitärer Lehrveranstaltungen [Opportunities for skills acquisition in university teaching: Relationships between students' evaluations and independent observations of relevant characteristics of university courses.]. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie, 43*, 22–28. doi:10.1026/0049-8637/a000029.

Braun, E., & Leidner, B. (2009). Academic course evaluation: Theoretical and empirical distinctions between self-rated gain in competences and satisfaction with teaching behavior. *European Psychologist, 14*, 297–306. doi:10.1027/1016-9040.14.4.297.

Cattell, R. B. (1946). *Description and measurement of personality*. New York: World Book Company.

Centra, J. (1979). *Determining faculty effectiveness: Assessing teaching, research and service for personnel decisions and improvement*. San Francisco: Jossey-Bass.

Centra, J. (1998). *The development of the Student Instructional Report II*. Princeton, NJ: Educational Testing Service [Retrieved from <http://www.ets.org/Media/Products/283840.pdf>].

Gaubatz, J., & Centra, N. B. (2005). *Student perceptions of learning and instructional effectiveness in college courses: A validity study of SIR II*. Princeton, NJ: Educational Testing Service [Retrieved from <http://www.ets.org/Media/Products/perceptions.pdf>].

Coates, H., & Edwards, D. (2009). *Engaging college communities: The impact of residential colleges in Australian higher education (AUSSE Research Briefing No. 4)*. Camberwell, Victoria: Australian Council for Educational Research [Retrieved from <http://www.acer.edu.au/documents/AUSSE_BriefingVolume409.pdf>].

Cole, J. S., & Gonyea, R. M. (2010). Accuracy of self-reported SAT and ACT test scores: Implications for research. *Research in Higher Education, 51*, 305–319. doi:10.1007/s11162-009-9160-9.

Conway, M., & Ross, M. (1984). Getting what you want by revising what you had. *Journal of Personality and Social Psychology, 47*, 738–748. doi:10.1037/0022-3514.47.4.738.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Fort Worth, TX: Harcourt Brace.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*, 297–334. doi:10.1007/BF02310555.

Cronbach, L. J. (1988). Five perspectives on validity argument. In H. Wainer & H. Braun (Eds.), *Test validity* (pp. 3–17). Hillsdale, NJ: Lawrence Erlbaum.

Cronbach, L. J., & Gleser, G. C. (1965). *Psychological tests and personnel decisions*. Urbana, IL: University of Illinois Press.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*, 281–302. doi:10.1037/h0040957.

Ebel, R. L. (1961). Must all tests be valid? *American Psychologist, 16*, 640–647. doi:10.1037/h0045478.

Edmunds, R., & Richardson, J. T. E. (2009). Conceptions of learning, approaches to studying and personal development in UK higher education. *British Journal of Educational Psychology, 79*, 293–309. doi:10.1348/000709908X368866.

European Association for Quality Assurance in Higher Education. (2005). *Standards and Guidelines for Quality Assurance in the European Higher Education Area*. Helsinki: Author [Retrieved from <http://www.eqar.eu/fileadmin/documents/e4/050221_ENQA_report.pdf>].

Fan, X., & Thompson, B. (2001). Confidence intervals about score reliability coefficients, please: An EPM guidelines editorial. *Educational and Psychological Measurement, 61*, 517–531. doi:10.1177/00131644014614001.

Gonyea, R. M., Kish, K. A., Kuh, G. D., Muthiah, R. N., & Thomas, A. D. (2003). *College Student Experiences Questionnaire. Norms for the fourth edition*. Bloomington, IN: Indiana University Center for Postsecondary Research, Policy and Planning [Retrieved from <http://cseq.iub.edu/pdf/intro_CSEQ_4th_Ed_Norms.pdf>].

Gosling, S. D., John, O. P., Craik, K. H., & Robins, R. W. (1998). Do people know how they behave? Self-reported act frequencies compared with on-line codings by observers. *Journal of Personality and Social Psychology, 74*, 1337–1349. doi:10.1037/0022-3514.74.5.1337.

Greenwald, A. G., & Gillmore, G. M. (1998). How useful are student ratings? *American Psychologist, 53*, 1228–1229. doi:10.1037/0003-066X.53.11.1228.
Guion, R. (1977). Content validity: The source of my discontent. *Applied Psychological Measurement, 1*, 1–10. doi:10.1177/014662167700100103.
Hannover, B. (2000). Das kontextabhängige Selbst oder warum sich unser Selbst mit dem sozialen Kontext verändert [The context-dependent self, or why our self changes with the social context]. In W. Greve (Ed.), *Psychologie des Selbst [Psychology of the self]* (pp. 227–238). Weinheim: Psychologie-Verlags-Union.
Hu, S., & Kuh, G. D. (2002). Being (dis)engaged in educationally purposeful activities: The influences of student and institutional characteristics. *Research in Higher Education, 43*, 555–575. doi:10.1023/A:1020114231387.
Joinson, A. (2001). Self-disclosure in computer-mediated communication: The role of self-awareness and visual anonymity. *European Journal of Social Psychology, 31*, 177–192. doi:10.1002/ejsp. 36.
Kalton, G., Collins, M., & Brook, L. (1978). Experiments in wording opinion questions. *Journal of the Royal Statistical Society (Series C), 27*, 149–161 [Retrieved from <http://www.jstor.org/stable/2346942>].
Kane, M. T. (2001). Current concerns in validity theory. *Journal of Educational Measurement, 38*, 319–342. doi:10.1111/j.1745-3984.2001.tb01130.x.
Kelley, T. L. (1927). *Interpretation of educational measurements*. New York: Macmillan.
Krosnick, J. (1999). Survey research. *Annual Review of Psychology, 50*, 537–567. doi:10.1146/annurev.psych.50.1.537.
Kuh, G. (2003). *The National Survey of Student Engagement: Conceptual framework and overview of psychometric properties*. Bloomington, IN: Indiana University Center for Postsecondary Research [Retrieved from <http://nsse.iub.edu/pdf/conceptual_framework_2003.pdf>].
Kuh, G. D. (2009). The National Survey of Student Engagement: Conceptual and empirical foundations. *New Directions for Institutional Research, 141*, 5–20. doi:10.1002/ir.283.
Kuh, G. D., Cruce, T. M., Shoup, R., Kinzie, J., & Gonyea, R. M. (2008). Unmasking the effects of student engagement on first-year college grades and persistence. *Journal of Higher Education, 79*, 540–563 [Retrieved from <http://www.jstor.org/stable/25144692>].
Lawless, C., & Richardson, J. T. E. (2004). Monitoring the experiences of graduates in distance education. *Studies in Higher Education, 29*, 353–374. doi:10.1080/03075070410001682628.
Liu, A., Sharkness, J., & Pryor, J. H. (2008). *Findings from the 2007 administration of Your First College Year (YFCY): National aggregates*. Los Angeles, CA: University of California, Higher Education Research Institute [Retrieved from <http://www.heri.ucla.edu/PDFS/YFCY_2007_Report05-07-08.pdf>].
Lord, F., & Novick, M. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.
Lucas, R., & Baird, B. (2006). Global self assessment. In M. Eid & E. Diener (Eds.), *Handbook of psychological measurement: A multimethod perspective* (pp. 29–42). Washington, DC: American Psychological Association.
Mabe, P. A., & West, S. G. (1982). Validity of self-evaluation of ability: A review and meta-analysis. *Journal of Applied Psychology, 67*, 280–296. doi:10.1037/0021-9010.67.3.280.
Marsh, H. W., & Roche, L. A. (1997). Making students' evaluations of teaching effectiveness effective: The critical issues of validity, bias, and utility. *American Psychologist, 52*, 1187–1197. doi:10.1037/0003-066X.52.11.1187.
Martin, E. (1996). *The effectiveness of different models of work-based university education*. Adelaide: Department of Employment, Education, Training and Youth Affairs [Retrieved from <http://www.dest.gov.au/archive/highered/eippubs/eip9619/front.htm>].
Masten, A., & Coatsworth, J. D. (1998). The development of competence in favorable and unfavorable environments: Lessons from research on successful children. *American Psychologist, 53*, 205–220. doi:10.1037/0003-066X.53.2.205.
Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York: American Council on Education and Macmillan.
Meyer, G. J., Finn, S. E., Eyde, L. D., Kay, G. G., Moreland, K. L., Dies, R. R., Eisman, E. J., Kubiszyn, T. W., & Read, G. M. (2001). Psychological testing and psychological assessment: A review of evidence and issues. *American Psychologist, 56*, 128–165. doi:10.1037/0003-066X.56.2.128.
Noftle, E., & Fleeson, W. (2010). Age differences in Big Five behavior averages and variabilities across the adult life span: Moving beyond retrospective, global summary accounts of personality. *Psychology and Aging, 25*, 95–107. doi:10.1037/a0018199.
Norenzayan, A., & Schwarz, N. (1999). Telling what they want to know: Participants tailor causal attributions to researchers' interests. *European Journal of Social Psychology, 29*, 1011–1020. doi:10.1002/(SICI)1099-0992(199912)29:8<1011::AID-EJSP974>3.0.CO;2-A.
Organisation for Economic Cooperation and Development. (2005). *The definition and selection of key competencies: Executive summary*. Paris: OECD, Directorate for Education [Retrieved from <http://www.oecd.org/dataoecd/47/61/35070367.pdf>].
Organisation for Economic Cooperation and Development. (2009). *Education at a glance 2009: OECD indicators*. Paris: OECD, Directorate for Education [Retrieved from <http://www.oecd.org/dataoecd/41/25/43636332.pdf>].
Pace, R., & Stern, G. (1958). An approach to the measurement of psychological characteristics of college environments. *Journal of Educational Psychology, 49*, 269–277. doi:10.1037/h0047828.
Pike, G. (2006). The convergent and discriminant validity of NSSE scalelet scores. *Journal of College Student Development, 47*, 551–564. doi:10.1353/csd.2006.0061.
Programme for International Student Assessment. (2005). *Problem solving for tomorrow's world: First measures of cross-curricular competencies from PISA 2003*. Paris: Organisation for Economic Cooperation and Development [Retrieved from <http://www.oecd.org/dataoecd/25/12/34009000.pdf>].
Prosser, M., & Trigwell, K. (2006). Confirmatory factor analysis of the approaches to teaching inventory. *British Journal of Educational Psychology, 76*, 405–419. doi:10.1348/000709905X43571.
Ramsden, P. (1991). A performance indicator of teaching quality in higher education: the Course Experience Questionnaire. *Studies in Higher Education, 16*, 129–150. doi:10.1080/03075079112331382944.
Richardson, J. T. E. (2000). *Researching student learning: approaches to studying in campus-based and distance education*. Buckingham, UK: SRHE & Open University Press.
Richardson, J. T. E. (2004). Methodological issues in questionnaire-based research on student learning in higher education. *Educational Psychology Review, 16*, 347–358. doi:10.1007/s10648-004-0004-z.
Roche, L. A., & Marsh, H. W. (1998). Workload, grades, and students' evaluations of teaching: Clear understanding sometimes requires more patient explanations. *American Psychologist, 53*, 1230–1231. doi:10.1037/0003-066X.53.11.1230.
Ross, M. (1989). The relation of implicit theories to the construction of personal histories. *Psychological Review, 96*, 341–357. doi:10.1037/0033-295X.96.2.341.
Rychen, D. S., & Salganik, L. (2003). *Key competencies for a successful life and a well-functioning society*. Cambridge, MA: Hogrefe & Huber Publishers.
Schuman, H., & Presser, S. (1981). *Questions and answers in attitude surveys*. New York: Academic Press.
Schwarz, N. (1999). Self-reports: How the questions shape the answers. *American Psychologist, 54*, 93–105. doi:10.1037/0003-066X.54.2.93.
Spain, J. S., Eaton, L. G., & Funder, D. C. (2000). Perspective on personality: The relative accuracy of self versus others for the prediction of emotion and behavior. *Journal of Personality, 68*, 837–867. doi:10.1111/1467-6494.00118.
Spencer, S. J., Steele, C. M., & Quinn, D. M. (1999). Stereotype threat and women math performance. *Journal of Experimental Social Psychology, 35*, 4–28. doi:10.1006/jesp. 1998.1373.
Steele, C. M., & Aronson, J. (1995). Stereotype threat and the intellectual test performance of African Americans. *Journal of Personality and Social Psychology, 69*, 797–811. doi:10.1037/0022-3514.69.5.797.
Strack, F., & Schwarz, N. (1992). Communicative influences in standardized question situations: The case of implicit collaboration. In K. Fiedler & G. Semin (Eds.), *Language and social cognition* (pp. 173–193). Cambridge, UK: Cambridge University Press.
Weinert, F. E. (2001). Concept of competence. A conceptual clarification. In D. Rychen & L. Salganik (Eds.), *Defining and selecting key competences* (pp. 17–31). Göttingen, Germany: Hogrefe.
Wilson, K., Lizzio, A., & Ramsden, P. (1997). The development, validation and application of the Course Experience Questionnaire. *Studies in Higher Education, 22*, 33–53. doi:10.1080/03075079712331381121.