Chapter 7

Online Assessment

Siegbert Reiss[1] and Ulf-Dietrich Reips[2]

[1]Department of Psychology, Goethe University Frankfurt, Germany
[2]Department of Psychology, University of Konstanz, Germany

Introduction

Until the end of the last century, a large part of the work in the area of technology-based assessment concentrated on the administration of conventional tests by computer. Afterwards, the universal presence of the Internet has spurred current development in assessment. The Internet has opened new ways of testing by facilitating the distribution of test materials, extending the options of item presentation, enhancing the time and locations for independent collection of data, and offering increased convenience for the examinees. Because in online-testing, neither the test taker nor the testing professional needs to be on-site for test administration and scoring, analyses of results and test evaluation can usually be automated. This approach is cost-effective and leads to considerable economic advantages after initial investment. In general, it seems to be easier to compile representative norming samples for different populations, see, for example, the multiple site entry technique (e.g., Birnbaum & Reips, 2005). Test-takers can be given their results immediately after completing a test and norms can be constantly updated. Second, a large database of items offers the possibility to produce IRT-scaled items and to conduct adaptive testing. Furthermore, advantages from computer-based testing as compared to paper-and-pencil testing are inherent in Internet-based testing. A paper-and-pencil approach may limit the way some abilities, skills and competences can be measured. In contrast, computer-based testing extends the available media and allows the presentation of various scenarios and the assessment of behavior under such scenarios. It encourages the development of innovative types of tasks and items, including automatically scored essay questions, simulations of laboratory science experiments, and other forms of constructed response items that require students to produce, rather than just select, their answers (for more information on item formats, see Chapter 5, this volume). Many of these efforts have sought to incorporate the most recent information technology innovations to expand the range of activities in which students can engage. Another aspect to consider in this context is the positive reaction of test takers towards Internet assessments. Salgado and Moscoso (2003) as well as Potosky and Bobko (2004) found that Internet-based tests were more positively perceived by test-takers than the paper-and-pencil version of the same tests.

Internet-based testing, as with the other three types of Internet-based research methods (non-reactive web-based methods, web surveys, and web experiments), requires attention to various peculiarities that sometimes substantially alter a test administrator's testing routines (Reips, 2006). The various advantages of online assessment are countered by a number of disadvantages that place specific requirements on their use. First, the usability of the websites on which

the test is executed must be guaranteed to ensure an unbiased and consistent performance of the test taker. Usability exceeds the requirement that the website must work without error. It is just as important to consider how easily and quickly the users can perform the tasks, how many errors users make when working on the website, and how satisfied users are with the appearance of and navigation on the website. Furthermore, data security measures are required to prevent the dissemination (and potential abuse) of confidential data.

Usually, psychometric tests are designed to be administered under controlled, standardized conditions. In contrast to traditional psychological testing that requires the presence of a test administrator, Internet-based tests are administrated to individuals outside a traditional proctored setting. An unproctored Internet test could be completed by applicants literally anywhere Internet is available. Test takers may complete assessment instruments in different locations, and under different physical and psychological conditions (e.g., Carstairs & Myors, 2009). In fact, it cannot be ruled out that any of these conditions interact with the measurement of the construct of interest. At least for personality measures, Arthur and colleagues (Arthur, Glaze, Villado, & Taylor, 2010, p. 15) demonstrated that unproctored Internet testing and proctored tests seem to reveal similar levels of cheating or response distortion as traditional testing situations. All in all, Richman, Kiesler, Weisband, and Drasgow (1999) could not find larger bias effects in their meta-analytic study of social desirability distortion in computer administered questionnaires than in traditional questionnaires (for Internet-based questionnaires, see Kaufmann & Reips, 2008). For a detailed discussion of open questions regarding unproctored Internet testing, see Tippins (2009).

However, if assessment results are to be used for an important purpose, one needs to establish that the results have not been biased by lack of standardization. While this necessity applies in general for paper-based tests in a proctored test environment as well, it should, however, be emphasized that Internet-based tests, of course, must comply with the same high psychometric quality requirements, (i.e., a sufficient level of validity), as other diagnostic instruments (Buchanan, 2001; Buchanan & Smith, 1999).

Numerous comparability studies have been conducted in the last years on a variety of psychological tests to examine whether test presentation mode – computer-based or paper-based – affects examinee performance (e.g., Davis, 1999; Texas Education Agency, 2008; Wang & Shin, 2010; Weigold, Weigold, & Russell, 2013). Whilst some studies have found benefits relating to computer administration and others have favored paper-based, the majority of recent comparability studies have indicated that computer-based and paper-based tests are comparable across delivery medium, at least in the multiple-choice format (Kingston, 2009; Mead & Drasgow, 1993; Noyes & Garland, 2008). The two methods produce similar statistical distributions (means, standard deviations, reliabilities, and standard errors of measurement) of test scores and are comparable in their predictive validity estimates. To minimize test irrelevant variance originating from an online administration, a number of documents relevant to standards and good practice for online testing have been published in the last years (e.g., Allan, Bulla & Goodman, 2003; Bailey, Schneider, & Ark, 2012; Bartram, 2006; International Test Commission, 2006; Lievens, 2006; Naglieri et al., 2004; Scheuermann & Guimarães Pereira, 2008). The objective of these guidelines is to complement the existing Standards for educational and psychological testing with a specific focus on computer-based and Internet-delivered testing. As a product of an intensive process of international collaboration and review, the International Test Commission (ITC) guidelines provides a valuable reference for best practices when conducting Internet-based testing. These guidelines are directed towards all stakeholders involved in the process of online testing such as the users, the developers and the publishers of comput-

er-based tests (CBTs) and Internet tests. The ITC guidelines (International Test Commission, 2006) address performance characteristics and technical limitations of the Internet as a test delivery medium (i.e., network speed, integrity, bandwidth, item presentation etc.), quality issues (i.e., psychometric qualities), control issues (i.e., control over the test conditions) and security issues (e.g., keeping scoring routines confidential, controlling access to test results, legal issues relating to data protection, privacy and storage).

Item Presentation and Response Acquisition

As mentioned, a computer-based presentation extends the media available for presentation and affects response modes, stimulus materials, and the input devices of the test (Dolan, Burling, Harms, Strain-Seymour, & Way, 2013; Educational Testing Service [ETS], 2012; Rosen & Tager, 2013; Sireci & Zenisky, 2006). Aside from single and multiple-choice format, fill-in-the-blank, highlighting, inserting text, and drag and drop techniques, various graphical response modes are available. Tasks and items may include animations, interactive and static graphics, embedded audio files and video clips. Input devices include the recent information technology hardware like keyboard, mouse, touch screen, light pen, joystick, graphics tablet, voice recorder, personal digital assistant or smartphone, etc., often equipped with speech and handwriting recognition software. An example of a task with video embedded is shown in Figure 7.1. The test administrator can choose whether the video should immediately begin to be played on the webpage or provide buttons to start, pause, and stop the video.

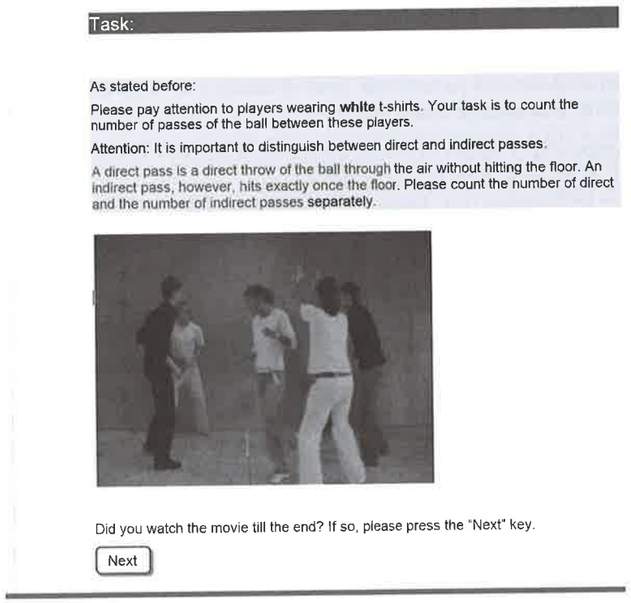

Figure 7.1. Web-based assessment task with video embedded.

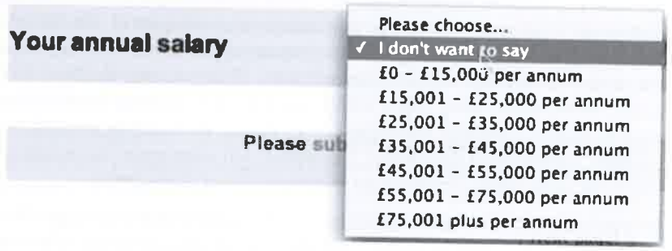

Figure 7.2. Drop-down menu item with preselected "I don't want to say" option in second position.

In a second example, Figure 7.2 shows some of the challenges in drop-down or pop-up menus (Reips, 2002). Options only appear after the menu is clicked with the mouse arrow, then one option needs to be selected my moving the mouse arrow and releasing the mouse button. Thus, item options may differ in physical and cognitive accessibility.

Despite the availability of many new options for response presentation and functionality Buchanan and Reips (2001) observed that respondent personality is associated with technology preferences and thus, recommended a low-tech philosophy of designing Internet-based questionnaires and tests to avoid systematic errors. This principle is further supported by findings showing difficulties for individuals with lower education levels when technologically challenging response formats (e.g., slider scales) are presented (Funke, Reips, & Thomas, 2011).

Reips (2010) categorizes design and formatting issues with questionnaires and tests on the Internet into two types: (a) blatant errors and (b) design decisions made consciously, but without the designer realizing that the format makes the test less useable or, in worst case scenarios, worthless for research purposes. He cites a typical example that shows the following errors (among others) that frequently appear in Internet-based questionnaires:

- Including preselected answers in drop-down menus or radio button lists;
- Volume of text to be entered in text fields is not limited;
- Lack of options that indicate reluctance to answer (e.g., "I don't want to answer");
- All items on one run-on webpage rather than following a one-item-one-screen (OIOS) design; and
- Background information on the survey accidentally being revealed via the URL of the survey.

Online Self-Assessment

One area where Internet testing has blossomed is the area of online self-assessment. These tests are usually embedded in the context of electronic application processes or of technology-enhanced learning environments and require diagnostic self-testing by the participants.

If we examine the research about how accurately people judge themselves, the usual finding is that people have low insight to objectively assess their own skills and character. A meta-analysis by Freund and Kasten (2012) found that self-estimates of cognitive ability were positive, but moderately correlated with psychometrically measured cognitive ability at a level of

roughly .33. One explanation for this weak correlation is that people do not have the knowledge and information necessary to assess their competence adequately. On the other hand, several studies suggest that integrating self-assessments improved the learning outcome and the accurate appraisal of their own performance (e.g., Chang, Liang, & Chen, 2013; Li & De Luca, 2012; Ross, 2006; Wang, 2007, 2009). So, it is not surprising that self-assessments are increasingly used in formative assessment (e.g., during the learning process) in order to improve students' attainment of the content. This area of self-assessment has benefited particularly well from the tremendous development in Internet-based assessments.

Kotter (1978, p. 7) characterizes self-assessment as "a systematic process designed to generate the type of accurate self-awareness needed to make rational job and career decisions." More recently, Frebort and Kubinger (2008) defined self-assessments as psychological relevant test batteries for the purposes of diagnostic self-testing, mostly presented via the Internet. In that sense, self-assessment is seen as one module in the process of a deliberate choice, where the test taker voluntarily engages in a diagnostic self-testing procedure and uses the feedback first and foremost to his or her own benefit. Usually the recordings are not accessible to third parties unless noted (e.g., results shared to obtain detailed psycho-diagnostic counseling). The results of a self-assessment are usually solely available to the test taker, who has a genuine desire to get veritable information about the tested content; therefore, distortion as discussed above is, under these conditions, rather unlikely. Moreover, recent research has indicated that people might be more willing to self-disclose in an online (self-)assessment than they would in person (Barak, & Gluck-Ofri, 2007; Suler, 2004). Thus, results of an online self-assessment may reflect personality aspects unfettered by normal social constraints and might lead to greater ecological validity. However, one of the implications that must be drawn from this is that normative data used in online self-assessments should be gathered from the appropriate online population.

Online self-assessment tools find their primary application in the area of career counseling, higher education institutions, and health care. All of these tools share the common goal to provide the test takers with feedback about their accordance of abilities, skills knowledge, expectations, interests and health.

Online Assessment in Career Counseling

Online assessments, and, in particular, self-assessments, are widespread in personnel consulting and personnel selection. Frequently these assessments are offered as part of electronic recruiting by internationally accessible job centers and globally active companies. Usually, it is these instruments' self-declared goal to inform job-seekers of their skills and personal suitability for a job, thus saving time and expense for the job-seekers and the company (Bartram, 2005; Psychological Testing Centre, 2006). These tests normally capture biographical information, abilities, and competencies and contain tasks that are adapted to the specific requirements and demands of the respective business sector. Additionally, intelligence tests and tests for specific cognitive abilities are mainly used in the course of recruiting trainees and job starters. In the area of personnel selection, tests primarily cover job-specific knowledge, and include personality scales, as well as scales measuring motivation, attitudes and interests to obtain information about behavior as it relates to an employees' daily professional life. Moreover, simulations of concrete daily tasks in the company are commonly applied in order to ascertain personal suitability to a job. Further, job applications may involve additional forms of electronically supported learning provided in in-company training and qualification tests.

In both cases, online self-assessments are generally used for screening and (self-)selection of suitable candidates. Nevertheless, the terms and conditions of an aptitude assessment still apply, meaning that the instruments employed should demonstrate general quality requirements for tests and demonstrate empirical evidence of the prognostic validity of the vocationally specific aptitudes measured.

Because online assessments in the employment context strive to report the test-taker's personal suitability for a job, a thorough analysis of the relevant personal conditions and requirements of the position to be filled is essential. It is good practice to have a stepwise procedure, which first covers essential information, such as possession of appropriate degrees/certifications, validated work experience and skill prescreens before finally testing abilities, competencies and personality variables. This approach reduces step-by-step the number of applicants who proceed into more time consuming parts of the assessment process (Russell, 2007).

Online Assessment in Higher Education

In the education sector, online assessments and self-assessments play a dual role. On the one hand, the assessments are used as a part of technology-enhanced learning environment; on the other hand, they are commonly embedded in the context of college and university admission.

Technology-enhanced learning environments are often used in higher education and in school education as a means of creating learning situations that require complex thinking, problem-solving and collaboration strategies. For the most part, these learning environments include self-assessment tools that are increasingly integrated into more complex and authentic problem contexts and a wider range of answer formats can be automatically scored. For example, electronic portfolios are already widely used in European schools and universities to support the formative and summative assessment of students' progress. Self- and peer-assessment can be powerful tools to assess students' learning. Additionally, immersive virtual environments, online simulations, virtual laboratories and games can recreate learning situations which require complex thinking and problem-solving, thus, allowing the development and assessment of skills and competences. For content areas such as science and technology, as well as for social and civic competencies, computer simulations and virtual laboratories provide opportunities for students to develop and apply skills and knowledge in more realistic contexts and can provide feedback in real time. Since learners' behavior in these electronic environments may be tracked, their individual learning results can be automatically assessed. Such computer-based problem-solving scenarios or *microworlds* have been successfully used to assess students' complex problem-solving behavior (Mayer et al., 2013; Sonnleitner, et al., 2012; Sonnleitner, Keller, Martin, & Brunner, 2013) and their strategies to move through the materials (Lengler & Reips, 2003). Additionally, the Internet reveals a pool of potential computer-based quizzes, games and tests which can be used for the assessment of competences in literacy, reading and text comprehension and mathematics, in the educational sector. As Redecker (2013) notes, the trends in technology-enhanced assessment is to make explicit testing obsolete:

> Learners will be continuously monitored and guided by the electronic environment which they use for their learning activities, so that diagnostic, formative and summative assessment will become embedded in the learning process. (p. 12)

Apart from technology-enhanced learning environments, online self-assessments are commonly used in the schools for issues related to student guidance and counseling. US high schools typically use career inventories in high school settings to provide feedback as to their

suitability for jobs/careers as students prepare to enter higher education or the work force. In contrast to the US, in several European countries, for example in Germany and Austria, applying to colleges or universities is not linked to the participation in standardized admissions tests. In order to ensure a substantial matching between the individual competencies of the students and the specific requirements of the intended subject, many universities offer online self-assessments to inform the prospective students as to their personal suitability to that subject. On the one hand, universities expect that the profound concern with the requirements of the subject will lead to a kind of self-selection, which, in turn, may reduce the number of students leaving a major or changing the area of study. On the other hand, aptitude-related feedback opens up the opportunity to deal with strengths and weaknesses on an individual basis. This assumes that the test applied should – apart from content validity – demonstrate the empirical evidence of its prognostic validity.

While benefits to such inventories may be realized, collecting data is difficult in the context of online self-assessment and academic success as sampling bias is quite likely because of the access conditions and voluntary participation of examinees. Nevertheless, there is empirical evidence of the prognostic validity of online self-assessments in the admission to universities (e.g., Hasenberg & Schmidt-Atzert, 2013; Reiss et al., 2009).

Online Assessment in Health Care

In the last decade, the new term *eHealth* was established to describe the use of emerging interactive technologies delivered via computer and/or the Internet, to enable disease prevention and disease management. Besides the economic advantages, this approach offers an increased user and supplier control of interventions. The reduction of geographically-based, as well as time- and mobility-based barriers (Griffiths, Lindenmeyer, Powell, Lowe, & Thorogood, 2006) are the most common reasons for delivering Internet-based monitoring and interventions (Ybarra & Eaton, 2005). In accordance, a variety of computer- and web-based patient monitoring and assessment systems have been developed. As in other domains, several studies have demonstrated that computer-based assessments can rival the validity of assessments using paper-and-pencil methods or trained interviewers (e.g., Achenbach, Krukowski, Dumenci, & Ivanova, 2005; Schulenberg & Yutrzenka, 2001; Vallejo, Jordán, Díaz, Comeche, & Ortega, 2007).

A specific interactive approach in the field of eHealth is termed *expert systems*. These expert systems typically involve a collection of characteristics and generate a feedback protocol tailored to the specific needs of the user on the basis of the users input. From this perspective, expert systems attempt to model methods similar to a real-world clinical encounter. A typical example is the web-based system Systematic Treatment Selection (STS; Nguyen, Bertoni, Charvat, Gheytanchi, & Beutler, 2007). This system is a cloud-based mental health clinical platform that provides a tailored and direct assessment of the patient problem and provides a written intake report, multiple measures of intake and outcome condition, and tailored self-help resources (http://www.innerlife.com/clinician/). The system is based on empirically derived principles that define the conditions under which different types of interventions are most likely to exert positive effects. Patient characteristics that fit to these principles are identified via a web-delivered self-report measure. Research has demonstrated the therapeutic efficacy of various treatment-planning dimensions in the STS/InnerLife system (Harwood et al., 2011).

Virtual reality is another technology that appeared online in recent years to be used as a clinical component in assessment and treatment of disorders and cognitive rehabilitation. The rationale

behind this approach is that real and virtual exposures elicit comparable reactions in participants. Virtual reality has been used, for instance, to develop suitable and virtual environments for areas of cognitive functions assessment and rehabilitation (Morganti, 2006), the assessment of body image (Villani, Gatti, Confalonieri, & Riva, 2012) and emotional responses (Gorini, Griez, Petrova, & Riva, 2010).

In addition, there is a growing number of online questionnaires related to clinical psychology, which stresses the importance of self-assessment in the mental health arena (e.g., Ritter, Lorig, Laurent, & Matthews, 2004). Because the assessment of adult psychopathology relies heavily on self-reports, it is important to determine how well self-reports agree with reports by informants who know the person being assessed. Achenbach and colleagues (2005) examined 51,000 articles published over 10 years in 52 peer-reviewed journals for correlations between self-reports and informants' reports about psychopathological characteristics, such as aspects of behavioral, emotional, cognitive, and personality functioning that may have been sufficiently abnormal, injurious, or troubling and warranting help from mental health professionals. They report a mean correlation of .453 across all recorded disorders. Correlations were larger for reports of substance use (.681) than for other kinds of problems (e.g., Internalizing, that included anxiety, depression, neuroticism, and suicidality, .428; Externalizing, that is, aggression, sociopathy, and antisocial behavior, .438). Achenbach et al.'s results indicate that information about adult behavior problems from self-assessment measures may differ from those reported by other informants. Accordingly, it appears reasonable to amend self-reports by including additional information for the clinical assessment of individuals.

Specifics of Constructing Online Assessments

To construct assessments online, a number of methodological specifics need to be considered. Some of these specifics are general to Internet-based research (e.g., Birnbaum & Reips, 2005; Reips, Buchanan, Krantz, & McGraw, 2016), some are specific to online assessments (Buchanan, Johnson, & Goldberg, 2005). General advice regarding the construction of online studies includes the implementation of techniques to avoid or control non-response (e.g., warm-up, seriousness check, high hurdle technique; see Reips, 2009), the use of paradata (e.g., Stieger & Reips, 2010), and measures for authentication or at least control for multiple submissions (Birnbaum & Reips, 2005).

Online assessments require specific validation as instruments that are delivered online. Buchanan (2001) writes:

> Online and offline versions of tests can be, and often are, equivalent in terms of what they measure and how well they do it. However, equivalence is not something which one may take for granted. If equivalence is considered important, it must be established for every instrument which is used. One cannot assume that an online test is reliable and valid simply because the offline test from which it was created is known to be reliable and valid. (pp. 67–68)

References

Achenbach, T. M, Krukowski, R. A., Dumenci, L., & Ivanova, M. Y. (2005). Assessment of adult psychopathology: Meta-analyses and implications of cross-informant correlations. *Psychological Bulletin, 131*, 361–382. http://doi.org/10.1037/0033-2909.131.3.361

Allan, J. M., Bulla, N., & Goodman, S. A. (2003). *Test access: Guidelines for computer administered testing*. Louisville, KY: American Printing House for the Blind. Retrieved from http://www.aph.org/tests/access/

Arthur, W., Jr., Glaze, R. M., Villado, A. J., & Taylor, J. E. (2010). The magnitude and extent of cheating and response distortion effects on unproctored Internet-based tests of cognitive ability and personality. *International Journal of Selection and Assessment, 18*, 1–16. http://doi.org/10.1111/j.1468-2389.2010.00476.x

Bailey, J., Schneider, C., & Ark, T. V. (2012). *Getting ready for online assessments. Digital learning now! Smart series*. Retrieved from http://www.digitallearningnow.com/wp-content/uploads/2013/01/Getting-Ready-for-Online-Asst.-Updated-Jan-2013.pdf

Barak, A., & Gluck-Ofri, O. (2007). Degree and reciprocity of self-disclosure in online forums. *Cyberpsychology and Behavior, 10*, 407–417. http://doi.org/10.1089/cpb.2006.9938

Bartram, D. (2005). Testing on the Internet: Issues, challenges and opportunities in the field of occupational assessment. In D. Bartram & R. K. Hambleton (Eds.), *Computer-based testing and the Internet: Issues and advances* (pp. 13–37). Hoboken, NJ: John Wiley & Sons.

Bartram, D. (2006). The internationalization of testing and new models of test delivery on the Internet. *International Journal of Testing, 6*, 121–131. http://doi.org/10.1207/s15327574ijt0602_2

Birnbaum, M. H., & Reips, U.-D. (2005). Behavioral research and data collection via the Internet. In R. W. Proctor & K.-P. L. Vu (Eds.), *The handbook of human factors in Web design* (pp. 471–492). Mahwah, NJ: Erlbaum.

Buchanan, T. (2001). Online personality assessment. In U.-D. Reips & M. Bosnjak (Eds.), *Dimensions of Internet science* (pp. 57–74). Lengerich, Germany: Pabst Science.

Buchanan, T., Johnson, J. A., & Goldberg, L. R. (2005). Implementing a five-factor personality inventory for use on the Internet. *European Journal of Psychological Assessment, 21*(2), 115–127. http://doi.org/10.1027/1015-5759.21.2.115

Buchanan, T., & Reips, U.-D. (2001). Platform-dependent biases in online research: Do Mac users really think different? In K. J. Jonas, P. Breuer, B. Schauenburg, & M. Boos (Eds.), *Perspectives on Internet research: Concepts and methods*. Retrieved from http://www.uni-konstanz.de/iscience/reips/pubs/papers/Buchanan_Reips2001.pdf

Buchanan, T., & Smith, J. L. (1999). Research on the Internet: Validation of a World-Wide Web mediated personality scale. *Behavior Research Methods Instruments & Computers, 31*(4), 565–571. http://doi.org/10.3758/BF03200736

Carstairs, J., & Myors, B. (2009). Internet testing: A natural experiment reveals test score inflation on a high-stakes, unproctored cognitive test. *Computers in Human Behavior, 25*, 738–742. http://doi.org/10.1016/j.chb.2009.01.011

Chang, C.-C., Liang, C., & Chen, Y.-H. (2013). Is learner self-assessment reliable and valid in a Web-based portfolio environment for high school students? *Computers & Education, 60*, 325–334. http://doi.org/10.1016/j.compedu.2012.05.012

Davis, R. N. (1999). Web-based administration of a personality questionnaire: Comparison with traditional methods. *Behavior Research Methods Instruments & Computers, 31*(4), 572–577. http://doi.org/10.3758/BF03200737

Dolan, R. P., Burling, K., Harms, M., Strain-Seymour, E., & Way, W. (2013). *A universal design for learning-based framework for designing accessible technology-enhanced assessments. Pearson's Research Reports*. Retrieved from http://www.pearsonassessments.com/research

Educational Testing Service. (2012). *PARCC item development (ITN 2012-31)*. Retrieved from http://www.parcconline.org

Frebort, M., & Kubinger, K. D. (2008). Qualitätsansprüche an ein Self-Assessment zur Studienwahlberatung: Der Wiener Ansatz [Quality requirements to a self-assessment for study guidance: The Vienna approach]. In H. Schuler (Ed.), *Studierendenauswahl und Studienentscheidung* (pp. 95–101). Göttingen, Germany: Hogrefe.

Freund, P. A., & Kasten, N. (2012). How smart do you think you are? A meta-analysis on the validity of self-estimates of cognitive ability. *Psychological Bulletin, 138*, 296–321. http://doi.org/10.1037/a0026556

Funke, F., Reips, U.-D., & Thomas, R. K. (2011). Sliders for the smart: Type of rating scale on the Web interacts with educational level. *Social Science Computer Review, 29*, 221–231. http://doi.org/10.1177/0894439310376896

Gorini, A., Griez, E., Petrova, A., & Riva, G. (2010). Assessment of the emotional responses produced by exposure to real food, virtual food and photographs of food in patients affected by eating disorders. *Annals of General Psychiatry, 9*(30), 1–10. http://doi.org/10.1186/1744-859X-9-30

Griffiths, F., Lindenmeyer, A., Powell, J., Lowe, P., & Thorogood, M. (2006). Why are health care interventions delivered over the Internet? A systematic review of the published literature. *Journal of Medical Internet Research, 8*(2), e10. http://doi.org/10.2196/jmir.8.2.e12

Harwood, T. M., Pratt, D., Beutler, L. E., Bongar, B. M., Samarea, L., & Forrester, B. T. (2011). Technology, telehealth, treatment enhancement, and celection. *Professional Psychology: Research and Practice, 42*, 448–454. http://doi.org/10.1037/a0026214

Hasenberg, S., & Schmidt-Atzert, L. (2013). Bessere Noten und zufriedenere Studierende? Das Marburger Self-Assessment für den Studiengang Biologie [Better grades and more satisfied students? The Marburg self-assessment for the biology program]. *Wirtschaftspsychologie, 1*, 25–33.

International Test Commission. (2006). International guidelines of computer-based and Internet-delivered testing. *International Journal of Testing, 6*, 143–171. http://doi.org/10.1207/s15327574ijt0602_4

Kaufmann, E., & Reips, U.-D. (2008). *Internet-basierte Messung sozialer Erwünschtheit* [Internet-based measurement of social desirability]. Saarbrücken, Germany: VDM Verlag Dr. Müller.

Kingston, N. M. (2009). Comparability of computer- and paper-administered multiple-choice tests for K-12 populations: A synthesis. *Applied Measurement in Education, 22*, 22–37. http://doi.org/10.1080/08957340802558326

Kotter, J. P. (1978). *Self-assessment and career development*. Englewood Cliffs, NJ: Prentice-Hall.

Lengler, R. & Reips, U.-D. (2003). "Readers" and "scanners" evaluate e-learning: Same overall grading despite large differences in user behaviour and experiences. In C. Jutz, F. Flückiger, & K. Wefler (Eds.), *5th International Conference on New Educational Environments* (pp. 151–154). Bern, Switzerland: Sauerländer.

Li, J., & De Luca, R. (2012). Review of assessment feedback. *Studies in Higher Education*, 1–16.

Lievens, F. (2006). The ITC guidelines on computer-based and Internet-delivered testing: Where do we go from here? *International Journal of Testing, 6*, 189–194. http://doi.org/10.1207/s15327574ijt0602_7

Mayer, H., Hazotte, C., Djaghloul, Y., Latour, T., Sonnleitner, P., Brunner, M., … Martin, R. (2013). Using complex problem solving simulations for general cognitive ability assessment: The genetics lab framework. *International Journal of Information Science and Intelligent System, 2*, 71–88.

Mead, A. D., & Drasgow, F. (1993). Equivalence of computerized and paper-and-pencil cognitive ability tests: A meta-analysis. *Psychological Bulletin, 114*, 449–458. http://doi.org/10.1037/0033-2909.114.3.449

Morganti, F. (2006). Virtual interaction in cognitive neuropsychology. In G. Riva, C. Botella, P. Légeron, & G. Optale (Eds.), *Cybertherapy. Internet and virtual reality as assessment and rehabilitation tools for clinical psychology and neuroscience* (pp. 85–101). Amsterdam, The Netherlands: IOS Press.

Naglieri, J. A., Drasgow, F., Schmit, M., Handler, L., Prifitera, A., Margolis, A., & Velasquez, R. (2004). Psychological testing on the Internet: New problems, old issues. *American Psychologist, 59*, 150–162. http://doi.org/10.1037/0003-066X.59.3.150

Nguyen, T. T., Bertoni, M., Charvat, M., Gheytanchi, A., & Beutler, L. E. (2007). Systematic Treatment Selection (STS): A review and future directions. *International Journal of Behavioral Consultation and Therapy, 3*, 13–29. http://doi.org/10.1037/h0100178

Noyes, J. M., & Garland, K. J. (2008). Computer- vs. paper-based tasks: Are they equivalent? *Ergonomics, 51*, 1352–1375. http://doi.org/10.1080/00140130802170387

Potosky, D., & Bobko, P. (2004). Selection testing via the Internet: Practical considerations and exploratory empirical findings. *Personnel Psychology, 57*, 1003–1035. http://doi.org/10.1111/j.1744-6570.2004.00013.x

Psychological Testing Centre. (2006). *Using online assessment tools for recruitment*. Retrieved from http://www.psychtesting.org.uk

Redecker, C. (2013). *The use of ICT for the assessment of key competences. Publications Office of the European Union, JRC76971*. Retrieved from http://publications.jrc.ec.europa.eu/repository/handle/111111111/27945

Reips, U.-D. (2002). Context effects in Web surveys. In B. Batinic, U.-D. Reips, & M. Bosnjak (Eds.), *Online social sciences* (pp. 69–80). Seattle, WA: Hogrefe & Huber.

Reips, U.-D. (2006). Web-based methods. In M. Eid & E. Diener (Eds.), *Handbook of multimethod measurement in psychology* (pp. 73–85). Washington, DC: American Psychological Association. http://doi.org/10.1037/11383-006

Reips, U.-D. (2009). Internet experiments: methods, guidelines, metadata. Proceedings of SPIE: Vol. 7240. *Human Vision and Electronic Imaging XIV, 724008*. http://doi.org/10.1117/12.823416. Retrieved from http://iscience.deusto.es/wp-content/uploads/2010/03/preprintSPIE_Reips_A4_color.pdf

Reips, U.-D. (2010). Design and formatting in Internet-based research. In S. Gosling & J. Johnson (Eds.), *Advanced methods for conducting online behavioral research* (pp. 29–43). Washington, DC: American Psychological Association.

Reips, U.-D., Buchanan, T., Krantz, J. H., & McGraw, K. (2016). Methodological challenges in the use of the Internet for scientific research: Ten solutions and recommendations. *Studia Psychologica*. Advance online publication.

Reiss, S., Tillmann, A., Schreiner, M., Schweizer, K., Krömker, D., & Moosbrugger, H. (2009). Online-Self-Assessments zur Erfassung studienrelevanter Kompetenzen [Online-self-assessments to measure study-related competences]. *Zeitschrift für Hochschulentwicklung, 4*, 60–71.

Richman, W. L., Kiesler, S., Weisband, S., & Drasgow, F. (1999). A meta-analytic study of social desirability distortion in computer administered questionnaires, traditional questionnaires, and interviews. *Journal of Applied Psychology, 84*, 754–775. http://doi.org/10.1037/0021-9010.84.5.754

Ritter, P., Lorig, K., Laurent, D., & Matthews, K. (2004). Internet versus mailed questionnaires: A randomized comparison. *Journal of Medical Internet Research, 6*(3), e29 http://doi.org/10.2196/jmir.6.3.e29

Rosen, Y., & Tager, M. (2013). Computer-based assessment of collaborative problem-solving skills: Human-to-agent versus human-to-human approach. *Pearson's Research Reports*. Retrieved from http://www.pearsonassessments.com/research

Ross, J. A. (2006). The reliability, validity, and utility of self-assessment. *Practical Assessment Research & Evaluation, 11*, 1–13.

Russell, D. P. (2007). Recruiting and staffing in the electronic age: A research-based perspective. *Consulting Psychology Journal: Practice and Research, 59*, 91–101. http://doi.org/10.1037/1065-9293.59.2.91

Salgado, J. F., & Moscoso, S. (2003). Internet-based personality testing: Equivalence of measures and assessees' perceptions and reactions. *International Journal of Selection and Assessment, 11*, 194–205. http://doi.org/10.1111/1468-2389.00243

Scheuermann, F., & Guimarães Pereira, A. (Eds.). (2008). *Towards a research agenda on computer-based assessment: Challenges and needs for European educational measurement*. Luxembourg City, Luxembourg: Office for Official Publications of the European Communities. Retrieved from http://publications.jrc.ec.europa.eu/repository/bitstream/111111111/907/1/reqno_jrc44526_report%20final%20version%5B2%5D.pdf

Schulenberg, S. E., & Yutrzenka, B. A. (2001). Equivalence of computerized and conventional versions of the Beck Depression Inventory–II (BDI-II). *Current Psychology, 20*, 216–230. http://doi.org/10.1007/s12144-001-1008-1

Sireci, S., & Zenisky, A. (2006). Innovative item formats in computer-based testing: In pursuit of improved construct representation. In S. Downing & T. Haladyna (Eds.), *Handbook of test development* (pp. 329–347). Mahwah, NJ: Lawrence Erlbaum.

Sonnleitner, P., Keller, U., Martin, R., & Brunner, M. (2013). Students' complex problem-solving abilities: Their structure and relations to reasoning ability and educational success. *Intelligence, 41*, 289–305. http://doi.org/10.1016/j.intell.2013.05.002

Sonnleitner, P., Brunner, M., Greiff, S., Funke, J., Keller, U., Martin, R., … Latour, T. (2012). The genetics lab: Acceptance and psychometric characteristics of a computer-based microworld assessing complex problem solving. *Psychological Test and Assessment Modeling, 54*, 54–72.

Stieger, S., & Reips, U.-D. (2010). What are participants doing while filling in an online questionnaire: A paradata collection tool and an empirical study. *Computers in Human Behavior, 26*(6), 1488–1495. http://doi.org/10.1016/j.chb.2010.05.013

Suler, J. (2004). The online disinhibition effect. *Cyberpsychology and Behavior, 7*, 321–326. http://doi.org/10.1089/1094931041291295

Texas Education Agency. (2008). *A review of literature on the comparability of scores obtained from examinees on computer-based and paper-based tests*. Retrieved from http://www.tea.state.tx.us/WorkArea/linkit.aspx?LinkIdentifier=id&ItemID=2147494120&libID=2147494117

Tippins, N. T. (2009). Internet alternatives to traditional proctored testing: Where are we now? *Industrial and Organizational Psychology, 2*, 2–10. http://doi.org/10.1111/j.1754-9434.2008.01097.x

Vallejo, M. A., Jordán, C. M., Díaz, M. I., Comeche, M. I., & Ortega, J. (2007). Psychological assessment via the Internet: A reliability and validity study of online (vs paper-and-pencil) versions of the General Health Questionnaire-28 (GHQ-28) and the Symptoms Check-List-90-Revised (SCL-90-R). *Journal of Medical Internet Research, 9*(1), e2. http://doi.org/10.2196/jmir.9.1.e2

Villani, D., Gatti, E., Confalonieri, E., & Riva, G. (2012). Am I my avatar? A tool to investigate virtual body image representation in adolescence. *Cyberpsychology, Behavior, and Social Networking, 15*, 435–440. http://doi.org/10.1089/cyber.2012.0057

Wang, H., & Shin, C. D. (2010). Comparability of computerized adaptive and paper-pencil tests. *Pearson Test, Measurement and Research Service Bulletin, 13*, 1–7.

Wang, T. H. (2007). What strategies are effective for formative assessment in an e-learning environment? *Journal of Computer Assisted Learning, 23*, 171–186. http://doi.org/10.1111/j.1365-2729.2006.00211.x

Wang, T. H. (2009). Web-based dynamic assessment: Taking assessment as teaching and learning strategy for improving students' e-Learning effectiveness. *Computers & Education, 54*, 1157–1166. http://doi.org/10.1016/j.compedu.2009.11.001

Weigold, A., Weigold, I. K., & Russell, E. J. (2013). Examination of the equivalence of self-report survey-based paper-and-pencil and internet data collection methods. *Psychological Methods, 18*, 53–70. http://doi.org/10.1037/a0031607

Ybarra, M. L., & Eaton, W. W. (2005). Internet-based mental health interventions. *Mental Health Services Research, 7*, 75–87. http://doi.org/10.1007/s11020-005-3779-8