Chapter 12

An Overview of Innovative Computer-Based Testing

Sebastiaan de Klerk

Abstract Driven by the technological revolution, computer-based testing (CBT) has witnessed an explosive rise the last decades, in both psychological and educational assessment. Many paper-and-pencil tests now have a computer-based equivalent. Innovations in CBT are almost innumerable, and innovative and new CBTs continue to emerge on a very regular basis. Innovations in CBT may best be described along a continuum of several dimensions. Parshall, Spray, Kalohn, and Davey (2002) describe five innovation dimensions in which CBTs can differ in their level of innovativeness: item format, response action, media inclusion, level of interactivity, and scoring method. This chapter provides a detailed description of the five innovation dimensions, including case examples. Furthermore, an overview of opportunities, risks, and future research will be given.

Keywords: computer-based testing, dimensions of innovation, opportunities and risks, future research

Introduction

The availability and utilization of personal computers has been growing explosively since the 1980s, and will continue to do so in the coming decades. The educational system has not been oblivious to the explosive rise of PCs and technology in general. For example, the development of high-speed scanners, or Optical Mark Recognition (OMR), halfway through the 1930s of the 20^{th} century introduced the possibility of automatically scoring multiple-choice tests. More recently, during the late 1970s, the first computer-delivered multiple-choice tests emerged, and computer-based testing (CBT) was born. Further improvements and cost reductions in technology made the application of large-scale, high-stake CBTs during the 1990s possible. Present advances in technology continue to drive innovations in CBT, and new CBTs are being designed on a regular basis by a whole range of educational institutions.

Nowadays, test developers can incorporate multimedia elements into their CBTs, and they can develop innovative item types, all under the continuing influence of technology improvements. This chapter provides an overview of innovations in CBT.

Because innovations in CBT are (almost) innumerable, and continue to emerge in many different forms, a dichotomous categorization of CBTs as innovative versus non-innovative is not possible. More specifically, however, innovation in CBTs may be seen as a continuum along several dimensions. For instance, some CBTs may be highly innovative (scoring innovativeness on multiple dimensions), while other CBTs are less innovative (scoring innovativeness on only one dimension). Inclusion of media (e.g., video, animation, or pictures), test format (e.g., adaptive), item format (e.g., drag- and drop-, matrix-, or ranking and sequencing questions), and construct measurement (e.g., skills or competencies) are all attributes upon which the innovativeness of a CBT can be determined. In general, using PCs or technology to develop creative ways of assessing test takers or to measure constructs that were previously impossible to measure is the most important dimension for innovations in computerized testing. Parshall et al. (2002) introduced five innovation dimensions of CBTs: item format, response action, media inclusion, level of interactivity, and scoring method. Each of the five innovation dimensions will be discussed below.

Dimensions of Innovation

Item Format

The first dimension is the item format, and this dimension makes reference to the response possibilities of the test taker. The multiple-choice item format probably is the most well-known item type, and can also be used in paper-and-pencil tests. Multiple-choice items fall into the category of so-called *selected response* formats. The characterizing feature of these formats is that the test taker is required to select one or multiple answers from a list of alternatives. In contrast, *constructed response* formats require test takers to formulate their own answers, rather than select an answer from a list of alternatives (Drasgow & Mattern, 2006). A fill-in-the-blank item type is an example of constructed response format, but essay questions and short answers are also constructed response items. All of the selected- and constructed-response item types can be administered by computer and, even more importantly, a growing amount of innovative item types are uniquely being designed for CBTs.

Scalise and Gifford (2006) present a categorization or taxonomy of innovative item types for technology platforms. The researchers have identified seven different item formats, and 28 corresponding item examples (four per category) after a profound literature search, and reported these item examples in their paper. Most of the 28 item types are deliverable via a PC; however, some item types have specific advantages when computerized. For example, categorization, matching, ranking and sequencing, and hot-spot items are item types that are most efficiently administered by computer, compared to paper-and-pencil administration. Innovations in item format demonstrate that innovation is actually twofold. On the one hand, we can create new item types to measure constructs differently (improved measurement). On the other hand, we can also create new item types to measure completely different constructs that were difficult to measure before. This will also hold for the other dimensions of innovation, as will become clear in the following sections.

Response Action

The second innovation dimension is response action, and this dimension represents the physical action(s) a test taker has to perform in order to answer a question. The most common response action is of course filling in an answer sheet of a multiple-choice test in a paper-and-pencil test, or mouse clicking in a CBT. However, computerized testing software and computer hardware offer some interesting features for response actions. For example, test takers can also report their answers by typing on the keyboard, or speak them into a microphone (possibly integrated with voice recognition software). These types of response actions can hardly be called innovative nowadays, because they have been available for quite some time now. However, they show the constant progress in educational testing, influenced by the technological revolution.

Response actions in CBTs of skill assessment have been studied for the last two decades, with researchers looking for possibilities to assess skill in a way such that the response action corresponds with the actual skill under investigation. For example, joysticks, light pens, touch screens, and trackballs were used by the test takers as tools for the response actions. This resulted in another stream of innovations in assessment. The current innovations in assessment show that a whole new movement of response actions is emerging.

Researchers are trying to unite response action and skill assessment, for example, through virtual environments, serious gaming, camera movement recognition, simulation software, and other innovative technologies that require test takers to physically perform a range of actions (e.g., a flight simulator). Van Gelooven and Veldkamp (2006) developed a virtual reality assessment for road inspectors. Because traffic density keeps on rising, road inspectors have taken over some tasks that used to be the duty of the traffic police, for instance, signaling to drivers, towing cars, and helping to fill in insurance documents after accidents. The test takers (road inspectors) are confronted with a virtual reality projected on a white screen. The director starts a specific case, and test takers can walk through the virtual environment with a joystick. During the assessment, all sorts of situations or problems develop, and the test takers are required to carry out actions with their joystick in the virtual environment. This example shows how assessments can be designed with innovative use of the response actions (controlling a joystick) a test taker has to perform.

Future innovations in CBT will repeatedly use more of these types of response actions, all the more because they unveil the possibility of measuring constructs that were difficult to measure before, or to measure constructs more accurately than we (as assessment experts) were able to in the past.

Media Inclusion

The third dimension is media inclusion, and indicates to what extent innovative CBTs incorporate (multi)media elements. Addition of media elements to CBTs can enhance the tests' coverage of the content area and may require test takers to use specific (cognitive) skills. Moreover, another key advantage of media inclusion is improved validity. Yet another advantage is that reading skills cannot be considered a confounding variable anymore. Media that are regularly found in CBTs are, among others, video, graphics, sound, and animations. The simplest form is providing a picture with an item stem, as is sometimes the case in paper-and-pencil tests. Ackerman, Evans, Park, Tamassia, and Turner (1999) have developed such a test of dermatological disorders that provides test takers with a picture of the skin disorder. Following presentation of the picture, the test taker is asked to select the disorder from a list on the right side of his screen.

The assessment remains rather “static”; however, it would be a more complex assessment form if test takers had to manipulate the picture provided with the item, for example, by turning it around or fitting it into another picture. Still more difficult are items in which test takers have to assemble a whole structure with provided figures or icons, for example, when they have to construct a model and the variables are provided.

Audio is most often used in foreign language tests, and usually requires test takers to put on headphones. However, other fields have also used audio in (computerized) testing. For example, the assessment of car mechanics sometimes relies upon sound. Test takers have to listen to recorded car engines and indicate which cars have engine problems. In addition, medical personnel are presented with stethoscope sounds during assessment, and they are asked which sounds are unusual. Another innovative application of sound in assessment is to present questions in sound for people who are dyslexic or visually-impaired.

Video and animations are other media elements that may be incorporated into CBTs. These media elements are highly dynamic, and are highly congruent with authentic situations that test takers will face outside of the assessment situation. Several researchers have carried out case studies in which assessment included video. Schoech (2001) presents a video-based assessment of child protection supervisor skills. His assessment is innovative because it incorporates video in the assessment, but it is not highly interactive. The test takers watch a video, and then answer (multiple-choice) questions about the video that they have just watched. Drasgow, Olson-Buchanan, and Moberg (1999) present a case study of the development of an interactive video assessment (IVA) of conflict resolution skills. Because they introduce an innovative idea for making a CBT relatively interactive, their study is described below, in the section about the level of interactivity (the fourth innovation dimension) of a CBT.

Level of Interactivity

Interactivity, the fourth dimension of innovation, indicates the amount of interaction between test taker and test. As such, paper-and-pencil tests have no interaction at all. All test takers are presented with the same set of items, and those do not change during the administration of the test. In contrast, CBTs may also be highly interactive because of an adaptive element. Computerized adaptive tests (CATs) compute which item should be presented to a test taker based upon the answers given to all previous items.

In that way, the CAT is tailored to the proficiency level of the test taker (Eggen, 2008, 2011). CATs are now widely used in assessment (both psychological and educational), but were initially a huge innovation made possible by the explosive growth of PCs and technology, and the introduction of Item Response Theory (IRT).

Another form of interactivity, also based on the concept of adaptive testing, is the incorporation of a two- or multistep branching function, possibly accompanied by video. Drasgow et al. (1999) present such a case study of an innovative form of a CBT. The CBT is structured upon two or more branches, and the answer(s) of the test taker form the route that is followed through the branches. The IVA of conflict resolution skills presented by Drasgow et al. required test takers to first watch a video of work conflict. Test takers then had to answer a multiple-choice question about the video. Following their answers, and depending upon their answers, a second video was started, and the cycle was completed once more. In essence, the more branches you create, the higher the assessment scores on interactivity, because it is highly unlikely that two test takers will follow exactly the same path.

Developing assessments that score high on the interactivity dimension is rather difficult, especially compared to some of the other innovation dimensions. Test developers are required to develop enough content to fill the branches in the adaptive interactive assessment. Another difficulty is the scoring of interactive CBTs. As test takers proceed along the branches of the interactive assessment, it becomes more difficult to use objective scoring rules, because many factors play a role, including weighing the various components of the assessment, and the dependency among the responses of the test taker. However, innovation in the level of interactivity has the potential to open up a wide spectrum of previously immeasurable constructs that now become available for measurement.

Scoring Method

The fifth and final innovation dimension is the scoring method. High-speed scanners were one of the first innovations in automatic scoring of paper-and-pencil multiple-choice tests. Automatic scoring possibilities have been developing rapidly, especially in the last two decades. Innovative items that score relatively low on interactivity and produce a dichotomous score are not too difficult to subject to automatic scoring.

Other innovative CBTs, for example, complex performance-based CBTs, may require scoring on multiple dimensions, and are much more difficult to subject to automatic scoring. In performance assessment, the process that leads to the product is sometimes equal to or even more important than the product itself; however, it is a complicated task to design an automatic scoring procedure for process responses as well as product responses in complex performance-based CBTs.

Consider, for example, the above-mentioned branching of CBTs that incorporate video as well. Response dependency can be an obstructive factor for the scoring of these types of CBTs. This means that test takers' responses on previous items may release hints or clues for subsequent items. An incorrect answer on an item, after a test taker has seen the first video in the IAV, releases another video that may give the test taker a hint to his mistake on the previous item. Another issue is the weighing of items in a multistep CBT. Do test takers score equal points for all items, or do they score fewer points for easier items that manifest themselves after a few incorrect answers by the test taker?

Automated scoring systems also demonstrate some key advantages for the grading process of test takers' responses. The number of graders can be reduced, or graders can be completely removed from the grading process, which will also eliminate interrater disagreement in grading. Researchers have found that automated scoring systems produced scores that were not significantly different from the scores provided by human graders. Moreover, performance assessment of complex tasks is especially costly; molding these assessments into a CBT is extremely cost and time efficient. Thus, computers offer many innovative possibilities for scoring test takers' responses. For example, the use of text mining in assessment or classification is possible because of innovations in computer-based scoring methods. Text mining refers to extracting interesting and useful patterns or knowledge from text documents. This technique provides a solution to classification errors, because it reduces the effects of irregularities and ambiguities in text documents (He, Veldkamp, & Westerhof, this volume). Yet another stream of innovation in scoring lies in test takers' behavior, and results in the scoring or logging of mouse movements, response times, speed-accuracy relationships, or eye-tracking. The key point that flows forth from the five innovation dimensions described above is twofold: not only do test developers become more capable of measuring constructs *better*, they also find themselves in a position to measure *new* constructs that were difficult to measure before.

Opportunities of Innovations in CBT

Performance Assessment and CBTs

One of the main objections to multiple-choice tests is that, although extremely accurate in measuring declarative knowledge, they are difficult to design to measure test takers' skills and abilities. Thus, integrating performance assessment with CBT is an interesting opportunity for innovative CBTs. Several research and educational institutions have started research programs to explore the possibilities of measuring test takers' skills in complex performance-based tasks. Clyman, Melnick, and Clauser (1999) found that the correlation between computer-based case simulations and declarative knowledge assessment was only 0.5. Traditional tests and interactive case simulations presented via the PC therefore measure different domains of knowledge, and even make it possible to measure completely different constructs. The challenge lies in using a PC's capability to provide assessment exercises that revolve around the core of the actual performance of a task. In that way, the link between CBT and performance assessment becomes stronger, and the construct under investigation corresponds in both assessment types. In other words, both assessments are measuring the same construct, although in different ways.

Media in CBTs

Another opportunity in innovative CBTs is the inclusion of media within the CBT. Above, media inclusion is briefly discussed as one of the five innovation dimensions. Media inclusion in itself does not make a CBT innovative. However—and this is really the starting point of adding media into a CBT—being innovative should improve measurement. As mentioned above, it should either enhance the measurement of constructs that are now measured by other assessment types, or it should enable us to measure constructs that we were previously unable to measure.

Consider, for example, the CBT designed by Ackermann et al. (1999), which is briefly described in the section about media inclusion, as one of the five innovation dimensions. The addition of pictures of dermatological disorders in their multiple-choice test made it possible to more effectively measure test takers' knowledge about the skin disorders under investigation. This example represents how media inclusion can enhance or improve the measurement of constructs that were previously measured only by a multiple-choice test. In this case, it is the test takers' ability to recognize and identify different types of skin disorders that can be more accurately measured, but the possibilities that media offers are almost infinite.

Novel Item Types in CBTs

A third opportunity that CBTs offer is the modification of item types, and the creation of novel item types. Again, the central idea is that they should improve the quality of measurement, or introduce new measurement possibilities. Paper-and-pencil tests impose restrictions on item construction. For example, it is almost impossible to administer drag-and-drop questions in a paper-and-pencil test, while computers offer great opportunities to integrate these questions within a CBT. Zenisky and Sireci (2002) describe 21 new types of item formats in their paper. Describing all item types is beyond the extent of this chapter; however, important to note here is that innumerable innovative item types have emerged in the literature on CBT item types in recent decades. Also, adaptations of these item types already exist, which shows that innovations and opportunities are almost countless. Computers offer opportunities such that new item types make it possible to access measurement to constructs that were very difficult or cost- and time-consuming to measure in the past.

Scoring of CBTs

Improved scoring is a fourth and final opportunity that flows forth from computerized tests. The emphasis in this section will be on automated scoring of CBTs that aims to measure performance-based constructs or score performance-based tasks. Typically, these types of assessments require human graders, and sometimes actors or even walk-ons. Cost and time constraints are some serious disadvantages that go hand in hand with performance assessment. Automated scoring systems are being developed, and some are already operational, that make it possible to remove human graders from the scoring process, or at least minimize human intervention in the scoring process. One opportunity is, of course, to design automated scoring systems that make it possible to score test takers' attributes or behaviors in a way that we were not able to do before. Another opportunity is that automated scoring systems are more consistent and are able to extract more information than human graders (for example, text mining: He and Veldkamp (2012).

Moreover, Williamson, Bejar, and Hone (1997, cited in Dodd & Fitzpatrick, 2002) found that human graders mostly let the automated score stand when provided with the analyses that resulted from the automated scoring.

The automated scoring of essays, for example, becomes more and more incorporated as a regular and common method for scoring essays. Burstein et al. (1998) found that there was an 87% to 94% agreement between the scores on essays that resulted from automated scoring systems and the scores awarded by the human graders. Above that, and equally important, this is not different from having two human graders score an essay independently. These findings support the idea that automated scoring systems are also applicable in the scoring of rather complex performance-based tasks. One concern that results from applying automated scoring systems in performance-based assessments is that it gives test takers the opportunity to actually lure the system into assigning a higher grade, for example, by using specific words that are scored by the system, but not properly used in the context of the essay. This is one risk that originates from the utilization of automated scoring systems, and innovations in CBTs in general. In the following section, other risks that might affect CBTs are discussed.

Risks of Innovations in CBT

CBTs offer a lot of opportunities, and the educational field can greatly benefit from these opportunities. However, every medal has two sides, and the other side of CBTs is that they are also subjected to several (potential) risks. For example, CBTs can be very costly and difficult to develop. Moreover, it may take a substantial amount of time to validate a CBT and, even if validated, questions about the efficiency of the CBT still remain. Finally, another threat to proper functioning of a CBT is test security and test disclosure.

Development

Guidelines on the development of classical multiple-choice tests are widely available in innumerable textbooks and publications. In contrast, such guidelines on the development of innovative CBTs did not exist up till now. Although many different guidelines exist that regulate the development of multiple-choice tests, it still takes a lot of time to develop a valid and reliable multiple-choice test. Consider the development of an innovative CBT without a frame of reference or any strong guidelines on these types of CBTs. It should now be possible to imagine that the development of a valid and reliable innovative CBT will require a lot of creativity, educational assessment expertise, and hard work. Such a process may endure multiple years and, by the time the assessment is completely developed, it may be outdated.

Or as Schoech (2001) subtly notes: Walk the fine line between current limitations and potentials. Exam development is often a multi-year process, yet technology changes rapidly in several years. Thus, a technology-based exam has the potential to look outdated when it is initially completed.

Validity of Scoring

Another risk imposed on innovations in CBTs is validity. Most concerns raised about the validity of CBTs are related to the scoring procedure. If test developers create CBTs that automatically score constructed responses (e.g., scoring of essays by a computer), they have to be cautious about some pitfalls. For example, Powers et al. (2002) invited several parties in their study to “challenge” (i.e., try to trick) *e-rater*, which is an automated essay scorer. Participants’ essays were also graded by humans, and the difference between the two types of scoring served as a dependent variable. The researchers found that the challengers were more successful in tricking the *e-rater* to award them higher scores than they deserved, based upon the human graders, than tricking the *e-rater* to award them lower scores than they deserved based upon the human graders. This short example stresses the importance of further research on automated scoring of (complex) tasks in CBTs.

Test Security

One of the most significant risks that stems from CBTs is test security. First of all, the ICT structure needs to be sufficient, thereby making it impossible for potential cheaters to hack a computer or network to get access to the questions in an item bank. Also, Internet-based CBTs make it possible to have test takers take tests at their own convenience, independent of time and place. This may expose the test to other test takers, and thereby impose a serious threat to test security. Another possible threat to the security of a test is test or item disclosure. Disclosure of a whole test or even several items may result in higher scores for subsequent test takers (Drasgow, 2002).

Finally, remaining in the same item bank for extended periods of time may result in diminished security of the test, because test takers are able to write down items, and may distribute them via the Internet, for example.

Future Research

The focus of future research in CBT will be on improving the measurement of skills and performance abilities. Computers enable test developers to create high-fidelity computer simulations that incorporate innovations on all of the dimensions discussed above. Those types of CBTs are designed with the goal of measuring skill and demonstrating performance. Additionally, they correspond to actual task performance to a great extent, which is defined as the authenticity of an assessment. Therefore, these CBTs rely more upon the concept of authenticity than multiple-choice tests do, for example. Integration of multimedia, constructed response item types, highly interactive designs, and new (automatic) scoring methods will lead to an assessment form that closely approximates performance assessment in its physical form. Future research should determine to what extent it is possible to have computers adopt tasks that were usually part of performance assessment.

Furthermore, the psychometric performance of innovative CBTs that integrate many of the innovation dimensions should be determined. For example, studies on the validity of CBTs that are based on the innovation dimensions are all but absent in the current literature. Other studies should try to make innovative CBTs equivalent with performance assessment. Reliability has always been an important issue in performance assessment, which usually relies upon human graders. Innovative CBTs may be able to integrate high reliability with authentic assessment forms. Therefore, researchers should also focus on the concept of reliability in innovative CBTs. Automated scoring systems that rely upon IRT algorithms, or the integration of computerized adaptive testing (CAT), are other interesting future research directions within CBT.

The research on innovations in CBTs that has been reported in the scientific literature mainly focuses on cognitive performance tasks, usually in higher education. Some case studies exist that have tried to measure particular constructs in skill-based professions, for example, in the medical professions or ICT. However, future research should also focus on measuring skill constructs in vocational professions that rely upon physical skills rather than cognitive or intellectual skills.

The continuing technological revolution makes it possible for test developers to further innovate, create, and revolutionize CBTs. The coming decade will be very interesting for the educational measurement field, and there is a whole new range of CBTs to look forward to.

References

Ackerman, T.A., Evans, J., Park, K.S., Tamassia, C., & Turner, R. (1999). Computer assessment using visual stimuli: A test of dermatological skin disorders. In F. Drasgow & J.B. Olson-Buchanan (Eds.), *Innovations in computerized assessment* (pp. 137-150). Mahwah, NJ: Erlbaum.

Burstein, J., Braden-Harder, L., Chodrow, M., Hua, S., Kaplan, B., Kukich, K., et al. (1998). *Computer analysis of essay content for automated score prediction* (ETS Research Report RR-98-15). Princeton, NJ: Educational Testing Service.

Clyman, S.G., Melnick, D.E., & Clauser, B.E. (1999). Computer-based case simulations from medicine: Assessing skills in patient management. In A. Tekian, C.H. McGuire, & W.C. McGahie (Eds.), *Innovative simulations for assessing professional competence* (pp. 29-41). Chicago, IL: University of Illinois, Department of Medical Education.

Dodd, B.G., & Fitzpatrick, S.J. (2002). Alternatives for scoring CBTs. In C.N. Mills, M.T. Potenza,

Drasgow, F. (2002). The work ahead: A psychometric infrastructure for computerized adaptive tests. In C.N. Mills, M.T. Potenza, J.J. Fremer, & W.C. Ward (Eds.), *Computer-based testing: Building the foundation for future assessments* (pp. 1-35). Mahwah, NJ: Erlbaum.

Drasgow, F., & Mattern, K. (2006). New tests and new items: Opportunities and issues. In D. Bartram & R.K. Hambleton (Eds.), *Computer-based testing and the internet: Issues and Advances* (pp. 59-75). Chichester: Wiley.

Drasgow, F., Olson-Buchanan, J.B., & Moberg, P.J. (1999). Development of an interactive videoassessment: Trials and tribulations. In F. Drasgow & J.B. Olson-Buchanan (Eds.), *Innovations in computerized assessment* (pp. 177-196). Mahwah, NJ: Erlbaum.

Eggen, T.J.H.M. (2008). Adaptive testing and item banking. In J. Hartig, E. Klieme, & D. Leutner (Eds.), *Assessment of competencies in educational contexts* (pp. 215-234). Götting en: Hogrefe.

Eggen, T.J.H.M. (2011, October). *What is the purpose of CAT?* Presidential address at the 2011 Meeting of the International Association for Computerized Adaptive Testing, Monterey, CA.

Fremer, J.J. & W.C. Ward (Eds.), *Computer-based testing: Building the foundation for future assessments* (pp. 215-236). Mahwah, NJ: Erlbaum.

Gelooven, D. van, & Veldkamp, B. (2006). Beroepsbekwaamheid van weginspecteurs: een virtual reality toets. In E. Roelofs & G. Straetmans (Eds.), *Assessment in* *actie:Competentiebeoordeling in opleiding en beroep* (pp. 93-122). Arnhem, the Netherlands: Cito.

He, Q & Veldkamp, B.P.(2012). Classifying unstructured textual data using the Product Score Model: an alternative text mining algorithm. In T.J.H.M. Eggen & B.P. Veldkamp (Eds.), *Psychometrics in Practice at RCEC:* (pp. 47-63). Enschede, Netherlands: RCEC.

Parshall, C.G., Spray, J.A., Kalohn, J.C., & Davey, T. (2002). *Practical considerations in computer-based testing.* New York, NY: Springer.

Powers, D.E., Burnstein, J.C., Chodorow, M., Fowles, M.E., & Kukich, K. (2002). Stumping *e-rater*: Challenging the validity of automated essay scoring. *Computers in Human Behavior,* 18, 103-134.

Scalise, K., & Gifford, B. (2006). Computer-based assessment in e-learning: A framework for constructing "intermediate constraint" questions and tasks for technology platforms. *Journal of Technology, Learning, and Assessment, 4*(6). Retrieved [March 20, 2012] from http://www.jtla.org.

Schoech, D. (2001). Using video clips as test questions: The development and use of a multimedia exam. *Journal of Technology in Human Services, 18*(3-4), 117-131.

Zenisky, A.L., & Sireci, S.G. (2002). Technological innovations in large-scale assessment. *Applied Measurement in Education, 15*, 337-362