Some Critical Reflections on the Special Issue: Current Innovations in Computer-Based Assessments

Samuel Greiff

University of Luxembourg, Luxembourg

Ronny Scherer

University of Oslo, Norway

Paul A. Kirschner

The Open University of the Netherlands, The Netherlands;

University of Oulu, Finland

Author Note

Samuel Greiff, Institute of Cognitive Science and Assessment (COSA), University of Luxembourg, 11, Porte des Sciences, 4366 Esch-sur-Alzette, Luxembourg, Email: samuel.greiff@uni.lu; Ronny Scherer, University of Oslo, Faculty of Educational Sciences, Centre for Educational Measurement at the University of Oslo (CEMO), Postbox 1161 Blindern, 0318 Oslo, Norway; Email: ronny.scherer@cemo.uio.no; Paul A. Kirschner, The Open University of the Netherlands, Heerlen, The Netherlands, Valkenburgerweg 177, 6419 AT Heerlen, The Netherlands; University of Oulu, Learning & Educational Technology Research Unit, Oulu, Finland; Email: paul.kirschner@ou.nl.

Some Critical Reflections on the Special Issue: Current Innovations in Computer-Based Assessments

Introduction

Technology and computers change and penetrate our lives to an extent that was unthinkable 30 years ago, and it is the mission of Computers in Human Behavior to advance our knowledge on how humans interact with, make use of, and are influenced by computers. The nine papers in this special issue “Current Innovations in Computer-Based Assessment” reflect the innovations and advances computers have brought to the ways we assess psychological attributes of a variety of populations including primary and secondary students, students in higher education, or adults in the work force.

This special issue covers four broad areas within the field of computer-based assessment (CBA): Assessment of new constructs or widening the assessment of existing constructs; use of log-file and multi-channel data; psychometric models and experiments that inform the measurement of complex skills and task construction; integration of assessment and learning. Obviously, these areas are neither representative in and by themselves, nor are they comprehensively and fully covered within this limited selection of nine papers. Even more importantly, the results reported in the contributions of this special issue do not represent a final destination, but are an intermediate step on a long journey to our understanding of computer-based assessment in which we are only making the first steps. In this discussion, we, as the guest editors of this special issue, highlight some points that we consider paramount to the development of the field and that might require close attention in the near future, namely: (1) Application across psychological sub-disciplines; (2) Adequate methodological approaches; (3) Theoretical and empirical foundation and validation; (4) Integration of assessment and learning, stealth assessment, and modern test design; and (5) Establishing CBA-specific, cognitive theory.

Application Across Psychological Sub-Disciplines

Assessments in one form or the other are relevant to virtually all sub-disciplines of psychology in that they are: the foundation of diagnosis and treatment in clinical psychology; used to identify gifted and special-needs students in education; widely employed in personnel selection and human resource development; and at the very heart of personality research and, in fact, any research on the human intellect. Interestingly, the use of CBAs as tool of innovation differs substantially across sub-disciplines: Educational psychology currently experiences a comprehensive shift toward CBAs that includes innovative item formats and advanced scoring procedures. This is also reflected in international large-scale assessments such as the Programme for International Student Assessment (PISA), in which innovative assessments have been implemented in over 50 countries and have found their way into widely perceived policy reports (OECD, 2014). In other areas, computer-based tools are used but with a smaller focus on assessment. For instance, complex simulations are used in industry contexts to increase safety (Kluge, Badura, & Rietz, 2013) and computer-based therapy is successfully offered to patients suffering from anxiety or depression (Ebert et al., 2015). Other areas use the computer only as an easier and more efficient tool to administer isomorphic adaptations of paper-and-pencil assessments, largely ignoring the potential computers offer as new tools to understand the very nature of assessment. The papers in this special issue also mirror the high prevalence of contributions in the field of educational psychology, but the innovations and the potential does not exclusively apply to a single or very few sub-disciplines. In fact, now emerging methods such as ambulatory assessment (Santaneglo, Bohus, & Ebner-Primer, 2014) and the wealth of additional information, for instance on test-taking processes, found in CBAs (Greiff, Niepel, Scherer, & Martin, 2016) are of relevance across sub-disciplines. However, this relevance is in stark contrast to the actual extent to which the potential of CBAs is exploited in some sub-disciplines. We consider the exploitation and integration of CBAs as a crucial development in the field over the next years and would welcome more contributions that provide integrations to other areas and that are not primarily or exclusively focused on educational psychology.

Adequate Methodological Approaches

Compared to standard paper-and-pencil assessment instruments, CBAs differ on several aspects. One of the most obvious differences is that CBAs offer rich data sets not only on performance (i.e., the correctness of responses), but also on the (observable) steps taken towards problem solutions (e.g., test-taking behavior). Given that multiple pieces of information are available for a single task (e.g., correctness of response, response time, sequence of activities), these data sources are usually nested within the tasks. This is not the case in paper-and-pencil assessments and, importantly, does not fit the frame of classical methodological approaches that assume independence between pieces of information. That is, the different modalities and the availability of log-file data that contain considerably more information than merely the correctness of an answer, including information on the number of mouse clicks and interactions, timing, and sometimes even eye movement, audio and video, cannot be handled within classical psychometric models; for example, what is clicked on by a learner or where (s)he is looking (i.e., her/his eye movements) is not independent of the answer that a person may have given and/or its correctness. Also, collaborative scenarios in which long sequences of actions are scored (von Davier, 2017) cannot be adequately integrated within more standard methodological approaches. Hence, more complex psychometric approaches are needed to describe the underlying (unobservable) constructs and these approaches need to take into account the complex interactions between students and task that take place and that often are considered an integral part of CBAs.

Research on methodology has made some headway over the last decade in providing suitable methods for the complex data patterns usually found in CBAs, such as response time-item response theory models (van der Linden, 2009) and data mining procedures (Baker, Martin, & Rossi, 2017), to name a few. Indeed, some contributions in this special issue tackle this topic, but there is still a long way to go until a comprehensive set of methods that can be adequately applied to the wealth of data from CBAs is available – both on a substantive-methodological and on a pragmatic-implementation basis.

Theoretical and Empirical Foundation and Validation

Despite the danger of stating the obvious, CBAs of psychological attributes require strong conceptual and empirical evidence regarding the psychological target construct, and this evidence cannot be inferred from other, non-CBA instruments. Put differently, developing items for CBAs, particularly when these items are complex and innovative, must be accompanied by evidence on the link between what they actually measure and what the assessment perpetuates, that is evidence on the validity of the assessment instrument, which is at the very heart of every measure. There are multiple approaches of how such evidence can be obtained including studies of student cognitive protocols, item and test performance, test-taking behavior based on log files, match between student learning and classroom instruction, and so forth (cf. Greiff & Iliescu, 2017; Pellegrino, DiBello, & Goldman, 2016). Of note, there has been a long and intensive discussion about how the mere transfer from a paper-and-pencil version of a test to a computer-based version of a test might change the underlying construct and the meaning of the scores (Mead & Drasgow, 1993). Obviously, this holds even more so for assessments that integrate multiple data channels, contain innovative item formats, or make use of log files. That is, the danger of any innovative technology is that the mere drive for innovation supersedes content and quality and, in this, muddies the clarity of the variables assessed. We consider it a threat to the validity of CBAs if the development of them is mainly driven by technological invention that will lead to pragmatic solutions without the necessary level of substantiation.

Integration of Assessment and Learning, Stealth Assessment, and Modern Test Design

Since the advent of computers, integration of assessment and learning within a single activity that serves two purposes has been highlighted as one, if not the major asset of the new technology (Williamson, Mislevy, & Bejar, 2006). Usually, assessment of learning (i.e., summative assessment) and assessment for learning (i.e., formative assessment) are distinguished in this context (for a recent overview on the benefit and the effectiveness of CBA for learning purposes consult Shute & Rahimi, 2017). The idea behind the drive for integration between the two is that in learning environments - mostly in the field of education and educational psychology - assessment could take place simultaneously with learning, possibly without the test taker even noticing (Shute & Ventura, 2013; Shute, Wang, Greiff, Zhao, & Moore, 2016). While this idea is very appealing, there are also good reasons why learning and assessment are kept apart. For instance, the theoretical frameworks for learning on the one hand and assessment on the other hand, in particular regarding some of the psychometric requirements, might not always sit well with each other and, in the end, create a situation in which neither learning nor assessment can be reliably captured (see also the next discussion point). In fact, there are great intelligent tutoring systems out there that are firmly based on theories of learning and cognitive science, but with little allusion to psychometrics and current standards of assessment and vice versa (cf. Koedinger, Corbett, & Perfetti, 2012). This might not be too much of an issue in situations where the assessment only serves the overarching purpose of learning, but at the very moment students (or test-takers in general) face any consequences in light of their performance, the soundness of the assessment approach becomes paramount. Even further, several new questions arise regarding stealth assessment, an evidence-centered, design-based assessment approach, in which assessments are directly and invisibly implemented into gaming environments (Shute & Rahimi, 2017). For instance, is it ethical and fair to score behaviors without telling students that not only their final and/or partial answers but also their actions will be relevant to their score? The PISA 2012 problem-solving assessment provides a good example for this: Certain behaviors and strategies have been shown to be beneficial for successfully solving computer-simulated complex-problem environments (Greiff, Wüstenberg, & Avvisati, 2015). However, students are only instructed to explore the problem situation, but are not told that their strategy is also scored. A second example taps the development of assessments “on the spot”: Using modern test design techniques and drawing from the psychometric advances of computer adaptive testing, tests can be individually tailored to test takers to increase the accuracy and efficiency of measurement (Zenisky & Luecht, 2016). At the same time, these approaches are based on test-takers' data that are extracted from their performance on a set of items. Two issues arise: First, once again, these data are evaluated implicitly, that is, without letting the test taker know. Second, test results and designs might no longer be fully comparable across students (e.g., due to different test lengths and items taken by different students), thus questioning test fairness. Again, this might be acceptable if the stakes are low for the students (e.g., in purely research-driven endeavors); yet, the moment there are practical implications the question of fairness arises. So, while we agree on the general potential of integrating learning and assessment in computer-based environments, this idea also faces potentially serious adverse effects, which do not always receive the attention they deserve or are shoved too easily aside.

Establishing CBA-Specific Cognitive Theory

Computers have sparked an entire new era of tools not only targeted at assessment, but also targeted at learning and instruction. Developments in the field of learning and instruction have been firmly based into theories on multimedia learning such as the cognitive theory of multimedia learning (CTMML; Mayer, 2005). Put differently, the way multimedia tools for learning and instruction are designed and how they are used is driven by evidence-based theory and research on design principles. However, this is not the case for CBA. Kirschner, Park, Malone, and Jarodzka (2016) argue that a comparable development has yet to take place in the field of CBA and for the type of tools and design principles that are (implicitly or explicitly) implemented into computer-delivered assessment tools. In fact, multimedia learning tools and the theory that underlies them have repeatedly shown how relevant design principles are for learning, for instance when it comes to the different types of cognitive load that a learning task might place on students (Plass, Moreno, & Brünken, 2010). Kirschner and colleagues highlight that it is not possible to draw direct inferences from theories motivated by a learning perspective to the multimedia elements employed in CBA. Indeed, CBAs need their own cognitive theory that informs the design and the use of these instruments, for instance to keep balanced the different kinds of cognitive demands, and this holds for both the assessment of and the assessment for learning. From our point of view, we currently know way too little about how multimedia in a very general sense impacts - and sometimes maybe even threatens - the assessment process and its validity. Thus, there is a rather urgent need for an integrated and comprehensive theoretical foundation that drives the design and the setup of CBAs and provides guidance for the entire process of developing, employing, interpreting, and making use of computer-delivered assessment instruments.

Conclusion

This discussion is meant to be a critical piece. The quality of the contributions to this special issue speak for itself, and there is no need to praise them any further here. Nevertheless, there is a need to point out that the field suffers from some blind spots, and that much work remains to be done. There is arguably a rather broad agreement that the field of

CBA is both an important and an interesting one, but its current state could be compared to complicated neuro-degenerative diseases such as dementia: While medical scientists slowly begin to understand the underlying mechanisms, there is hardly anything that can be done at bedside. Similarly, scientists – and this special issue is a great example of this current state – begin to understand the mechanisms and the potential underlying CBA, but broader applications and larger exploitation are still some distance away. This, however, is by no means meant as a discouragement but rather an encouragement to the field, and some big leaps as well as steady development are likely to continue throughout the near future. We conclude with the hope that this special issue serves as food for thought and inspiration to you and your research, both in terms of which important knowledge has already been established and which areas deserve further attention.

References

Baker, R. S., Martin, T., & Rossi, L. M. (2017). Educational data mining and learning analytics. In A. A. Rupp & J. P. Leighton (Eds.), *The Wiley handbook of cognition and assessment* (pp. 379-396). Chichester, West Sussex: John Wiley & Sons, Ltd. doi:10.1002/9781118956588.ch16

Ebert, D. D., Zarski, A.-C., Christensen, H., Stikkelbroek, Y., Cuijpers, P., Berkinh, M., & Riper, H. (2015). Internet and computer-based cognitive behavioral therapy for anxiety and depression in youth. A meta-analysis of randomized controlled outcome trials. *PLoS One, 10*. doi:10.1371/journal.pone.0119895

Greiff, S. & Iliescu, D. (2017). A test is much more than just the test. Some thoughts on adaptations and equivalence. *European Journal of Psychological Assessment, 33*, 145-148. doi:10.1027/1015-5759/a000428

Greiff, S., Niepel, C., Scherer, R., & Martin, R. (2016). Understanding students' performance in a computer-based assessment of complex problem solving. An analysis of behavioral data from computer-generated log files. *Computers in Human Behavior, 61*, 36-46. doi:10.1016/j.chb.2016.02.095

Greiff, S., Wüstenberg, S., & Avvisati, F. (2015). Computer-generated log-file analyses as a window into students' minds? A showcase study based on the PISA 2012 assessment of problem solving. *Computers & Education, 91*, 92-105. doi: 10.1016/j.compedu.2015.10.018

Kirschner, P. A., Park, B., Malone, S., & Jarodzka, H. (2016). Toward a cognitive theory of multimedia assessment (CTMMA). In J. M. Spector, B. B. Lockee, & M. D. Childress (Eds.), *Learning, design, and technology* (pp. 1-23). Cham: Springer. doi:10.1007/978-3-319-17727-4_53-1

Kluge, A., Badura, B., & Rietz, C. (2013). Communicating production outcomes as gains or losses, operator skill and their effects on safety-related violations in a simulated production context. *Journal of Risk Research, 16,* 1241–1258. doi:10.1080/13669877.2013.788059

Koedinger, K. R., Corbett, A. C., & Perfetti, C. (2012). The knowledge-learning-instruction (KLI) framework. Bridging the science-practice chasm to enhance robust student learning. *Cognitive Science, 36*, 757-798. doi:10.1111/j.1551-6709.2012.01245.x

Mayer, R. E. (2005). *The Cambridge handbook of multimedia learning*. New York, NY: Cambridge. doi:10.1017/cbo9780511816819

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

OECD (2014). *PISA 2012 results: Creative problem solving*. Paris: OECD Publishing. doi: 10.1787/9789264208070-en

Pellegrino, J. W., DiBello, L. V., & Goldman, S. R. (2016). A framework for conceptualizing and evaluating the validity of instructionally relevant assessments. *Educational Psychologist, 51*, 59-81. doi:10.1080/00461520.2016.1145550

Plass, J. L., Moreno, R., & Brünken, R. (2010). *Cognitive load theory*. New York, NY: Cambridge University Press. doi:10.1017/cbo9780511844744.001

Santangelo, P., Bohus, M., & Ebner-Priemer U. W. (2014). Ecological momentary assessment in borderline personality disorder. A review of recent findings and methodological challenges. *Journal of Personality Disorders, 28*, 555-576. doi:10.1521/pedi_2012_26_067

Shute, V. J., & Rahimi, S. (2017). Review of computer-based assessment for learning in elementary and secondary education. *Journal of Computer Assisted Learning, 33*, 1-19. doi:10.1111/jcal.12172

Shute, V. J., & Ventura, M. (2013). *Measuring and supporting learning in games. Stealth assessment.* Cambridge, MA: The MIT Press.

Shute, V., Wang, L., Greiff, S., Zhao, W., & Moore, G. (2016). Measuring problem solving skills via stealth assessment in an engaging video game. *Computers in Human Behavior, 63*, 106-117. doi:10.1016/j.chb.2016.05.047

Van der Linden, W. J. (2009). Conceptual issues in response-time modeling. *Journal of Educational Measurement*, *46*, 247-272. doi:10.1111/j.1745-3984.2009.00080.x

Von Davier, A. A. (2017). Computational psychometrics in support of collaborative educational assessments. *Journal of Educational Measurement*, *54*, 3-11. doi:10.1111/jedm.12129

Williamson, D. M., Mislevy, R. J., & Bejar, I. I. (Eds.). (2006). *Automated scoring of complex tasks in computer-based testing*. Mahwah, N.J.: Lawrence Erlbaum Associates.

Zenisky, A. L., & Luecht, R. M. (2016). The future of computer-based testing. In C. S. Wells & M. Faulkner-Bond (Eds.), *Educational measurement. From foundations to future* (pp. 221-238). New York, NY: Guilford Press.

Acknowledgements

The guest editors would like to thank the journal managers and the journal staff of *Computers in Human Behavior* for their support in all practicalities. The guest editors would also like to thank the authors and reviewers without whom this special issue would have never come to pass.

This research was supported by the FINNUT Young Research Talent Project Grant (NFR-254744 “ADAPT21”) awarded to Ronny Scherer by The Research Council of Norway, and by the Fonds National de la Recherche Luxembourg (ATTRACT “ASKI21”) awarded to Samuel Greiff. All statements expressed in this article are those of the authors and do not reflect the official opinions or policies of the authors’ host affiliations or any of the supporting institutions.