Personnel Assessment and Decisions

Volume 5
Issue 3 *Advanced Technologies in Assessment: A Science–Practice Concern*

Article 1

2019

Introduction to the Special Issue on Advanced Technologies in Assessment: A Science-Practice Concern

Tara S. Behrend
The George Washington University

Richard N. Landers
University of Minnesota

Introduction to the Special Issue on Advanced Technologies in Assessment: A Science–Practice Concern

Tara S. Behrend[1] and Richard N. Landers[2]

1. The George Washington University
2. University of Minnesota

Technological innovation in personnel selection and assessment has developed rapidly, with new products now introduced to the market continuously. At present, providers are pitching assessments that make use of new technologies like gamification, artificial intelligence, virtual reality, and biometric measures. These products are first scrutinized in the public eye, and questions have been raised about their efficacy, fairness, and legality. In some ways, the promises and pitfalls of these new technologies are no different from those that emerged from technologies now a generation old that have become comfortable and familiar. In other ways, however, new technological capabilities have introduced unexpected challenges and raised special concerns we have never before needed to consider.

For example, many modern assessments now make use of *trace data*, which refers to behavioral data collected from user interactions with assessment tools, such as mouse movements, facial expressions, word choice, or reaction times. Although the collection of trace data has been possible for quite some time, recent advances in data science technologies have increased the potential financial return for their collection and use. As a result, trace data are now sometimes used in complex evaluative algorithms that are not transparent to users, and sometimes not even transparent to assessment practitioners. The public has become both fascinated and concerned with how such data are used in assessment as the industry expands with significant speed, far beyond the speed of academic research and also often beyond the understanding of the ostensibly responsible practitioners. This pattern of rapid invention and deployment is a common and understandable one for technologists (i.e., Gartner's hype cycle; Dedehayir & Steinert, 2016) but is less commonly understood in the world of talent assessment.

To better tackle intersection points like these between talent assessment and technology, we have curated five articles in this special issue of *Personnel Assessment and Decisions.* In our call for papers, we described the lack of alignment between science and practice, and sought papers that took rigorous empirical approaches to understanding these issues. Each article in the present issue thus considers related aspects of recent assessment technologies: artificial intelligence, asynchronous video interviews, natural language processing, crowdsourced ratings, and responsive web design.

Articles and Emergent Themes

One of the most significant challenges for talent assessment academics and practitioners is to understand the value proposition for emergent assessment technologies, to determine where investment is worthwhile. Whereas Myspace and fax machines have all but disappeared from the modern cultural landscape, for example, PDFs and smartphones are clearly here to stay. Yet such clarity is difficult to gain before the market surrounding these technologies has settled, and this often takes several years after their initial introduction. We contend that the role of researchers is not simply to wait until these technologies have been evaluated by the market, only afterward to evaluate their claims more rigorously, but instead to (begin to) evaluate them as they are introduced and perhaps even inform their development in the first place. Talent assessment technology research should inform the development of talent assessment technology, and the only way to do that is to remain at the forefront of such introductions, asking questions that will themselves lead to better technologies.

The authors in this special issue were encouraged to evaluate emerging technologies to this end; we identified three overarching themes in their approaches. First, three papers explored older assessment technologies that are being remade due to recent technological progress, such as the evolution of in-person interviews into synchronous video interviews into asynchronous video interviews. Second, three papers explored truly novel technologies that are currently being used for assessment in ways that less clearly resemble existing approaches. Third, four papers touched on the psychometric concerns associated with assessment using these technologies, exemplifying the use of estab-

lished evaluative techniques to these new technologies, whether evolutions or novel entities. Each will be described in turn.

Theme 1: The Old Made New. Basch and Melchers (this issue) explore how people's reactions to asynchronous interviews change as a function of how the organization justified using them instead of traditional interviews. Hickman, Tay, and Woo (this issue) similarly examine interviews but examine from the perspective of convergence of personality scoring using a natural language processing API versus self-report methods. Grelle and Gutierrez (this issue) tackled the problem of redesigning traditional measures for new formats using responsive design techniques, a critical evolution of simple binary comparisons between "traditional" and "new" assessment technologies. Each of these papers carefully considers how traditional assessment methods can, do, and should change as a function of new technologies.

Theme 2: Authentically Novel Technologies. In contrast to the first theme, the second considers new capabilities created by new technology. Gonzalez, Capman, Oswald, Theys, and Tomczak (this issue) discuss the potential of artificial intelligence, and more specifically machine learning, in talent assessment broadly. Machine learning can be considered an evolution of traditional statistical approaches, but it also fundamentally changes the types of prediction questions that can be asked in validation, creating new possibilities for assessment not previously explored. In the context of interviews, which is certainly an old assessment technology, Hickman et al. (this issue) explore a novel statistical approach from data science, natural language processing. Importantly, they explicitly explore the role of database curation and algorithm design that lead to the predictions made by commercial natural language processing APIs, and in doing so highlight how the details and implications of these design choices are not always clear. Landers, Brusso, and Auer (this issue) examine the validity of crowdsourced ratings as an alternative to traditional survey-based data collection, reflecting a source of data not commonly seen in the assessment literature. Each article explores a novel technology currently used in practice, one that was born of new technological innovation but for which there is limited scholarly research and mixed practitioner expertise.

Theme 3: Applying Psychometrics. Gonzalez et al. (this issue) argue convincingly that we must maintain traditional psychometric rigor in AI approaches and understand the various sacrifices and trade offs inherent to an AI-based approach. Hickman et al. (this issue) attempt to explore NLP from the perspective of psychometrics but find that convergent validity as we usually understand it might mean something different in this context. Grelle and Gutierrez (this issue) use a traditional measurement equivalence framework in their study but informed by a responsive design mindset. Landers et al. (this issue) approach convergence not at the individual but group level, using this to validate Internet-sourced organizational ratings. Each of these studies explores ways we can use psychometrics to understand new technologies but also illustrates how somewhat broader and more flexible approaches are needed than are often employed.

The Big Picture: Agenda for Future Work

As we edited this special issue, several broader ideas about the "big picture" state of our research literature and its potential future emerged that we believed important to consider moving forward. Specifically, as technology research increases in the context of the talent assessment literature, "staying the course" in relation to established and comfortable research methodologies could cause significant long-term damage in terms of developing an authentic understanding of the technologies involved. We present these concerns here.

First, in the present environment of rapid evolution and change, it is clear that traditional equivalence studies cannot be the way that we accumulate evidence about assessment technologies. Despite significant flaws in such research questions, it is still common to ask simple equivalence questions like, "How are computer-based assessments different from paper-and-pencil assessments?" and "How are skype interviews different from in-person interviews?" Even asking such questions conveys limited appreciation and expertise in how the design of technologies influence the answers to such questions. We simply cannot continue to ask questions in this way; it is a waste of everyone's time, from researcher to technologist. Technologies are not psychological constructs (Landers & Behrend, 2017); each is a tool constructed by humans for a particular purpose, and the particular combination of effects created by that construction process and final product means that the direct comparison of any two technologies often masks hundreds or thousands of meaningful smaller effects. The only way to avoid this is to carefully control the design process to target equivalence (cf. Grelle and Gutierrez, this issue) and then engage in redesign until that goal is achieved, or even further, to realize that equivalence is often not the goal of assessment development and therefore should not be evaluated on that standard. Our theories of assessment are simply not sufficiently developed to describe differences between technologies as they presently exist, and our research methods are often uninformative to meaningful, practical questions. There are many differences within the category of "video interview," for instance, that dramatically change a candidate's experience. Which variables are collected? How are they collected? How is this information combined to make a decision? How is information communicated to a

candidate? What are the physical and social aspects of the environment in which the data are collected? How was the interview experience designed in each case? What processes if any were used to ensure a similar experience between the two? All of these issues have profound effects on candidate reactions, reliability, and validity.

Second, developers of advanced employee selection technologies are confronted with a number of practical questions that our current literature does little to address. A popular model of building AI-driven selection tools, for example, is to start with a very large number of variables, from trace data or otherwise, and develop a model to determine which of those many variables relate to a criterion. Then in a second step, any variables that appear to contribute to adverse impact or have other undesirable characteristics are systematically removed, leaving a final model that appears to maximize predictive validity while meeting fairness standards. However, optimizing a model this way may have other costs that negate the value of using the assessment in the first place; the uncritical removal of all "problematic" variance may harm predictive validity such that these measures are even less effective than traditional construct-based assessment approaches. The question of incremental validity becomes all important, as do questions of opportunity costs and unintended consequences.

The lack of cross-area expertise that leads to this situation reflects a broader problem of researcher interdisciplinary fluency. We have an obligation to conduct research on topics that do not waste our participants' or the public's time, and this obligation is not met when different sets of researchers are studying the same problems from different perspectives without communicating with each other. The most carefully designed study is of no value if it answers the wrong question. How do we make sure that we are not wasting resources, whether time or money? The landscape of researchers exploring questions of technology and assessment is diverse, and people often use different vocabularies to describe similar phenomena. Becoming fluent in multiple fields is challenging, but it is essential that the talent assessment community foster boundary-spanning work, incorporating human resources, industrial-organizational psychology, data science, and human–computer interaction. This enables us to not only advise better decision making to practice by borrowing from other fields but also to meaningfully share that knowledge back across disciplinary barriers to build better understanding together than is possible when siloed.

As a clear example of both the potential and danger here, in the swirling mass of new vendors and new promises, sound psychometric practices remain a meaningful bulwark against poor quality assessment, technology enhanced or not, to help us evaluate promises made regarding quality. At the same time, it is likely that our psychometric practices will need to be amended and expanded to describe new kinds of data. We should not assume that the methods of the past century are sufficient as they stand, and we should not dismiss new approaches without careful consideration. Navigating this balance successfully, through meaningful interdisciplinary scholarship, will emerge as the primary concern of high quality assessment researchers for the coming generation.

Third, we must continue to broaden the domain of criteria we consider when evaluating assessment technologies. In addition to applicant reactions and predictive validity, we need to also think about security and privacy, whether for ethical, legal (as in the case of GDPR or HIPAA), or practical reasons. Data that are anonymized can still be used to identify individuals when combined and triangulated with other data points, and case studies show us that publicizing this data can be harmful to individuals (Kosinski, Stillwell, & Graepel, 2013). Other fields have a head start in considering these issues. As assessment researchers, we should think about how to conduct and promote research that makes the best use of these insights from other fields. For instance: How do we collect data responsibly? How do we protect individuals and organizations from harm? How do we address issues of consent when data collected for one purpose are used for other purposes? How does the use of advanced technologies alter the fundamental relationship between employer and employee?

In sum, we hope that the papers in this special issue spark new lines of research and serve as references for scientists and practitioners using advanced technologies in their work. We hope to encourage broader conceptualizations of meaningful research in technology-enhanced talent assessment, to encourage truly integrative interdisciplinary work, and to encourage more complete mental models of "what's important" for researchers and practitioners to explore. We expect that this body of work will grow rapidly in the coming years, and as long as researchers keep an eye toward these concerns, we are optimistic about the future of the field.

REFERENCES

Dedehayir, O. & Steinert, M. (2016). The hype cycle model: A review and future directions. Technological Forecasting and Social Change, 108, 28-41.

Kosinski, M., Stillwell, D., & Graepel, T. (2013). Private traits and attributes are predictable from digital records of human behavior. Proceedings of the National Academy of Sciences, 110(15), 5802-5805.

Landers, R., & Behrend, T. (2017). When are models of technology in psychology most useful? Industrial and Organizational Psychology, 10(4), 668-675. doi:10.1017/iop.2017.74