Beyond Empirical Equivalence

Carter Gibson, Daly Vaughn, and Mike Hudy
Shaker

We wonder whether theory alone can solve problems and answer questions faced by practitioners working on the front lines of assessment innovation. Stated another way, to what degree can current theories influence the application of our work to new technology when it comes available? We are speaking as practitioners working in selection, the area in which technology has been studied most commonly in industrial-organizational (I-O) psychology (e.g., King, Ryan, Kantrowitz, Grelle, & Dainis, 2015). More specifically, we focus on the impact of mobile technology on our selection systems. We are excited for the focal article (Morelli, Potosky, Arthur, & Tippins, 2017) on theory development relative to technological advancement because much of the work we do in this area has not been discussed significantly in the literature. Our goal in this commentary is to review what we have learned about the implications of technology from our experience building and validating innovative prehire assessments.

Theory Meets Practice

At the start of the millennium, it was a selling point if a selection vendor sold assessments that could be administered online. These days it is an expectation. Although many I-O psychologists had concerns at the time, the market moved forward, and expectations for assessments now to be mobile compatible are increasingly urgent. Although many of the questions raised by the authors in the focal article are academically interesting, by the time new research or theory development hits press, it is often too late for the early applications of the new technology. Instead, we believe that peer-reviewed theory development often serves the important role of supporting decisions that have already been made in practice. Theory can also support revisions, additions, or modifications to our existing design and scoring approaches. It is our experience that the rapid evolution of technology in our space means the research studying new technology can rarely serve the role of leading advancements. In addition, many of the factors examined by the three conceptual frameworks are used to understand and explain issues that are, and always have been, beyond our control. We agree that new technology

Carter Gibson, Shaker; Daly Vaughn, Shaker; Mike Hudy, Shaker.
Correspondence concerning this article should be addressed to Carter Gibson, Shaker, 3201 Enterprise Parkway, Suite 360, Cleveland, OH 44122. E-mail: carter.gibson@shakercg.com

https://doi.org/10.1017/iop.2017.75 Published online by Cambridge University Press

introduces potentially new sources of error, but as these are rarely within our control, theories speaking to them may have limited practical value.

One of the biggest modern concerns in our industry is increasing accessibility. Each year more and more people are taking our assessments on mobile devices. We also know that several protected classes are more likely to be using mobile devices as their primary or only source of connectivity when completing an assessment (Aon Hewitt, 2016). There may come a day when not offering a mobile-optimized assessment will be considered a discriminatory act itself. We recognize that new technology brings new threats to the validity and equivalency of our tests, but we must balance the purity of our assessments with the accessibility needs of our applicants. We went down this road when the field was first confronted with the concept of unproctored Internet testing (UIT), a concept that is an expectation for high-volume prehire assessment today. It is valuable to consider the implications of technology on our field, but as practitioners looking to lead the path forward, we cannot await the development and dissemination of new theory before taking practical action toward adopting new modalities for test delivery.

To be clear, our intent is not to discredit the role of theory development in informing good practice. In fact, theoretical considerations drive many of the basic design elements of our assessments. But we believe the most pragmatic way to remain relevant and on the cutting edge of assessment development is to take a more holistic, multifaceted approach, which allows for faster product development and implementation. We believe that an evolving practical guide to best practices would be tremendously helpful. Such a guide could draw upon ideas from theory and published research but also be substantially influenced by other sources of information, such as work carried out in other disciplines with less theoretical orientation (e.g., basic user design principles), our own collective past practitioner experiences (including findings from internally conducted studies), our collective expert judgment, and our willingness to be open to piloting new approaches and technology. This could take the form of Society for Industrial and Organizational Psychology (SIOP)-sponsored articles by leading practitioners or new article submission formats to our journals. There would likely be value in brief case studies that cover a unique applied challenge and the thought process and steps taken to address it. Such a tool could even be used in a classroom environment where students could talk through the situation and get practice thinking scientifically about real-world problems.

Empirical Versus Experience Equivalence

Many of the studies reviewed by Morelli et al. (2017) are concerned with empirical equivalence. The authors rightly critique this approach as Sisyphean. In our view, creation of these somewhat arbitrary or convenient device-type

https://doi.org/10.1017/iop.2017.75 Published online by Cambridge University Press

classifications highlights and exacerbates an existing challenge that has been evident from the beginning of UIT. Although understanding the underlying factors deserves continued study, a practical approach is to also highlight the design considerations, instructional elements, and content validity of the assessment. This approach helps ensure that our tests provide a fair opportunity for candidates and minimize or mitigate as much of the variance not attributable to job relevant differences between candidates. In the real world, we are pushed to mobile delivery to meet the applicants where they are. But even before the increased adoption of smartphones, not all applicants were completing an identical assessment on equal footing. Two people taking an online assessment built for a PC could have two very different testing environments. One might be somewhere that makes concentration difficult or interrupts the assessment, whereas another candidate may minimize or eliminate all potential distractions before starting. We can tell an applicant to block off time to focus but, practically speaking, there is nothing we can do to guarantee that the testing environments are controlled and equal. Mobile testing has brought these fears back to our minds, prompting us to imagine scenarios such as an applicant taking our assessment on a crowded bus while the Internet connection goes in and out as the bus travels.

Although we agree with limitations raised by the authors regarding past research, we propose moving forward using a more pragmatic approach reliant upon guidelines and design principles from a number of sources, including psychological theory. Past work on empirical equivalence has argued that some constructs (e.g., personality) and methods (e.g., situational judgement) demonstrate empirical equivalence but others do not (e.g., cognitive ability; Arthur, Doverspike, Munoz, Taylor, & Carr, 2014). We argue that this may be an issue more strongly tied to content design and is not specific to the constructs assessed in a vacuum. Furthermore, many of these studies have been done on live applicant samples and often fail to recognize significant selection effects when it comes to device choice. For example, we see differences across applicants using mobile phones, tablets, and PCs on objective biodata items related to level of education, career stability, and beliefs regarding tardiness for work. It is unlikely that increased the permissibility of devices or reductions in screen size would impact stated level of education. These differences suggest that the only true way (in any practical sense) to even look at empirical equivalence is to randomly assign devices during a concurrent validation study.

Speaking of practical concerns, the issue of empirical device equivalence is not a question many of our clients are able or willing to answer. We have performed large-scale studies of device equivalence on concurrent validations with random assignment to devices, and the logistical hurdles are

https://doi.org/10.1017/iop.2017.75 Published online by Cambridge University Press

immense. You need a large incumbent sample. You also need the resources to bring the devices to the applicants and the internal influence to support such an initiative. The truth is that this is a very difficult sell to most clients for what they view as an academic issue.

Given these significant practical limitations and theoretical concerns for measuring empirical equivalence, our answer to this problem is to design assessments to the lowest common denominator and scale up—the exact opposite of what many assessment providers have done over the years (Morelli, Mahan, & Illingworth, 2014). We agree with the authors that we should not aggregate into simple categories (e.g., PC, tablet) because these categories are more transient than they first appear. A large mobile phone with a modern processor may provide a better experience than a slower tablet with a lower resolution built several years ago. Also, some tablet users have a keyboard, whereas others may not. Although most people do not draw a distinction between a large monitor on a desktop and a small monitor on an ultrabook (both being categorized as a PC), many people would differentiate between a large phone and a small tablet. We prefer to think of devices along a continuum of capability, where we imagine the lowest common denominator for a particular solution in terms of screen size, processing power, resolution, and so forth, and design our assessments to that device (e.g., specifications consistent with a mobile phone released several years ago).

For example, rather than take a large cognitive measure and shrink it to fit on a phone or force an applicant to scroll to see the entire page, we believe measures should be designed to the limitations of the phone and then the assessment presented the same way on the most capable device. By optimizing to the lowest common denominator and scaling to other device types, you effectively ensure what we would call *experience equivalence*. Our tests used for selection or other high-stakes purposes should be device agnostic to the extent that is practically feasible. We design each assessment so that any observed differences in the data best capture and reflect true differences in standing on the construct of interest. In practice, this means shorter situational judgment items (to fit all text onto one screen), simpler cognitive exercises, and some constructs that simply cannot be measured at all (e.g., typing speed or typing accuracy) if your lowest common denominator lacks a keyboard.

Conclusion

Assessments have moved to mobile, and that's not going away. Building theory regarding the impact that new technology has on our measurement may be academically interesting, but as practitioners, we often are forced to determine the best path forward before new theory is readily available, and this often leads to similar conclusions provided by new theory creation. It is

https://doi.org/10.1017/iop.2017.75 Published online by Cambridge University Press

important that we continue to effectively balance the practical needs of market demands and technology evolution with the theory-driven orientation of our field. In this manner, published research and theory support continued evolution and improvement but rarely can lead the way. Although we may take a pessimistic view of the value added by theoretical advancements to early-stage innovation, we greatly appreciate efforts to move forward thinking on this important topic. Many firms in the pre-hire space are coming from fields other than I-O psychology, and we believe that our theories, research, and desire to more deeply understand the data position our field to create quality assessments that help organizations better leverage their human capital.

References

Aon Hewitt, (2016). Selection testing: A look at trends in mobile device usage. Retrieved from http://www.aon.com/unitedkingdom/attachments/trp/Mobile_Testing_Whitepaper_2016.pdf

Arthur, W., Jr., Doverspike, D., Munoz, G. J., Taylor, J. E., & Carr, A. E. (2014). The use of mobile devices in high-stakes remotely delivered assessments and testing. *International Journal of Selection and Assessment, 22*(2), 113–223.

King, D. D., Ryan, A. M., Kantrowitz, T., Grelle, D., & Dainis, A. (2015). Mobile Internet testing: An analysis of equivalence, individual differences, and reactions. *International Journal of Selection and Assessment, 23*(4), 382–394.

Morelli, N. A., Mahan, R. P., & Illingworth, A. J. (2014). Establishing the measurement equivalence of online selection assessments delivered on mobile verses nonmobile devices. *International Journal of Selection and Assessment, 22*(2), 124–138.

Morelli, N., Potosky, D., Arthur, W., Jr., & Tippins, N. (2017). A call for conceptual models of technology in I-O psychology: An example from technology-based talent assessment. *Industrial and Organizational Psychology: Perspectives on Science and Practice, 10*(4), 634–653.

Integrating Technology Into Models of Response Behavior

Dev K. Dalal and Jason G. Randall
University at Albany, State University of New York

Morelli, Potosky, Arthur, and Tippins (2017) are correct in calling for more conceptual models explicitly linking technology to industrial-organizational (I-O) psychology. As these authors note, in the absence of models and theories of technology to guide the research and practice of I-O psychology, the field runs the risk of chasing the impacts of specific technological

Dev K. Dalal, University at Albany, State University of New York; Jason G. Randall, University at Albany, State University of New York.

Correspondence concerning this article should be addressed to Dev K. Dalal, University at Albany, State University of New York, 1400 E. Washington Ave., SS-399, Albany, NY 12222. E-mail: ddalal@albany.edu

https://doi.org/10.1017/iop.2017.75 Published online by Cambridge University Press