Human Factors and Simulation, Vol. 30, 2022, 18–26
https://doi.org/10.54941/ahfe1001487

Toward a Consequential Validity Perspective for Selecting Participant Groups in Testing and Evaluation Studies for Complex Systems

Nathan Lau[1] and Ron L. Boring[2]

[1]Grado Department of Industrial Engineering, Virginia Tech Blacksburg, VA 24061, USA
[2]Human Factors and Reliability Department, Idaho National Laboratory Idaho Falls, ID 83415, USA

ABSTRACT

Testing and evaluation of technology design for complex systems cannot readily attain conclusive results. This is because skilled professionals are often not available for testing while non-professionals may not be capable of operating the actual systems or high-fidelity simulators. Thus, practitioners and applied scientists can be challenged with decisions on selecting participant groups, which can severely constrain choices in the experimental tasks. This article presents the perspective of consequential validity, highlighting that general validity or rules to participant selection probably do not exist. Most importantly, the validity of a testing method or an empirical finding critically rests on the decisions of interest that must take into account nuances or idiosyncrasies of specific situations and desired outcomes. This perspective stands in contrast to how the literature predominantly portrays validity of testing methods or empirical findings as universal rather than focusing on outcomes within the confines of the study methods. The perspective of consequential validity calls for studies on how classical metrics of reliability and validity could manifest in consequence of specific decisions informed by empirical testing.

Keywords: Testing and evaluation, Consequential validity, Generalization, Application

INTRODUCTION

Empirical research is central to the science and engineering of human factors because human response has been anything but predictable in the real world. There is no shortage of evidence, stories and even sayings on how scientists, designers and engineers might have discounted the diversity of human beings and behaviors. A number of accidents with semi-automated driving vehicles involved drivers engaging in dangerous behaviors, such as watching a movie (National Transportation Safety Board, 2017), that the designers argued were either unanticipated or unpreventable.

We sometimes hear that psychology is a "white male science" with predominately white males as both researchers and participants. These research results have been haphazardly generalized to other population groups and sometimes resulted in serious inequity (e.g., Perez, 2019; Roberts et al.,

2020). The issues of sampling and generalization have remained common with unfortunate practical ramifications. For example, natural language processing in many services fails to recognize speech equally across races, potentially limiting access and productivity of various population segments (e.g., Koenecke et al., 2020). Given the increasing reliance on data models or machine learning in decision making, the equity ramification in misapplying empirical data cannot be overstated (e.g., Lau et al., 2018; Lau et al., 2020; O'Neil, 2016). The promising light has been the increasing awareness of diversity and inclusion in society, causing scientists and practitioners to ensure validity of their work in science and engineering.

For safety-critical and complex industrial systems, empirical research has been a constant challenge, albeit from a slightly different perspective. Domain professionals are rarely available to participate in research studies, leaving convenience sampling of human participants as the predominant strategy in many experimental studies. That is, skilled professionals are not readily available for testing, while non-professionals are not readily capable to operate actual systems or high-fidelity simulators. Arguably, empirical findings from a larger and more generic population group might be confounded by fewer meaningful idiosyncrasies that do not apply to the specific population of interest.

Given these constraints, testing and evaluation for complex systems often rely on a combination of studies with participants of varying skill levels and tasks of varying complexities/representativeness to maximize confidence in the design and qualification decisions based on those empirical results. However, generalization and application of empirical findings still rely extensively on expertise and past practice rather than science. Given the continual discoveries of inequitable policies and designs, a deeper, and ideally a more systematic, examination into participant sampling is warranted to support human factors scientists and practitioners in making methodological decisions on test and evaluation studies as well as drawing conclusions from the results under recruitment and other study constraints.

CONSEQUENTIAL VALIDITY

The generalization or application of empirical results can be examined from the perspective of the long-standing research on reliability and validity in psychological testing. In particular, reliability and validity research in psychological testing has undergone three phases (Murphy, 2009). The *first phase* focused on test sensitivity and reliability to ensure consistent outcomes across time, items, and/or people. Reliability ensures that the measurements are not merely noise without any systematic variance in differentiating individual differences or other factors of interest. Reliability is thus a precursor to validity, as tests must consistently differentiate on some properties or dimensions to inform any decisions.

The *second phase* is the pursuit for construct or criterion validity to establish that the results of a psychological test can predict some specific performance or outcomes in the real world. For example, significant effort has been dedicated to validating that NASA TLX is indeed measuring workload,

which has been postulated and tested to correlate with performance and other constructs (e.g., Hart & Staveland, 1988; Rubio et al., 2004). The pursuit of construct validity of various measures or metrics often dominates the discussion in human factors research on testing and evaluation. For example, the dominant discussion on measurements of situation awareness (SA) is to establish various types of reliability and validity to validate the construct (c.f., Durso et al., 2006; Endsley, 2000; Lau et al., 2016a; Lau et al., 2014; Pew, 2000; Salmon et al., 2009). Attaining construct validity is an important scientific endeavor that provides clarity on what is being measured and how the measurements can be generalized.

The *third phase* is the pursuit of consequential validity in recognition that why and how psychological testing methods are used vary across situations and play a significant role in validity in practice. That is, the validity of psychological testing in practice is associated with the specific decisions informed by the test results and the consequences given the decisions. Hence, construct validity of a test or measure established in scientific research is no guarantee of desirable consequences, as every specific decision has nuances that inevitably deviate from the circumstances of the validation studies in scientific research. Let's consider a perfectly reliable test that erroneously claims to measure empathy but in fact only measures general intelligence. Though lacking construct validity, this test is invaluable and possesses great consequential validity for screening hires whose job is to answer IQ questions. This rather absurd example illustrates how seemingly construct-valid testing methods may not necessarily be deemed valid in practice, and vice versa. In fact, human beings often operate on consequential validity, such as the use of heuristics, rather than formal, construct validity. Consequential validity is thus the focus in practice or applications when there are very specific situations and desired outcomes.

The three research phases of psychological testing highlight that universal validity for testing and evaluation methods does not truly exist. At best, only the findings on lacking reliability and validity may be generalizable to all cases. The nuance turns out significant between science seeking the truth and practice seeking desirable consequences. In particular, science strives to partition out idiosyncrasies across situations for generalization, whereas practice strives to account for idiosyncrasies of a specific situation for desired outcomes. Thus, the validity of test and evaluation study results does depend on the decision to be made and consequences to be desired.

Examples of Participant Sampling on Consequential Validity

The concept of consequential validity is associated with all aspects of testing and evaluation methods, including participant sampling. Applying any particular empirical result for a decision in practice must take into account the methodological details of the study. We present three examples of how decisions and consequences can influence the applications of our own research results in relation to participant sampling.

The first example compares two decisions on adopting Ecological Interface Design (EID; Vicente & Rasmussen, 1992), a two-phase framework

for designing user interfaces to support operators in managing unanticipated events in complex systems. Empirical research on EID is primarily based on a small-scale, representative thermohydraulic process simulator called Dual Reservoir System Simulation II (DURESS; Vicente, 1999) and college students (Vicente, 2002) to demonstrate the benefits of ecological interfaces relative to conventional ones. Subsequent research includes a number of highly representative studies employing high-fidelity simulators and professional operators (see e.g., Bennett & Flach, 2011; Jamieson, 2007).

Let's consider two decisions with potentially different consequences based on these research results: (1) should EID be recommended for the new construction of a nuclear power plant (NPP), and (2) should EID be recommended for replacing user interfaces during a digital upgrade of the control system for an operating NPP? Although these two questions sound similar, the validity of applying EID empirical test results to support the first decision appears much higher than the second one. The consequence of ecological interfaces being effectively integrated into operations and thus promoting safety of a new plant appears likely, because a new build generally entails lengthy commissioning processes and equipment (including new hires and training programs) to resolve or accommodate idiosyncrasies accompanied with new designs. In contrast, replacing user interfaces of an operating plant with ecological interfaces presents a difficult decision because the empirical results do not and cannot address the idiosyncrasies being accommodated by the unique operating processes, staff and equipment. For example, it is very difficult to assess whether control room operators of an existing plant would have their performance degraded for an extended time due to negative transfer of learning from the legacy user interfaces. Even though the two decisions concern the development of user interfaces for NPP control rooms, the consequences of those decisions are sufficiently different. In a representative study involving a high-fidelity simulator of an operating NPP and professional operators licensed for that NPP, ecological interfaces were indeed found to support situation awareness and task performance for beyond-design basis events, providing the evidence for adopting EID even in existing NPPs (Burns et al., 2008; Lau, Jamieson, et al., 2008; Lau, Veland, et al., 2008). The empirical findings specific to the NPP seem necessary to advocate for introducing ecological interfaces into the operating plant during a control system upgrade.

The second example compares two decisions on the use of eye-gaze measurements in healthcare. The literature contains a large number of empirical studies illustrating how various eye-gaze measurements are significantly different between experts and novices of medical professionals (Ashraf et al., 2018; Gegenfurtner et al., 2011; Tien et al., 2014). Along with increasing ease of use and decreasing costs, eye tracking is slowly being adopted in various healthcare applications for assessment.

Let's consider two decisions with potentially different consequences based on these research results: (1) should fixations on areas of interest (AOIs) be added as a criterion to decide on whether a trainee pass a skill test in the Fundamental of Laparoscopic Surgery (FLS) curriculum, and (2) should fixations on AOIs be used as an early indicator of trainee potential to decide on

whether to remove someone from the training program? The consequence of failing someone in a skill test is clearly different from removing someone from training (prior to formal testing). The validity for introducing a gaze-based criterion in the first decision appears strong, especially when attending doctors, residents, and medical students have participated in the research to suggest sensitivity, reliability and validity of the gaze metric at differentiating two skill levels. Adopting the same metric as a gauge of deviation and thus potentially passing a FLS skill test would also seem reasonable extension of the gaze metric. However, our recent research examining skill progression of medical students on a FLS practice task indicates that trainees seem to have faster completion time but exhibit worse gaze metrics after some training before attaining the best completion time and gaze metrics (Deng et al., 2021). We postulated this non-linear relationship is probably because trainees might commit "errors" in order to shift their gaze behaviors from feedback to feedforward control as they gain some familiarity and skills with the task. In this case, the problematic issue for the second decision is not the representativeness of the medical task and participant sampling but rather the specifics on how proficiency is acquired for a particular skill.

The final example considers the need to gather empirical data to support human reliability analysis (HRA). HRA has developed a wide range of methods for estimating human error probabilities (HEPs), but the basis of these methods often stems from expert estimation. Thus, there is a strong desire and need to validate these HEPs, especially as they are now used to make design decisions for new interface technologies in advanced reactors. There are three main ways to solve this problem of a shortage of data: (1) conduct studies using full-scope simulators and licensed operators, (2) gather data from simulators at plants used for training, or (3) develop surrogate technologies like simplified simulators that can be used with student operators. Each has tradeoffs: (1) the full-scope simulator studies cannot economically or feasibly be conducted with sufficient sample sizes to account for low probability events, (2) training simulators may not afford sufficient operational control to provide useful and complete data, and (3) simplified simulators and students may not sufficiently generalize back to the target population. All three approaches prove reliable measures, and the first two approaches have good construct validity. A good deal of research has gone into establishing the construct validity of the simplified simulator (e.g., Park et al., 2022), ensuring that the results apply to the target population and domain. Further research has established the limits of simplified simulators in terms of the suitability of different degrees of simulator fidelity and types of scenarios (Boring et al., 2019) to begin addressing consequential validity. These considerations have helped pave the way for using the cheaper simplified simulators and more readily available student operators for validating HRA results.

Advancing the Consequential Validity Perspective

The three phases of research on psychological testing and our research experience suggest that the concept of validity in science is not necessarily identical

in practice. While scientific studies on reliability and construct validity target generalization and produce invaluable knowledge relevant for practice, applied scientists, designers, engineers and regulators mostly operate in the application space where idiosyncrasies can, or even should, play a crucial role in both decisions and consequences. For this reason, the perspective of consequential validity is pertinent in making detailed methodological decisions in application driven empirical studies and translating empirical results for decision-making. That is, any testing methods that produce reliable results can be valid for some decisions, and the decisions and corresponding consequences dictate the validity of a testing method or empirical evaluation. Validity, at least consequential validity, cannot be established prior to specifying the decision.

Presently, there is a paucity of research on consequential validity in human factors to investigate how specific decisions based on an empirical result contribute to consequences of those decisions. Thus, a framework from the consequential validity perspective is pre-mature for selecting participants or making other methodological decisions. Nevertheless, our rumination on the concept of consequential validity in connection to our research experience highlights considerations and research needs on participant sampling, experimental task design, and measurement selection catering to applications. Given a specific decision to be made, the primary consideration in the testing method is the *consequences for whom,* so that the degree of idiosyncrasies to be addressed can be explicit. High specificity of the target population (e.g., operators of a single vs all NPPs) would likely demand highly representative testing to account for the idiosyncrasies contributing to the consequences of the decision.

Second, the design of the experimental tasks, which are defined by the testbeds and scenarios, must match the participant characteristics driven by the first consideration. Matching experimental tasks to participant characteristics is nothing new; however, research is lacking on equating, especially quantitively, one problem space to another. What is the equivalence in terms size and complexity of the problems space for college students to the operational space of a power plant for professional operators (Boring et al., 2019)? The size of the problem space is extremely important in accounting for the diversity of human behaviors that scientific research often unknowingly eliminates for the sake of high confidence in experimental hypotheses. Research has started looking into the implications of matching between participant sampling and experimental tasks by comparing empirical findings involving high fidelity simulator with professionals to those involving medium fidelity simulators with non-professionals (Park et al., 2022). Such research is invaluable for quantifying how the methodological differences should impact the application of empirical results.

Finally, selection of measures represents a serious concern with great uncertainty. Many human performance measures are being used across a wide range of study representativeness and domain applications that seemingly indicate high reliability and validity. However, many measures are highly customized to individual studies (see discussion on queried based SA measures, Lau et al., 2016b; Lau et al., 2012; Lau et al., 2013). On one hand,

customizing measures for specific study methods affirms the concept of consequential validity. On the other, the degree to which the psychometrics of a measure determined mainly by generalization studies would transfer to specific application testing and evaluation are often unknown. Methodological and empirical research needs to guide selection of measures across situations and decisions.

CONCLUSION

The concept of consequential validity in psychological testing research highlights that general validity or rules to participant selection probably do not exist, and more importantly, that the validity of a testing method or an empirical finding critically rests on the decisions of interest. Though this may be of no surprise to seasoned practitioners and researchers, the literature focuses primarily on generalization, portraying validity of testing methods or empirical findings as universal rather than outcomes within confines of the study methods. Given nuances between every application, research should also turn to developing systematic processes or procedures to apply testing methods and results for specific decisions. There need to be more investigations on how classical metrics of reliability and validity would manifest in the consequences of different decisions. The perspective of consequential validity sheds light on bridging generalization-focused studies in science to application-focused decisions in practice.

REFERENCES

Ashraf, H., Sodergren, M. H., Merali, N., Mylonas, G., Singh, H., & Darzi, A. (2018). Eye-tracking technology in medical education: A systematic review. *Medical Teacher*, *40*(1), 62–69. https://doi.org/10.1080/0142159X.2017.1391373

Bennett, K. B., & Flach, J. M. (2011). *Display and Interface Design: Subtle Science, Exact Art*. CRC Press.

Boring, R. L., Ulrich, T. A., Lew, R., & Rasmussen, M. (2019). Parts and Wholes: Scenarios and Simulators for Human Performance Studies. *Advances in Human Error, Reliability, Resilience, and Performance*, 116–127. https://doi.org/https://doi.org/10.1007/978-3-319-94391-6_12

Burns, C. M., Skraaning Jr., G., Jamieson, G. A., Lau, N., Kwok, J., Welch, R., & Andresen, G. (2008). Evaluation of ecological interface design for nuclear process control: situation awareness effects. *Human Factors*, *50*, 663–679.

Deng, S., Kulkarni, C., Wang, T., Hartman-Kenzler, J., Barnes, L. E., Henrickson Parker, S., Safford, S. D., Rajamohan, S., & Lau, N. K. (2021). Differentiating Laparoscopic Skills of Trainees with Computer Vision Based Metrics. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *65*(1), 304–308. https://doi.org/10.1177/1071181321651263

Durso, F. T., Bleckley, M. K., & Dattel, A. R. (2006). Does Situation Awareness Add to the Validity of Cognitive Tests? *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *48*(4), 721–733. https://doi.org/10.1518/001872006779166316

Endsley, M. R. (2000). Direct measurement of situation awareness: Validity and use of SAGAT. In M. R. Endsley & D. J. Garland (Eds.), *Situation awareness: analysis and measurement* (pp. 147–174). Lawrence Erlbaum Associates.

Gegenfurtner, A., Lehtinen, E., & Säljö, R. (2011). Expertise Differences in the Comprehension of Visualizations: a Meta-Analysis of Eye-Tracking Research in Professional Domains. *Educational Psychology Review*, *23*(4), 523–552. https://doi.org/10.1007/s10648-011-9174-7

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theorectical research. In P. A. Hancock & N. Meshkati (Eds.), *Human Mental Workload* (pp. 139–183). Elsevier Science Publisher.

Jamieson, G. A. (2007). Ecological interface design for petrochemical process control: An empirical assessment. *IEEE Trans. Systems, Man and Cybernetics*, *37*(6), 906–920.

Koenecke, A., Nam, A., Lake, E., Nudell, J., Quartey, M., Mengesha, Z., Toups, C., Rickford John, R., Jurafsky, D., & Goel, S. (2020). Racial disparities in automated speech recognition. *Proceedings of the National Academy of Sciences*, *117*(14), 7684–7689. https://doi.org/10.1073/pnas.1915768117

Lau, N., Fridman, L., Borghetti, B. J., & Lee, J. D. (2018). Machine Learning and Human Factors: Status, Applications, and Future Directions. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *62*(1), 135–138. https://doi.org/10.1177/1541931218621031

Lau, N., Hildebrandt, M., & Jeon, M. (2020). Ergonomics in AI: Designing and Interacting With Machine Learning and AI. *Ergonomics in Design*, *28*(3), 3–3. https://doi.org/10.1177/1064804620915238

Lau, N., Jamieson, G. A., & Skraaning, G. (2016a). Empirical evaluation of the Process Overview Measure for assessing situation awareness in process plants. *Ergonomics*, *59*(3), 393–408. https://doi.org/10.1080/00140139.2015.1080310

Lau, N., Jamieson, G. A., & Skraaning, G. (2016b). Situation awareness acquired from monitoring process plants – the Process Overview concept and measure. *Ergonomics*, *59*(7), 976–988. https://doi.org/10.1080/00140139.2015.1100329

Lau, N., Jamieson, G. A., & Skraaning Jr, G. (2012). Situation Awareness in Process Control: A Fresh Look. *Proceedings of the 8th American Nuclear Society International Topical Meeting on Nuclear Plant Instrumentation & Control and Human-Machine Interface Technologies (NPIC & HMIT)*, 1511–1523.

Lau, N., Jamieson, G. A., & Skraaning Jr, G. (2013). Distinguishing three accounts of Situation Awareness based on their domains of origin. *Proc. of the 52nd Annual Meeting of the Human Factors and Ergonomics Society*, 220–224.

Lau, N., Jamieson, G. A., & Skraaning Jr., G. (2014). Inter-rater reliability of query/probe-based techniques for measuring situation awareness. *Ergonomics*, *57*(7), 959–972.

Lau, N., Jamieson, G. A., Skraaning Jr., G., & Burns, C. M. (2008). Ecological interface design in the nuclear domain: An empirical evaluation of ecological displays for the secondary subsystems of a boiling water reactor plant simulator. *IEEE Trans. Nuclear Science*, *55*(6), 3597–3610.

Lau, N., Veland, ∅., Kwok, J., Jamieson, G. A., Burns, C. M., Braseth, A. O., & Welch, R. (2008). Ecological interface design in the nuclear domain: An application to the secondary subsystems of a boiling water reactor plant simulator. *IEEE Trans. Nuclear Science*, *55*(6), 3579–3596.

Murphy, K. R. (2009). Validity, Validation and Values [Article]. *Academy of Management Annals*, *3*(1), 421–461. https://doi.org/10.1080/19416520903047525

National Transportation Safety Board. (2017). *Highway accident report: Collision between a car operating with automated vehicle control systems and a tractor-semitrailer truck near Williston, Florida, May 7, 2016*.

(NTSB/HAR-17/02). https://www.ntsb.gov/investigations/accidentreports/reports/har1702.pdf

O'Neil, C. (2016). *Weapons of Math Destruction: How big data increases inequality and threatens democracy* [Book]. Broadway Books. http://login.ezproxy.lib.vt.edu/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=nlebk&AN=1109940&site=eds-live&scope=site

Park, J., Boring, R. L., Ulrich, T. A., Lew, R., Lee, S., Park, B., & Kim, J. (2022). A framework to collect human reliability analysis data for nuclear power plants using a simplified simulator and student operators. *Reliability Engineering & System Safety*, *221*, 108326. https://doi.org/https://doi.org/10.1016/j.ress.2022.108326

Perez, C. C. (2019). *Invisible women: Data bias in a world designed for men*. Abrams Press.

Pew, R. W. (2000). The state of situation awareness measurement: Heading toward the next century. In M. R. Endsley & D. J. Garland (Eds.), *Situation Awareness Analysis and Measurement* (pp. 33–47). Lawrence Erlbaum Associates.

Roberts, S. O., Bareket-Shavit, C., Dollins, F. A., Goldie, P. D., & Mortenson, E. (2020). Racial Inequality in Psychological Research: Trends of the Past and Recommendations for the Future. *Perspectives on Psychological Science*, *15*(6), 1295–1309. https://doi.org/10.1177/1745691620927709

Rubio, S., Diaz, E., Martin, J., & Puente, J. M. (2004). Evaluation of subjective mental workload: A comparison of SWAT, NASA-TLX, and workload profile methods. *Applied Psychology: an International Review*, *53*(1), 61–86. http://dx.doi.org/10.1111/j.~1464--0597.2004.00161.x

Salmon, P. M., Stanton, N. A., Walker, G. H., Jenkins, D., Ladva, D., Rafferty, L., & Young, M. (2009). Measuring Situation Awareness in complex systems: Comparison of measures study. *International Journal of Industrial Ergonomics*, *39*(3), 490–500. https://doi.org/10.1016/j.ergon.2008.10.010

Tien, T., Pucher, P. H., Sodergren, M. H., Sriskandarajah, K., Yang, G.-Z., & Darzi, A. (2014). Eye tracking for skills assessment and training: a systematic review. *Journal of Surgical Research*, *191*(1), 169–178. https://doi.org/http://dx.doi.org/10.1016/j.jss.2014.04.032

Vicente, K. J. (1999). *Cognitive work analysis: Toward safe, productive, and healthy computer-based work*. Lawrence Erlbaum Associates.

Vicente, K. J. (2002). Ecological interface design: Progress and challenges. *Human Factors*, *44*(1), 62–78.

Vicente, K. J., & Rasmussen, J. (1992). Ecological interface design: theoretical foundations. *IEEE Trans. Systems, Man and Cybernetics*, *22*(4), 589–606. https://doi.org/10.1109/21.156574