Behavioral and Brain Sciences

Measurement practices exacerbate the generalizability crisis: Novel digital measures can help

--Manuscript Draft--

Manuscript Number:	
Full Title:	Measurement practices exacerbate the generalizability crisis: Novel digital measures can help
Short Title:	Measurement practices exacerbate the generalizability crisis
Article Type:	Open Peer Commentary
Corresponding Author:	Brittany Davidson, PhD University of Bath UNITED KINGDOM
Corresponding Author Secondary Information:	
Corresponding Author's Institution:	University of Bath
Corresponding Author's Secondary Institution:	
First Author:	Brittany Davidson, PhD
First Author Secondary Information:	
Order of Authors:	Brittany Davidson, PhD
	David A Ellis, PhD
	Clemens Stachl, PhD
	Paul J Taylor, PhD
	Adam N Joinson, PhD
Order of Authors Secondary Information:	
Abstract:	Psychology’s tendency to focus on confirmatory analyses before ensuring constructs are clearly defined and accurately measured is exacerbating the generalizability crisis. Our growing use of digital behaviors as predictors has revealed the fragility of subjective measures and the latent constructs they scaffold. However, new technologies can provide opportunities to improve conceptualizations, theories, and measurement practices.

1 Target Article

Tal Yarkoni – The Generalizability Crisis

2 Word Counts

Abstract	55
Main text	994
References	863
Entire Text	2,055

3 Title

Measurement practices exacerbate the generalizability crisis: Novel digital measures can help

4 Authors

Brittany I. Davidson[1,2]

David A. Ellis[1]

Clemens Stachl[3]

Paul J. Taylor[4]

Adam N. Joinson[1]

5 Intuitions, Addresses, Telephone Numbers

[1]University of Bath, Claverton Down, Bath, BA2 7AY, UK

[2]University of Bristol, One Cathedral Square, Bristol, BS1 5DD, UK

[3]Stanford University, 450 Serra Mall, Stanford, CA 94305, USA

[4]Lancaster University, Bailrigg, Lancaster, LA1 4YW, UK

6 Email addresses & websites

BID bid23@bath.ac.uk
https://www.brittanydavidson.co.uk/

DAE dae30@bath.ac.uk
http://www.davidaellis.co.uk/

CS cstachl@stanford.edu

https://www.clemensstachl.com

PJT p.j.taylor@lancaster.ac.uk
https://pauljtaylor.com/

ANJ aj266@bath.ac.uk
http://www.joinson.com/home/Welcome.html

7 Abstract

Psychology's tendency to focus on confirmatory analyses before ensuring constructs are clearly defined and accurately measured is exacerbating the generalizability crisis. Our growing use of digital behaviors as predictors has revealed the fragility of subjective measures and the latent constructs they scaffold. However, new technologies can provide opportunities to improve conceptualizations, theories, and measurement practices.

8 Main Text

Yarkoni (2020) highlights the disconnect between psychology's descriptive theories and its inferential tests—a problem we argue is exacerbated by inadequate measurement. The primacy of measurement in psychology's history has ebbed-and-flowed, from the absolute focus on what was observable and quantifiable that defined behaviorist approaches (Hayes & Brownstein, 1986; Skinner, 1963, 1976) to the overreliance on button presses and mouse clicks that characterizes some modern research (Baumeister et al., 2007). Today, digital trace data provides new opportunities for rich measurement that captures behavioral, situational, and environmental/contextual factors simultaneously (Lazer et al., 2020; Mischel, 2004). For instance, smartphones are a powerful data source—a collection of sensors and logging routines that we carry with us for large swathes of the day—that psychologists are utilizing to predict a variety of outcomes, from social interaction, personality, mood, to general health (Davidson, 2020; Ellis, 2020; Miller, 2012; Piwek et al., 2016; Harari et al., 2019; Stachl et al., 2020).

Improved methodology alone will not result in rapid progress for the behavioral sciences (see Kaplan, 1964 and Uttal, 2001). For example, digital trace data has re-ignited problems with traditional operationalizations of latent variables. Research demonstrating associations between new and old measures often fails to articulate why a connection between a latent measure (e.g., mood disturbance) and behavioral (digital) predictor (e.g., keystroke speed) should exist in advance of an analysis (Davidson, 2020; Zulueta et al., 2018). Without specification or theory, the focus on prediction over explanation restricts generalizability further. A related challenge is the disconnect between subjective and objective (e.g., Taylor et al. 2021), where predictive studies find their survey data predicts an outcome, but objective measures do not (Eisenberg et al., 2019). Here, the problem is an overreliance on subjective methodologies to measure both latent and observable constructs. For example, the gold standard for personality measurement relies on surveys (e.g., HEXACO, OCEAN, Big 5) and remains contested (Cattell, 1958; Kagan, 2001). Similarly, other measures including estimates of everyday behavior rarely align with reality (Parry et al., 2020). While latent measurement remains core to psychological science, many constructs are developed rapidly, with little standardization, and rely on face validity alone (e.g., ‘internet addiction’, despite being sardonic in origin, has spawned 100s of technology addiction scales; Howard & Jayne, 2015). New digital sources need to avoid these issues if they are to prosper.

Illuminating the complex relationship between generalizability and measurement further—observations of behavior via digital traces will often only explain (or predict) part of a broad latent construct. At face value, predicting part of extraversion may appear straightforward from digital recordings of speech, or time spent using social apps. However, there are other sub-components of extraversion that this data will struggle to explain (e.g., feeling indifferent to social activities). Other personality factors such as openness and agreeableness remain conceptually more challenging to map onto (a single) digital behavior (Hinds & Joinson, 2019, Stachl et al., 2020). Hence, it is critically important psychology shifts away from predictive validity alone as evidence for successful operationalization and parameterization, especially from new data sources (Boyd et al., 2020). Any new digital measure has to be developed incrementally, where researchers first describe how it conceptually aligns with an existing latent construct (Glewwe & van der Gaag, 1990). Assuming that digital traces are behavioural expressions of latent variables, researchers should be able to qualitatively express links at a more general level first across contexts, then move to specifics, which would enhance generalizability.

Of course, re-focusing on actual behavior via digital traces will not be a panacea. Some digital traces may be ‘objective’, but they are rarely error-free (Sen et al., 2019). For example, a microphone-based audio classifier can detect whether ambient conversations are taking place around an individual, but it may not distinguish real conversations from someone watching television. Similarly, little consideration is given to how measurement variance might be reduced or maximized for a new digital source. For example, while some assessments in psychology (e.g., cognitive tasks) do not produce reliable individual differences, others (e.g., mood) purposefully reflect variations in individual responses (Hedge et al., 2018). Hence, it is critical to find ways to share *raw* data, processing pipelines, and analysis scripts for digital trace research, as the degrees of freedom are vast, which causes large variance in conclusions made from the same data (Silberzahn et al., 2018; Towse et al., 2020). Validation procedures are likely to reflect the disparity of digital data sources, but combining small and large-scale approaches (e.g., N=1 sample, case studies) can successfully quantify errors associated with smartphone sensing-based methods (Geyer et al., 2020; Sen et al., 2019; Szot et al., 2019). Only then can related work explore how signals from multiple systems may be combined to improve data efficiency. Failure to ensure this basic research is completed will result in little progress as research agendas risk shifting in the wrong direction if the grounding principles are weak, particularly in applied settings, such as security and health, which are increasingly interested in digital traces (Davidson, 2020; Guttman & Greenbaum, 1998).

Moreover, we acknowledge that research in this space remains challenging to conduct because data derived from digital sources can be difficult to access, handle, and interpret (DeMasi et al., 2017). This challenges the way psychologists are trained and incentivized (not) to publish descriptive findings in an interdisciplinary landscape. However, we are hopeful that new methods and emerging forms of data will complement psychology's diverse measurement practices. Collectively termed the *Internet of Things*, the future potential for data linkage that could further leverage real-world research remains an exciting prospect. In the long term, taking time to understand how behavioral, situational, and environmental/contextual factors can be extracted from objective digital data will allow psychology to develop robust contextualized and comprehensive theory (Lazer et al., 2020).

Our muse are people and psychology should critically consider how it moves forward and merges old and new. Generalizability requires sound measures first, but there is still little agreement between psychologists on what is worth measuring.

9 Conflicts of Interest

This work was part-funded by the Centre for Research and Evidence on Security Threats (ESRC Award: ES/N009614/1 to PJT; ANJ; DAE), www.crestresearch.ac.uk.

10 References

Baumeister, R. F., Vohs, K. D., & Funder, D. C. (2007). Psychology as the science of self- reports and finger movements: Whatever happened to actual behavior? *Perspectives on Psychological Science, 2*(4), 396–403. https://doi.org/10.1111/j.1745-6916.2007.00051.x

Boyd, R. L., Pasca, P., & Lanning, K. (2020). The Personality Panorama: Conceptualizing Personality through Big Behavioural Data. *European Journal of Personality*, *34*(5), 599–612. https://doi.org/10.1002/per.2254

Cattell, R. B. (1958). What is 'objective' in 'objective personality tests'? *Journal of Counseling Psychology, 5*(4), 285. https://doi.org/10.1037/h0046268

Davidson, B. I. (2020). The crossroads of digital phenotyping. *General Hospital Psychiatry*. https://doi.org/10.1016/j.genhosppsych.2020.11.009

DeMasi, O., Kording, K., & Recht, B. (2017). Meaningless comparisons lead to false optimism in medical machine learning. *PLOS ONE, 12*(9), e0184604. https://doi.org/10.1371/journal.pone.0184604

Eisenberg, I. W., Bissett, P. G., Zeynep Enkavi, A., Li, J., MacKinnon, D. P., Marsch, L. A., & Poldrack, R. A. (2019). Uncovering the structure of self-regulation through data-driven ontology discovery. *Nature Communications*, *10*(1), 2319. https://doi.org/10.1038/s41467-019-10301-1

Ellis, D. A. (2020). *Smartphones within Psychological Science*. Cambridge University Press.

Geyer, K., Ellis, D. A., Shaw, H., & Davidson, B. I. (2020). *Open source smartphone app and tools for measuring, quantifying, and visualizing technology use* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/eqhfa

Glewwe, P., & van der Gaag, J. (1990). Identifying the Poor in Developing Countries: Do Different Definitions Matter? *World Development*, *18*(6), 803–814. https://doi.org/10.1016/0305-750X(90)90003-G

Guttman, R. & Greenbaum, C.W. (1998). Facet Theory: It's Development and Current Status. *European Psychologist*, 3, 13-36. https://doi.org/10.1027/1016-9040.3.1.13

Harari, G. M., Müller, S. R., Stachl, C., Wang, R., Wang, W., Bühner, M., Rentfrow, P. J., Campbell, A. T., & Gosling, S. D. (2020). Sensing sociability: Individual differences in young adults' conversation, calling, texting, and app use behaviors in daily life. Journal of Personality and Social Psychology, 119(1), 204–228. https://doi.org/10.1037/pspp0000245

Hayes, S. C., & Brownstein, A. J. (1986). Mentalism, Behavior-Behavior Relations, and a Behavior-Analytic View of the Purposes of Science. *The Behavior Analyst*, *9*(2), 175–190. https://doi.org/10.1007/BF03391944

Hedge, C., Powell, G., & Sumner, P. (2018). The reliability paradox: Why robust cognitive tasks do not produce reliable individual differences. Behavior Research Methods, 50(3), 1166–1186.

Hinds, J., & Joinson, A. (2019). Human and computer personality prediction from digital footprints. Current Directions in Psychological Science, 28(2), 204-211.

Howard, M. C., & Jayne, B. S. (2015). An analysis of more than 1,400 articles, 900 scales, and 17 years of research: the state of scales in cyberpsychology, behavior, and social networking. Cyberpsychology, Behavior, and Social Networking, 18(3), 181-187.

Kagan, J. (2001). The Need for New Constructs. *Psychological Inquiry*, *12*(2), 84–103. https://doi.org/10.1207/S15327965PLI1202_03

Kaplan, A. (1964). The Conduct of Inquiry: Methodology for. *Behavioral Science. Scranton PA.*

Lazer, D. M. J., Pentland, A., Watts, D. J., Aral, S., Athey, S., Contractor, N., Freelon, D., Gonzalez-Bailon, S., King, G., Margetts, H., Nelson, A., Salganik, M. J., Strohmaier, M., Vespignani, A., & Wagner, C. (2020). Computational social science: Obstacles and opportunities. *Science*, *369*(6507), 1060–1062. https://doi.org/10.1126/science.aaz8170

Miller, G. (2012). The Smartphone Psychology Manifesto. *Perspectives on Psychological Science*. https://doi.org/10.1177/1745691612441215

Mischel, W. (2004). Toward an Integrative Science of the Person. *Annual Review of Psychology*, *55*(1), 1–22. https://doi.org/10.1146/annurev.psych.55.042902.130709

Parry, D. A., Davidson, B. I., Sewall, C., Fisher, J. T., Mieczkowski, H., & Quintana, D. (2020). A Systematic Review and Meta-Analysis of Discrepancies Between Logged and Self- Reported Digital Media Use. *PsyArXiv*. 10.31234/osf.io/f6xvz

Piwek, L., Ellis, D. A., Andrews, S., & Joinson, A. (2016). The Rise of Consumer Health Wearables: Promises and Barriers. *PLOS Medicine*, *13*(2), e1001953. https://doi.org/10.1371/journal.pmed.1001953

Sen, I., Floeck, F., Weller, K., Weiss, B., & Wagner, C. (2019). A Total Error Framework for Digital Traces of Humans. *ArXiv:1907.08228 [Cs]*. http://arxiv.org/abs/1907.08228

Skinner, B. F. (1963). Behaviorism at Fifty. *Science*, *140*(3570), 951–958.

Skinner, B. F. (1976). *About Behaviorism*. Vintage Books.

Stachl, C., Au, Q., Schoedel, R., Gosling, S. D., Harari, G. M., Buschek, D., Völkel, S. T., Schuwerk, T., Oldemeier, M., Ullmann, T., Hussmann, H., Bischl, B., & Bühner, M. (2020). Predicting personality from patterns of behavior collected with smartphones. *Proceedings of the National Academy of Sciences*, *117*(30), 17680–17687. https://doi.org/10.1073/pnas.1920484117

Szot, T., Specht, C., Specht, M., & Dabrowski, P. S. (2019). Comparative analysis of positioning accuracy of Samsung Galaxy smartphones in stationary measurements. *PLOS ONE*, *14*(4), e0215562. https://doi.org/10.1371/journal.pone.0215562

Silberzahn, R., Uhlmann, E. L., Martin, D. P., Anselmi, P., Aust, F., Awtrey, E., ... & Carlsson, R. (2018). Many Analysts, One Data Set: Making Transparent How Variations in Analytic Choices Affect Results. Advances in Methods and Practices in Psychological Science, 1(3), 337-356.

Taylor, P. J., Banks, F., Jolley, D., Ellis, D. A., Watson, S. J., Weiher, L., Davidson, B., & Julku, J. (2021). Oral hygiene effects verbal and nonverbal displays of confidence. *Journal of Social Psychology*. doi:10.1080/00224545.2020.1784825

Towse, J. N., Ellis, D. A., & Towse, A. S. (2020). Opening Pandora's Box: Peeking inside Psychology's data sharing practices, and seven recommendations for change. *Behavior Research Methods*. https://doi.org/10.3758/s13428-020-01486-1

Uttal, W. R. (2001). *The new phrenology: The limits of localizing cognitive processes in the brain* (pp. xv, 255). The MIT Press.

Yarkoni, T. (2020). The generalizability crisis. *Behavioral and Brain Sciences*, 1–37. https://doi.org/10.1017/S0140525X20001685

Zulueta, J., Piscitello, A., Rasic, M., Easter, R., Babu, P., Langenecker, S. A., ... & Leow, A. (2018). Predicting mood disturbance severity with mobile phone keystroke metadata: A biaffect digital phenotyping study. Journal of medical Internet research, 20(7), e241.