to analyze and interpret them (Kukla, 1989). Furthermore, *pace* Yarkoni, theory cannot be read off of empirical data: theory needs to be *developed*, which requires a set of skills different from that of the experimenter. Like many other sciences, psychology needs specialized theorists whose work visibly contributes to experimental research (MacKay, 1988).

Let us close by calling attention to important similarities between the generalizability crisis and the replication crisis. Both have been with us for quite some time and both involve widespread violation of fundamental and well-known principles of scientific investigation. It is fairly obvious, for example, that the findings of a single small study may very well be false positives, especially after some p-value hacking. It is equally obvious that the inferences we draw from obtained data should be warranted. Arguably, researchers do not need Yarkoni to educate them about the need for conservative conclusions: they know the rules – they just do not follow them. This suggests that we should explore measures focused on changing the research culture (Nosek, Spies, & Motyl, 2012). But although many practices advocated by the open science movement, such as data sharing and improved quality of reporting (Hensel, 2020; Miłkowski, Hensel, & Hohol, 2018), can help to enhance both reproducibility and generalizability (the latter, by enabling high-quality re- and meta-analysis), it is also necessary to strengthen theorizing and work toward consistently incorporating theoretical results into experimental research. Without that, psychology will be a headless rider doomed to face ever new crises.

Financial support. This work was supported by the National Science Centre (Poland) research grant (MM, grant number 2014/14/E/HS1/00803).

Conflict of interest. None.

References

Alogna, V. K., Attaya, M. K., Aucoin, P., Bahník, Š., Birch, S., Birt, A. R., … Zwaan, R. A. (2014). Registered replication report: Schooler and Engstler-Schooler (1990). *Perspectives on Psychological Science*, *9*(5), 556–578. https://doi.org/10.1177/1745691614545653.

Cartwright, N. (2020). Middle-range theory: Without it what could anyone do?. *THEORIA. An International Journal for Theory, History and Foundations of Science*, *35*(3), 269–323.https://doi.org/10.1387/theoria.21479.

Hensel, W. M. (2020). Double trouble? The communication dimension of the reproducibility crisis in experimental psychology and neuroscience. *European Journal for Philosophy of Science*, *10*, 44. https://doi.org/10.1007/s13194-020-00317-6.

Irvine, E. (2021). The role of replication studies in theory building. *Perspectives on Psychological Science*, *16*(4), 844–853. https://doi.org/10.1177/1745691620970558.

Kukla, A. (1989). Nonempirical issues in psychology. *American Psychologist*, *44*(5), 785–794. https://doi.org/10.1037/0003-066X.44.5.785.

MacKay, D. G. (1988). Under what conditions can theoretical psychology survive and prosper? Integrating the rational and empirical epistemologies. *Psychological Review*, *95*(4), 559–565. https://doi.org/10.1037/0033-295X.95.4.559.

Miłkowski, M., Hensel, W. M., & Hohol, M. (2018). Replicability or reproducibility? On the replication crisis in computational neuroscience and sharing only relevant detail. *Journal of Computational Neuroscience*, *45*(3), 163–172. https://doi.org/10.1007/s10827-018-0702-z.

Muthukrishna, M., & Henrich, J. (2019). A problem in theory. *Nature Human Behaviour*, *3*, 221–229. https://doi.org/10.1038/s41562-018-0522-1.

Nosek, B. A., Spies, J. R., & Motyl, M. (2012). Scientific utopia: II. Restructuring incentives and practices to promote truth over publishability. *Perspectives on Psychological Science*, *7*(6), 615–631. https://doi.org/10.1177/1745691612459058.

Patton, M. Q. (2005). Qualitative research. In Everitt, B. S., & D. C. Howell (Eds.), *Encyclopedia of statistics in behavioral science* (Vol. 3, pp. 1633–1636). John Wiley & Sons. https://doi.org/10.1002/0470013192.bsa514.

Shadish, W. R. (1993). Critical multiplism: A research strategy and its attendant tactics. *New Directions for Program Evaluation*, *60*, 13–57. https://doi.org/10.1002/ev.1660.

Shadish, W. R., Cook, T., & Campbell, D. (2002). *Experimental and quasi-experimental design for generalized causal inference*. Houghton-Mifflin.

Weidman, A. C., Steckler, C. M., & Tracy, J. L. (2017). The jingle and jangle of emotion assessment: Imprecise measurement, casual scale usage, and conceptual fuzziness in emotion research. *Emotion*, *17*(2), 267–295. https://doi.org/10.1037/emo0000226.

Citizen science can help to alleviate the generalizability crisis

Courtney B. Hilton[a] and Samuel A. Mehr[a,b,c]

[a]Department of Psychology, Harvard University, Cambridge, MA 02138, USA; [b]Data Science Initiative, Harvard University, Cambridge, MA 02138, USA and [c]School of Psychology, Victoria University of Wellington, Kelburn Parade, Wellington 6012, New Zealand.
courtneyhilton@g.harvard.edu; sam@wjh.harvard.edu; https://themusiclab.org

doi:10.1017/S0140525X21000352, e21

Abstract

Improving generalization in psychology will require more expansive data collection to fuel more expansive statistical models, beyond the scale of traditional lab research. We argue that citizen science is uniquely positioned to scale up data collection and, that in spite of certain limitations, can help to alleviate the generalizability crisis.

Yarkoni argues that common statistical practices in psychology fail to quantitatively support the generalizations psychologists care about. This is because most analyses ignore important sources of variation and, as a result, unjustifiably generalize from narrowly sampled particulars.

Is this problem tractable? We are optimists, so we leave aside Yarkoni's suggestions to "do something else" or "embrace qualitative research," and focus instead on his key prescription: the adoption of mixed-effects modeling to estimate effects at the level of a factor (e.g., stimulus), to be interpreted as one of a population of potential measurements, licensing generalization over that factor.

Yarkoni is correct that far too few studies do this. In our field of the psychology of music, many inaccurately generalize, for example, from a single musical example to *all* music; or from a set of songs from a particular context (e.g., pop songs) to *all* songs; or from the music perception abilities of a particular subset of humans to *all* humans.

Consider the "Mozart effect": a notorious positive effect of listening to a Mozart sonata on spatial reasoning that was overgeneralized to "all Mozart" and eventually "all music." While replicable under narrow conditions, the original result was, in fact, specific to neither spatial reasoning, Mozart, nor music generally – the effect was the result of generic modifications to arousal and mood (Thompson, Schellenberg, & Husain, 2001).

Modeling random effects for stimuli and other relevant factors, however, brings with it a substantial challenge: researchers will need far more stimuli and participants, sampled more broadly and deeply, and with far more measures, than is typically practical. Psychologists already struggle to obtain sufficient statistical power for narrowly sampled, fixed-effect designs (Smaldino & McElreath, 2016).

How, then, can we alleviate the generalizability crisis? We think *citizen science* can help.

Citizen science refers to a collection of research tools and practices united by the alignment of interests between participants and the aims of the project, such that participation is intrinsically motivated (e.g., by curiosity in the topic) rather than by extrinsic factors (e.g., money or course credit). The results are studies that cheaply recruit thousands or even millions of diverse participants via the internet. Studies take many forms, ranging from "gamified" experiments that go viral online, such as our "Tone-deafness test" (current $N > 1.2$ million; https://themusiclab.org); to collective/collaborative field reporting, such as New Zealand's nationwide pigeon census (the Great Kererū count, https://www.greatkererucount.nz/).

The potential of citizen science is staggering. For example, the Moral Machine Experiment (Awad et al., 2018) collected 40 million decisions from millions of people (representing 10 languages and over 200 countries) on moral intuitions about self-driving cars. Such massive scale enabled the quantification of cross-country variability in moral intuitions, and how it was mediated by cultural and economic factors particular to each country, with profound real-world implications.

Further, when citizen science is coupled with corpus methods, generalizability across stimuli can be effectively maximized. We previously investigated high-level representations formed during music listening, by asking whether naïve listeners can infer the behavioral context of songs produced in unfamiliar foreign societies (Mehr et al., 2018, 2019). Each iteration of a viral "World Music Quiz" played a random draw of songs from the *Natural History of Song* corpus, a larger stimulus set that representatively samples music from 86 world cultures.

As such, the findings of the experiment – that listeners made accurate inferences about the songs' behavioral contexts – can be accurately generalized (a) to the populations of songs the stimulus subsets were drawn from (e.g., lullabies); (b) more weakly, to music, writ large (insofar as the subpopulations of songs represented by those categories sample from other categories); and (c) to the population of listeners from whom our participants were drawn (i.e., members of internet-connected societies). All of these factors can be explicitly modeled with random effects.

The same reasoning applies to studying subpopulations of participants (measured in terms of any characteristic) and even subsets of corpora. For example, in a study of acoustic regularities in infant-directed vocalizations across cultures, we model random effects of listener characteristics, speaker/singer (i.e., the producers of the stimuli) characteristics, and stimulus categories of interest (e.g., infant-directed vs. adult-directed speech). This is only possible with large datasets (in our case, nearly 1 million listener judgements; Hilton, Moser, et al., 2021). Other under-used analyses also become more practical with big citizen-science data, including radical randomization (Baribault et al., 2018), prediction with cross-validation (Yarkoni & Westfall, 2017), and matching methods for causal inference (Stuart, 2010).

Citizen-science methods are limited, however, by the need to factor in participants' interests and incentives; the need to avoid factors that might dissuade participation (e.g., clunky user interfaces, boring time-consuming tasks), which can require graphic design and web development talent for "gamification" (e.g., Cooper et al., 2010); the risks of recruiting a biased population subset (i.e., those with internet access; Lourenco & Tasimi, 2020); and the trade-offs between densely sampling stimuli across- versus within-participants, given the typically short duration of citizen-science experiments.

Indeed, while our efforts to recruit children at scale online via citizen science show promising results (Hilton, Crowley de-Thierry, Yan, Martin, & Mehr, 2021), rare or hard-to-study populations may be difficult to recruit en masse (cf. Lookit, a platform for online research in infants; Scott & Schulz, 2017). As Yarkoni notes, alternative approaches like multisite collaborations (e.g., ManyBabies Consortium, 2020) could be calibrated to maximize generalizability across stimuli rather than directly replicating results with the same stimuli.

All that being said, thanks to a growing ecosystem of open-source tools (e.g., de Leeuw, 2015; Hartshorne, de Leeuw, Goodman, Jennings, & O'Donnell, 2019; Peirce et al., 2019); the availability of large-scale, naturalistic corpora from industry partners (e.g., Spotify Research; Way, Garcia-Gathright, & Cramerr, 2020); and calls for collaborative, field-wide investment in citizen-science infrastructure (Sheskin et al., 2020) – addressing these limitations has never been easier.

As such, we think that citizen science can play a useful role as psychologists begin to address the generalizability crisis.

Acknowledgment. We would like to thank Max Krasnow, Mila Bertolo, Stats Atwood, Alex Holcombe, and William Ngiam for feedback on drafts of this commentary.

Financial support. C.B.H. and S.A.M. are supported by NIH DP5OD024566. S.A.M. is supported by the Harvard Data Science Initiative.

Conflict of interest. None.

References

Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Sharff, A., … Rahwan, I. (2018). The moral machine experiment. *Nature*, *563*(7729), 59–64.

Baribault, B., Donkin, C., Little, D. R., Trueblood, J. S., Oravecz, Z., van Ravenzwaaij, D., … Vandekerckhove, J. (2018). Metastudies for robust tests of theory. *Proceedings of the National Academy of Sciences*, *115*(11), 2607–2612.

Cooper, S., Khatib, F., Treuille, A., Barbero, J., Lee, J., Beenen, M., … Players, F. (2010). Predicting protein structures with a multiplayer online game. *Nature*, *466*(7307), 756–760.

de Leeuw, J. R. (2015). jsPsych: A JavaScript library for creating behavioral experiments in a Web browser. *Behavior Research Methods*, *47*(1), 1–12.

Hartshorne, J. K., de Leeuw, J., Goodman, N., Jennings, M., & O'Donnell, T. J. (2019). A thousand studies for the price of one: Accelerating psychological science with Pushkin. *Behavior Research Methods*, *51*(4), 1–22.

Hilton, C., Crowley de-Thierry, L., Yan, R., Martin, A., & Mehr, S. (2021). Children infer the behavioral contexts of unfamiliar songs. *PsyArXiv*. doi: 10.31234/osf.io/rz6qn.

Hilton, C. B., Moser, C. J., Bertolo, M., Lee-Rubin, H., Amir, D., Bainbridge, C. M., … Mehr, S. A. (2021). Acoustic regularities in infant-directed vocalizations across cultures. *bioRxiv*. doi: 10.1101/2020.04.09.032995

Lourenco, S. F., & Tasimi, A. (2020). No participant left behind: Conducting science during COVID-19. *Trends in Cognitive Sciences*, *24*(8), 583–584.

ManyBabies Consortium. (2020). Quantifying sources of variability in infancy research using the infant-directed-speech preference. *Advances in Methods and Practices in Psychological Science*, *3*, 24–52.

Mehr, S. A., Singh, M., Knox, D., Ketter, D., Pickens-Jones, D., Atwood, S., … Glowacki, L. (2019). Universality and diversity in human song. *Science*, *366*(6468), eaax0868.

Mehr, S. A., Singh, M., York, H., Glowacki, L., & Krasnow, M. M. (2018). Form and function in human song. *Current Biology*, *28*(3), 356–368.e5.

Peirce, J., Gray, J. R., Simpson, S., MacAskill, M., Höchenberger, R., Sogo, H., … Lindeløv, J. K. (2019). PsychoPy2: Experiments in behavior made easy. *Behavior Research Methods*, *51*(1), 195–203.

Scott, K., & Schulz, L. (2017). Lookit (Part 1): A new online platform for developmental research. *Open Mind*, *1*(1), 4–14.

Sheskin, M., Scott, K., Mills, C. M., Bergelson, E., Bonawitz, E., Spelke, E. S., … Schulz, L. (2020). Online developmental science to foster innovation, access, and impact. *Trends in Cognitive Sciences*, *24*(9), 675–678.

Smaldino, P. E., & McElreath, R. (2016). The natural selection of bad science. *Royal Society Open Science*, *3*(9), 160384.

Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical Science*, *25*(1), 1–21.

Thompson, W. F., Schellenberg, E. G., & Husain, G. (2001). Arousal, mood, and the Mozart effect. *Psychological Science*, *12*(3), 248–251.

Way, S. F., Garcia-Gathright, J., & Cramerr, H. (2020). Local trends in global music streaming. *Proceedings of the Fourteenth International AAAI Conference on Web and Social Media*, 10.

Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science*, *12*(6), 1100–1122.

Look to the field

Rumen Iliev[a], Douglas Medin[b] and Megan Bang[c]

[a]Toyota Research Institute, Los Altos, CA 94022, USA; [b]Department of Psychology and School of Education and Social Policy, Northwestern University, Evanston, IL 60208, USA and [c]Learning Sciences and Department of Psychology, Northwestern University, Annenberg Hall, Evanston IL 60208, USA
rumen.iliev@tri.global
medin@northwestern.edu
Megan.bang@northwestern.edu

doi:10.1017/S0140525X21000509, e22

Abstract

Yarkoni's paper makes an important contribution to psychological research by its insightful analysis of generalizability. We suggest, however, that broadening research practices to include field research and the correlated use of both converging and complementary observations gives reason for optimism.

We agree with Yarkoni's thesis that there is a "generalizability crisis" and that the mapping between verbal theoretical constructs and measures and models is the source of many difficulties. In particular, the limited variation in procedures, stimuli, contexts, and measures represents a significant challenge to generalizability. Yarkoni summarizes these concerns by suggesting that "a huge proportion of the quantitative inferences drawn in the published psychology literature are so weak as to be at best questionable and at worst utterly nonsensical."

Although Yarkoni's arguments are compelling, we don't fully agree with the somewhat gloomy picture he paints. The generalizability crisis creates something of a paradox: If generalization claims are on such shaky grounds, why is it that many phenomena are so robust that they make for reliable classroom demonstrations and/or have been shown to have substantial practical significance?

With respect to the former, examples include a number of judgment and decision biases identified and analyzed by Kahneman, Tversky, Fischhoff, Slovic, Loewenstein, Weber, and others (e.g., availability heuristic, loss aversion, framing effects, quantity insensitivity). With respect to the latter, Cialdini (2009a, 2009b) has demonstrated simple but effective manipulations that increase environmentally friendly behaviors (e.g., hotel guests reusing towels). Similarly, implementing changes default assumptions (Thaler & Sunstein, 2008) has been shown to facilitate policy goals such as increasing organ donation.

Field versus lab

We suggest that attention to the field is a critical factor supporting both relevance and generalizability. Those involved in lab research usually aim to demonstrate the presence of a particular effect, and tend to be motivated to create a specific environment or context to observe it. Lab researchers have an unlimited number of levers to establish conditions which will maximize the chances for observing desired effects. Rigorous control procedures can be implemented that are not feasible outside the lab. But this precise control may be exactly what limits generalizability.

Field researchers face the opposite problem. They typically work in environments which can be changed very little, and with populations they rarely can preselect. Field/applied researchers are routinely motivated to search for effects and manipulations which are robust enough to work in their specific context. Field research may operate as a "generalizability filter" separating tenuous effects from interventions with a higher chance for success.

Judgment and decision-making research may have benefited from the fact that much of it has been done in business schools. Business school faculty rarely have access to a "subject pool" and they tend to rely on both studies in classrooms and in the field. The participants in business school studies often are students who have experience in the business world and are seeking MBAs (or PhDs). This is just one factor that serves to increase the likelihood that research by business school faculty will make connections with corporate contexts.

Consider, for example, "sunk cost" effects. Sunk costs refer to situations where commitment of resources is continued and escalated beyond any rational considerations because one doesn't want to "waste" the prior investment. This is sometimes referred to as "throwing good money after bad." The interest in sunk cost effects originated with real-world examples. But a careful analysis of generalizability suggests that there are other situations where the opposite of sunk cost effects can be shown (prematurely withdrawing an investment just before it starts to pay off; e.g., Drummond, 2014; Heath, 1995). Instead of undermining the sunk costs construct, such findings invite attention to what factors are associated with each type of outcome. For instance, sunk cost effects for money may be different from sunk cost effects for time (Cunha, Marcus, & Caldieraro, 2009; Soman, 2001).

Field research may also serve as a direct test of generalizability of lab findings. For example Hofmann, Wisneski, Brant, and Skitka (2014) used text messaging at varied times to assess everyday moral and immoral acts and experiences. They found moral experiences to be common and, they observed both moral licensing and moral contagion, effects that previously had been shown in lab studies.

This interplay between lab and field is useful to both. Although generalizability is important, it could be argued that variability is even more fundamental. At the heart of social science is the search for patterned variation, variation that our theories seek to understand. Attention to the field may serve to increase attention to potential interactions and undermine a main effect focus.

Field as a source of complementary evidence

As Yarkoni notes, conceptual replications (as opposed to exact replications) put assumptions of generalizability to the test and represent an effective research strategy. They also are a key tool in establishing construct validity (e.g., Grahek, Schaller, & Tackett, 2021), linking theory and measures.

Field observation offers a complementary form of converging measure that can be an important research tool. For example, lab studies suggesting that participants see nature as incompatible with human presence (nature is pristine and humans can enjoy it but are not part of it) can be complemented by analyses using Google images. For example, a search of images for "ecosystems" found that humans were present only two percent of the