Running Head: Commentary on Yarkoni (in press)

Exposing and overcoming the fixed effect fallacy through crowd science

Wilson Cyrus-Lai
INSEAD

Warren Tierney
INSEAD

Martin Schweinsberg
ESMT Berlin

Eric Luis Uhlmann
INSEAD

(Commentary on Tal Yarkoni, "The generalizability crisis")

ABSTRACT: 56 words
MAIN TEXT: 994 words
REFERENCES: 428 words
ENTIRE TEXT: 1652 words

Authors' Note: Please address correspondence to Wilson Cyrus-Lai (wilson-cyrus.lai@insead.edu), Warren Tierney (warren.tierney@insead.edu), Martin Schweinsberg (martin.schweinsberg@esmt.org) and Eric Uhlmann (eric.luis.uhlmann@gmail.com)

CONTACT INFORMATION:

Wilson Cyrus-Lai
Organisational Behaviour Area
INSEAD
1 Ayer Rajah Avenue 138676
Singapore
Phone: 65 9022 0155
E-mail: wilson-cyrus.lai@insead.edu

Warren Tierney
Organisational Behaviour Area / Marketing Area
INSEAD
1 Ayer Rajah Avenue 138676
Singapore
Phone: 353 87329150
E-mail: warren.tierney@insead.edu

Martin Schweinsberg
Organisational Behaviour Area
ESMT Berlin
Schlossplatz 1, 10178
Berlin Germany
Phone: 49 30 212 31-1549
E-mail: martin.schweinsberg@esmt.org

Eric Luis Uhlmann
Organisational Behaviour Area
INSEAD
1 Ayer Rajah Avenue 138676
Singapore
Phone: 65 8468 5671
E-mail: eric.luis.uhlmann@gmail.com

Abstract: By organizing crowds of scientists to independently tackle the same research questions, we can collectively overcome the generalizability crisis. Strategies to draw inferences from a heterogeneous set of research approaches include *aggregation*, for instance meta-analyzing the effect sizes obtained by different investigators, and *parsing*, attempting to identify theoretically meaningful moderators that explain the variability in results.

Yarkoni (in press) highlights the *fixed effect fallacy*, arguing that many if not most research findings are unlikely to prove robust to stimulus sampling and task operationalizations. Experimental studies in psychology and related fields are exposed to the possibility that the effect is specific to the stimulus set in question, such that alternative approaches could have attenuated or even reversed the reported finding. Recent initiatives to crowdsource the analyses of complex datasets (Bastiaansen et al., 2020; Botvinik-Nezer et al., 2020; Silberzahn et al., 2018; Schweinsberg et al., 2021), and design of experiments (Baribault et al., 2018; Landy et al., 2020) provide strong quantitative evidence for these assertions. When different scientists independently analyze the same dataset to try and answer the same research question, or separately create their own experimental design to test the same hypothesis, a wide range of results are obtained.

These large-scale crowd science projects illustrate two key approaches to drawing robust conclusions and building strong theory through diversity in approaches and results. One strategy to overcoming the generalizability challenge is ***aggregation***, for example simply meta-analyzing across the estimates obtained by independent analysts or from different experimental designs. Another is ***parsing***, or attempting to find meaningful moderators that explain why some approaches yield large estimates and others small to null estimates or even estimates reversed in sign.

The parsing strategy is in harmony with the perspectivist approach to theoretical progress, which assumes that most phenomena in the social sciences are massively moderated (McGuire, 1973, 1983). From this perspective, “the opposite of a great truth is also true” (Banaji, 2003), and thus it is unsurprising that different empirical approaches to testing the same idea can return effect size estimates that are opposed in sign. The fundamental task of researchers, from a perspectivist standpoint, is to untangle this web by identifying moderators that will allow us to predict when effects emerge, disappear, and reverse. However, we suggest that aggregation and parsing can be complementary rather than competing: meta-scientists can both meta-analyze across crowdsourced approaches and seek to meaningfully explain variability in effect sizes.

In an illustration of the aggregation strategy, Landy et al. (2020) recruited up to thirteen research teams to independently create experimental stimulus sets to test the same set of five original hypotheses, all supported in unpublished research by the original authors (e.g., “working for no reason is morally praised”, “deontologists are happier than consequentialists”). Over 15,000 research participants were randomly assigned to the different study designs. All five original effects directly replicated using the same stimulus set the original authors had used. However, four of five hypotheses had different material-makers created designs that returned statistically significant effects in opposite directions from one another. At the same time, two out of five original hypotheses proved conceptually robust when meta-analyzing the results across the experimental designs from the different teams of researchers. This maps on closely to predictions by Yarkoni (in press) and others, that even when directly replicable, only a minority of findings in social psychology and related fields will prove generalizable across contexts and approaches.

Employing both the aggregation and parsing strategies together, Schweinsberg et al. (2021) asked up to fifteen independent researchers to test two hypotheses using the same dataset capturing gender and status dynamics in intellectual debates. Not only statistical choices (e.g., covariates), but also the operationalization of variables (e.g., status) were left unconstrained and up to the individual researchers' discretion. For example, an analyst could choose to identify high vs. low status academics using job rank, citation counts, PhD institution rank, or a combination of indicators. No two researchers employed the same specification. For both hypotheses, independent analysts reported statistically significant estimates in opposite directions despite relying on the same dataset. Hypothesis 1 (women speak more in the presence of other women) was supported while aggregating across different measurement and testing approaches, whereas Hypothesis 2 (high status academics speak more) was comparatively not, with estimates distributed around zero in the latter case. Leveraging a Boba multiverse analysis (Liu et al., 2020; see also Steegen et al., 2016) to identify key analyst choice points, Schweinsberg et al. (2021) further demonstrate that differing variable operationalizations directly contributes to this radical dispersion in estimates across different analysts. For example, researchers who operationalized status as job rank consistently returned negative estimates for H2, whereas those operationalizing status using ranking of doctoral institution returned consistently positive estimates. This illustrates how the parsing strategy treats variability across different approaches as clues to meaningful moderation, rather than error to be averaged away.

In order to draw generalizable conclusions, Tierney, Cyrus-Lai, et al. (2021) assigned teams of doctoral students and professors to separately create conceptual replication designs testing for backlash against angry women. The original study finds that although male managers who

express anger (relative to sadness or neutral emotions) experience a boost in status, female managers who express anger are accorded less social status and respect (Brescoll & Uhlmann, 2008). Participants in this ongoing data collection across over 50 laboratories are randomly assigned to one of 27 study designs (the original design and 26 conceptual replication designs) testing the hypothesized interaction between target gender and emotion expression. The employed methods include scenarios, ostensive newspaper stories, audio recordings, video recordings, and storyboards with illustrated characters as well as a myriad of different ways of expressing anger. In addition to a pre-registered meta-analysis of the results across designs, we will systematically test potential moderators of the results across designs; among these are anger extremity, dominance displays, and the salience of target gender.

In summary, we can collectively overcome the generalizability crisis by organizing crowds of scientists to tackle the same research questions independently. Doing so will further expose the fixed effect fallacy that a single analysis and research paradigm are sufficient for drawing strong theoretical inferences. Scientists can rely on the wisdom of the crowd by aggregating results across independent investigators, and seek to identify meaningful moderators of the results across different approaches, in the perspectivist spirit.

Funding/financial support statement

This research was supported by an R&D grant from INSEAD to Eric Uhlmann.

Conflict of interests declaration

Conflicts of interest: none.

References

Banaji, M. R. (2003). The opposite of a great truth is also true: Homage of Koan #7. In J. Jost, D. Prentice, & M. R. Banaji (Eds.), *The yin and yang of progress in social psychology: Perspectivism at work* (pp. 127-140). Washington, DC: American Psychological Association.

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.

Bastiaansen, J.A., Kunkels, Y.K., Blaauw, F.J., et al. (2020). Time to get personal? The impact of researchers choices on the selection of treatment targets using the experience sampling methodology. *Journal of Psychosomatic Research, 137,* 110211.

Botvinik-Nezer, R., Holzmeister, F., Camerer, C. F., Dreber, A., Huber, J., Johannesson, M., et al. (2020). Variability in the analysis of a single neuroimaging dataset by many teams. *Nature,* 582, 84–88.

Brescoll, V., & Uhlmann, E.L. (2008). Can angry women get ahead? Status conferral, gender, and workplace emotion expression. *Psychological Science, 19*, 268-275.

Landy, J. F., Jia, M., Ding, I. L., Viganola, D., Tierney, W., et al.,… & Uhlmann, E.L. (2020). Crowdsourcing hypothesis tests: Making transparent how design choices shape research results. *Psychological Bulletin*, *146*(5), 451–479.

Liu, Y., Kale, A., Althoff, T., & Heer, J. (2020). Boba: Authoring and Visualizing Multiverse

Analyses. *IEEE Transactions on Visualization and Computer Graphics (Proc. VAST)*.

McGuire, W. J. (1973). The yin and yang of progress in social psychology: Seven koan. *Journal of Personality and Social Psychology, 26*(3), 446-456.

McGuire, W.J. (1983). A contextualist theory of knowledge: Its implications for innovations and reform in psychological research. In L. Berkowitz (Ed.), *Advances in Experimental Social Psychology* (Vol. 16, pp. 1-47). New York, NY: Academic Press.

Schweinsberg, M., Feldman, M., Staub, N., …Tierney, W., … Cyrus-Lai, W., … & Uhlmann, E. (2021). Radical dispersion of effect size estimates when independent scientists operationalize and test the same hypothesis with the same data. *Organizational Behavior and Human Decision Processes*.

Silberzahn, R., Uhlmann, E. L., Martin, D., Anselmi, P., Aust, F., Awtrey, E., et a., & Nosek, B.A. (2018). Many analysts, one dataset: Making transparent how variations in analytical choices affect results. *Advances in Methods and Practices in Psychological Science, 1*, 337–356.

Steegen, S., Tuerlinckx, F., Gelman, A., & Vanpaemel, W. (2016). Increasing transparency through a multiverse analysis. *Perspectives on Psychological Science, 11*(5), 702–712.

Tierney, W., Cyrus-Lai, W., et al., & Uhlmann, E.L. (2021). *Who respects an angry woman? A pre-registered re-examination of the relationships between gender, emotion expression, and status conferral*. Crowdsourced research project in progress.