Running Head: Commentary on Almaatouq et al. (in press)

Test Many Theories in Many Ways

Wilson Cyrus-Lai
INSEAD

Warren Tierney
INSEAD

Eric Luis Uhlmann
INSEAD

(Commentary on Almaatouq et al., "Integrative Experiment Design in the Social and Behavioral Sciences")

ABSTRACT: 60 words
MAIN TEXT: 994 words
REFERENCES: 747 words
ENTIRE TEXT: 1949 words

Authors' Note: Please address correspondence to Wilson Cyrus-Lai (wilson-cyrus.lai@insead.edu)

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

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

Demonstrating the limitations of the one-at-a-time approach, crowd initiatives reveal the surprisingly powerful role of analytic and design choices in shaping scientific results. At the same time, cross-cultural variability in effects is far below the levels initially expected. This highlights the value of "medium" science, leveraging diverse stimulus sets and extensive robustness checks to achieve integrative tests of competing theories.

Almaatouq et al. (in press) argue that the “one-at-a-time” approach to scientific research has led to collections of atomized findings of unclear relevance to each other. They advocate for an integrative approach in which stimuli are varied systematically across theoretically important dimensions. This allows for strong inferences (Platt, 1964) regarding which theory holds the most explanatory power across diverse contexts, as well as the identification of meaningful moderators.

Our research group has addressed this challenge by examining the analytic and design choices that naturalistically emerge across independent investigators as well as the implications for the empirical results (Landy et al., 2020; Schweinsberg et al., 2021; Silberzahn et al., 2018). These crowdsourced many analysts and many designs initiatives reveal dramatic dispersion in estimates due to researcher choices, empirically demonstrating the limitations of the one-at-a-time approach (see also Baribault et al., 2018; Botvinik-Nezer et al., 2020; Breznau et al., 2022; Menkveld et al., 2021). At the same time, we have sought to further increase the already high theoretical value of replications by leveraging them for competitive theory testing. Rather than test the original theory against the null hypothesis, we include new conditions and measures allowing us to simultaneously examine the pre-registered predictions of different theoretical accounts (Tierney et al., 2020, 2021). In this manner, we can start to prune the dense theoretical landscape (Leavitt et al., 2010) found in areas of inquiry characterized by many atomized findings and narrow theories.

In contrast, a striking and unexpected *lack* of variability has emerged in the results when many laboratories collect data using the same methods. In such crowd replication initiatives, cross-site heterogeneity in estimates is far below what one would expect based on intuition and theory (Olsson-Collentine et al., 2020). From a perspectivist (McGuire, 1973) standpoint,

psychological phenomena should emerge in some contexts and be nonexistent or even reversed in others (see also Henrich et al., 2010). And yet, effects seem to either fail to replicate across all populations sampled or emerge again and again (see also Delios et al., 2022).

Bringing many designs, analyses, theories, and data collection teams together, we recently completed a crowdsourced initiative that qualifies as the type of comprehensive integrative test that Almaatouq et al. (in press) envision. Tierney et al. (2023) systematically re-examined the relationships between anger expression, target gender, and status conferral. In the original research, women who displayed anger in professional settings suffered steep drops in the status and respect they were accorded by social perceivers (Brescoll & Uhlmann, 2008). In the original investigations, only a single set of videos featuring one female and one male target were employed as stimuli, and all participants were from Connecticut. In contrast, the crowdsourced replication project featured 27 experimental designs, a multiverse capturing many defensible analytic approaches, and 68 data collection sites in 23 countries. We further tested the original prescriptive stereotype account against competing theories predicting that anger signals status similarly for women and men, that anger has vastly different status implications in Eastern and Western cultures, and that feminist messaging has successfully reduced or even reversed gender biases. As Almaatouq et al. (in press) recommend, we probed the dose-response relationship between anger and status conferral by both experimentally manipulating and measuring the extremity of emotion expressions across different designs.

The crowd initiative finds that anger increases status by signaling dominance and assertiveness, while also diminishing it by projecting incompetence and unlikability,

aggregating across a wide range of research approaches and populations. Critically, this same pattern emerged for both female and male targets, social perceivers of different genders, and in both Eastern and harmony-oriented cultures and Western and more conflict-oriented ones. Highlighting the value of deploying diverse research approaches, six of the 27 designs found favoritism towards men in status conferral, but one design pointed to the opposite conclusion. Similarly, in a multiverse with 32 branches, there existed just two specifications that supported the original gender-and-anger backlash effect. Had we employed a one-at-a-time approach, we could have accidentally happened upon or strategically chosen narrow methods yielding non-representative conclusions (e.g., of pro-female status bias or gender backlash). Overall, the intellectual returns on including many designs, many analyses, and many theories was high. In contrast, and consistent with past crowd initiatives, collecting data across many places revealed minimal cross-site heterogeneity and no interesting cultural differences.

Thus, we envision a diverse scientific ecology consisting of many “small” and “medium” projects and just a few huge international efforts. The one-at-a-time approach is an efficient means to introduce initial evidence for promising new hypotheses. However, as a theoretical space becomes increasingly cluttered, intellectual returns are maximized by sampling stimuli widely and employing many analyses to provide severe tests of competing theories (Mayo, 2018). Although this could involve a crowd of laboratories, a single team could carry out a multiverse (Steegen et al., 2016) and operationalize key variables in a variety of ways. A small team might sample just one or two participant populations that are easily accessible to them. Finally, a subset of findings of particularly high theoretical and practical importance should be selected for crowdsourced data collections across many nations as a systematic test of cross-cultural generalizability. When numerous sites are not available, the researchers

might carry out the first generalizability test in the most culturally distant population available (Muthukrishna et al., 2020). If the effect is still observed, this represents initial evidence of universality (Norenzayan & Heine, 2005).

In sum, an ironic legacy of the movement to crowdsource behavioral research may be showing that scaling science to such a massive level might be neither efficient nor strictly necessary for most research findings. The sorts of integrative tests Almaatouq et al. (in press) envision can also be accomplished by a small team that actively ensures a diversity of analyses and stimuli, and yet collects data locally or across a few carefully selected cultures rather than globally. In the future, our greatest intellectual returns on investment may come from "medium" science that prioritizes testing many theories in many ways.

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

Almaatouq, A., Griffiths, T.L., Suchow, J.W., Whiting, M.E., Evans, J., & Watts, D.J. (in press). Beyond playing 20 questions with nature: Integrative experiment design in the social and behavioral sciences. *Behavioral and Brain Sciences.*

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.

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. L., & Uhlmann, E. L. (2008). Can an angry woman get ahead? Status conferral, gender, and expression of emotion in the workplace. *Psychological Science, 19*, 268–275.

Breznau, N., Rinke, E. M., Wuttke, A., Nguyen, H. H., Adem, M., Adriaans, J., ... & Van Assche, J. (2022). Observing many researchers using the same data and hypothesis reveals a hidden universe of uncertainty. *Proceedings of the National Academy of Sciences, 119(44)*, e2203150119.

Delios, A., Clemente, E., Wu, T., Tan, H., Wang, Y., Gordon, M., Viganola, D., Chen, Z., Dreber, A., Johannesson, M., Pfeiffer, T., Generalizability Tests Forecasting Collaboration, & Uhlmann, E.L. (2022). Examining the context sensitivity of research findings from archival data. *Proceedings of the National Academy of Sciences*, *119*(30), e2120377119.

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences*, *33*, 61-83.

Landy, J. F., Jia, M., Ding, I. L., Viganola, D., Tierney, W., Dreber, A., Johanneson, M.,

Pfeiffer, T., . . . & Uhlmann, E. L. (2020). Crowdsourcing hypothesis tests: Making transparent how design choices shape research results. *Psychological Bulletin, 146*(5), 451–479.

Leavitt, K., Mitchell, T., & Peterson, J. (2010). Theory pruning: Strategies for reducing our dense theoretical landscape. *Organizational Research Methods, 13*, 644-667.

Mayo, D. G. (2018). *Statistical inference as severe testing: How to get beyond the statistics wars.* Cambridge University Press.

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.

Menkveld, A. J., Dreber, A., Holzmeister, F., Huber, J., Johannesson, M., Kirchler, M., Razen, M., Weitzel, U., Abad, D., … Wu, Z.-X. (2021). *Non-standard errors*. Tinbergen Institute Discussion Paper TI 2021-102/IV. https://papers.tinbergen.nl/21102.pdf

Muthukrishna, M., Bell, A.V., Henrich, J., Curtin, C.M., Gedranovich, A., McInerney, J. & Thue, B. (2020). Beyond western, educated, industrial, rich, and democratic (WEIRD) psychology: Measuring and mapping scales of cultural and psychological distance. *Psychological Science*, *31*(6), 678-701.

Norenzayan, A., & Heine, S. J. (2005). Psychological universals: What are they and how can we know? *Psychological Bulletin, 135*, 763-784.

Olsson-Collentine, A., Wicherts, J.M., & van Assen, M.A.L.M. (2020). Heterogeneity in direct replications in psychology and its association with effect size. *Psychological Bulletin*, *146*(10), 922-940.

Platt, J. R. (1964). Strong inference. *Science, 146*, 347-353.

Schweinsberg, M., Feldman, M., Staub, N., van den Akker, O., van Aert, R., van Assen, M.,

Liu, Y., … & 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, 165,* 228-249.

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,* 702–712.

Tierney, W., Cyrus-Lai, W., et al.,… & Uhlmann, E.L. (2023). *Who respects an angry woman? A pre-registered re-examination of the relationships between gender, emotion expression, and status conferral*. Unpublished manuscript.

Tierney, W., Hardy, J. H., III., Ebersole, C., Leavitt, K., Viganola, D., Clemente, E., Gordon, M., Dreber, A.A., Johannesson, M., Pfeiffer, T., Hiring Decisions Forecasting Collaboration, & Uhlmann, E. (2020). Creative destruction in science. *Organizational Behavior and Human Decision Processes, 161,* 291-309.

Tierney, W., Hardy, J. H., III., Ebersole, C. R., Viganola, D., Clemente, E. G., Gordon, M., Hoogeveen, S., Haaf, J., Dreber, A. , Johannesson, M., Pfeiffer, T., Huang, J. L., Vaughn, L. A., DeMarree, K.G., Igou, E., Chapman, H., Gantman, A., Vanaman, M., Wylie, J., Storbeck J., Andreychik, M. R., McPhetres, J., Culture and Work Forecasting Collaboration, & Uhlmann, E. L. (2021). A creative destruction approach to replication: Implicit work and sex morality across cultures. *Journal of Experimental Social Psychology, 93,* 104060.