Running Head: DEMOGRAPHIC REQUIREMENTS

Concerns that demographic reporting requirements are deleterious and unethical to psychology

David Trafimow

Michael C. Hout

Andrew R. A. Conway

New Mexico State University

Direct correspondence to:

David Trafimow
Department of Psychology, MSC 3452
New Mexico State University
P. O. Box 30001
Las Cruces, NM 88003-8001
dtrafimo@nmsu.edu

Abstract

According to the journal scope statement of the *Journal of Experimental Psychology: Human Perception and Performance* (JEP:HPP), researchers are now required to report demographics and justify their sample compositions. However, we feel that the requirement is indefensible on both conceptual and ethical grounds. Conceptually, the requirement wrongly emphasizes generalizing findings rather than generalizing theories without recognizing the crucial role auxiliary assumptions play in the generalization process. Moreover, it distracts researchers with a focus on theoretically irrelevant measures, fails to distinguish between including demographics as moderators in analyses versus as mere classification percentages, encourages researchers to commit the fallacy of using interindividual summary statistics to draw conclusions at the intraindividual level, and potentially reduces sampling precision. Ethically, the requirement places poor or minority researchers at a disadvantage and has the potential to create unnecessary anxiety for participants. It also pushes European and British researchers to violate the General Data Protection Regulation that operates in Europe and the UK, thereby placing those researchers in an untenable situation.

Keywords: demographics; auxiliary assumptions; generalizability; ethics

Word count: 3679

In their journal scope statement, the *Journal of Experimental Psychology: Human Perception and Performance* (JEP:HPP) states the following as a requirement for authors (https://www.apa.org/pubs/journals/xhp).

> It is widely recognized that at least some psychological results are correlated with demographics. Thus, this journal follows APA position that basic demographics be provided in a manuscript because without this information the reader would have no basis for knowing to whom to generalize the findings.

The statement concerns us not only because it sets policy for a premier journal in the area of human perception and performance, but additionally because it is APA policy and the "journal follows APA position". The implication is that psychology authorities at APA, not psychology experts on the editorial board at JEP:HPP, determined it is vital to know "to whom to generalize findings". This is cause for concern. More specifically, we argue that the new policy statement indicates an indefensible philosophical commitment that harms psychological science and is ethically problematic.

Testing theories and models of human perception and performance

JEP:HPP is a basic research journal that publishes studies on perception, control of action, perceptual aspects of language processing, and related cognitive processes. More generally, JEP:HPP solicits articles that propose and test theories or models of human perception and performance. In psychological science, theory testing is a tricky endeavor. This is because theories and models contain nonobservational terms that refer to unobservable entities. For example, the Nobel Laureate, Leon Lederman (1993) asserted that (a) Newton's $force = mass \cdot acceleration$ is the most important equation in the history of physics, but that (b) $mass$ is not directly observable and does not have an independent definition. Mass should not be confused with weight, which is observable; the same object would have the same mass—but different

weights—on Earth or the Moon. In psychology, cognitive processes pertaining to attention, perception, motor control, language (and so on) are also not directly observable.

Because theoretical terms are nonobservational, there is no way to test theories directly. There is a gap between nonobservational theoretical terms and observational empirical terms in an empirical hypothesis. A physicist who wishes to use weight measurements to draw conclusions about object mass is forced to make additional assumptions connecting weights to mass. Likewise, a psychologist who wishes to relate observable reaction times to unobservable cognitive processes is forced to make additional assumptions too, though these may be tacit. Science philosophers and philosophically-oriented psychologists have long recognized the necessity to make additional assumptions, not in the theory proper, to traverse the distance between nonobservational terms in theories, and observational terms in empirical hypotheses researchers use to test theories (Duhem, 1954; Lakatos, 1978; Meehl, 1990; Popper, 1963; 1972; Quine, 1952 Trafimow, 2009; 2012; 2023; Whewell, 1840). These additional assumptions are often termed *auxiliary assumptions*. To test a theory or model, a researcher must have (a) the theory or model and (b) auxiliary assumptions. In addition, auxiliary assumptions can serve to set initial experimental conditions (Hempel, 1965), though this is not our main concern here.

Although replication failures can be attributed to problems with the theory (or model) itself, replication failures can alternatively be attributed to inappropriate auxiliary assumptions. For example, consider a theory of selective attention (in healthy adults) which states that the processing of irrelevant information is determined by the perceptual “load” of relevant information (Lavie, 1995); low load tasks (e.g., those involving simple stimuli, easy discrimination/identification) require minimal attentional resources and high load tasks (e.g., those with complex stimuli, difficult discrimination/identification) require more attentional

resources. Imagine that a research team randomly assigns participants to a low or high perceptual load condition. Consistent with the theory and empirical hypothesis, participants in the low perceptual load condition process irrelevant "distractor" information to a large degree (presumably because attentional capacity was not fully utilized, and therefore "spilled over" to the secondary task/info), and participants in the high load condition (whose attentional capacity was presumably fully utilized) did not process the irrelevant information to the same degree. Then, another research team attempts to replicate this work and uses participants from a different culture (or demographic) and fails to find the same effects. How serious is this generalization failure?

One possible reason for the generalization failure, of course, is that the theory is wrong, but an alternative reason is that at least one auxiliary assumption that was formerly appropriate is inappropriate in the new culture/demographic. For example, perhaps due to different perceptually relevant experiences, the perceptual load manipulation is more effective in the culture sampled by the original study (Culture A) than it is in the culture sampled by the study attempting a replication (Culture B). In that case, the generalization failure need not discommode proponents of the theory. We'd argue that the generalization failure could even be advantageous for the theory.

To see how, imagine another experiment is conducted in Culture B, and the researchers perform a culturally appropriate perceptual load manipulation that works as hypothesized. We would expect theories to predict correctly when combined with appropriate auxiliary assumptions and not to predict correctly when combined with inappropriate ones. In the present case, the theory predicts correctly when combined with appropriate auxiliary assumptions, but

not when combined with inappropriate ones. Thus, the case for the theory is stronger with the ostensible generalization failure than without it.

We can now make a preliminary statement concerning the JEP:HPP requirements. A failure to generalize a finding is unimportant if it can be attributed to a problematic auxiliary assumption. Trafimow (2023) distinguished between Sense 1 external validity that pertains to generalizing findings, versus Sense 2 external validity that pertains to generalizing theories, and provided numerous illustrations that Sense 2 external validity is much more important than Sense 1 external validity. Unfortunately, the JEP:HPP requirements reverse progress and throw the field back to an unsophisticated, findings-based conception of external validity when the focus ought to be on theories (or models if a fully-articulated theory is unavailable).

What demographics?

Lest the present argument be misconstrued, our point is not that it never matters when findings fail to generalize. Rather, our point is that the reason it matters, in those cases where it does matter, is because of implications for the theory. If findings fail to generalize because an auxiliary assumption that used to be appropriate no longer is, this is not a problem for the theory.

The present argument can be interpreted as militating against the emphasis of JEP:HPP on generalizing findings (as opposed to theory), but not against the conclusion that researchers should report comprehensive demographic information. It is possible that reporting comprehensive demographic information is excellent for theory testing because it provides more opportunities to find failures of findings to generalize which, in turn, provides opportunities to distinguish cases where the failures are due to problematic auxiliary assumptions versus problematic theories.

If this were the argument, the explicit recognition of the cruciality of theory and distinction between theoretical and auxiliary assumptions would render us significantly less concerned with the JEP:HPP statement, though there would remain an important problem. To see the problem, consider again the example of perceptual load theory (Lavie, 1995). Perhaps it is the case that artists can tolerate more of a perceptual load than non-artists. Perhaps extraversion, neuroticism, or other personality traits moderate the perceptual load effect. There are an indefinite number of possible moderators and it is obviously impossible to measure all of them.

Consider next the typical 'experimental hour' that most experimental psychology researchers have within which they can test participants. The more time spent assessing potential effect moderators, the less the time is available for the researcher to collect behavioral data from their participant. In turn, fewer observations per participant will reduce precision. For instance, suppose a researcher collects either 25 or 50 observations for each participant, under normality. If $n = 25$, there is a 95% probability that a participant's mean will be within 0.395 standard deviations of that person's population mean, whereas if $n = 50$, there is a 95% probability that a participant's mean will be within 0.278 standard deviations. The difference between 0.395 standard deviations and 0.278 standard deviations is 0.117 standard deviations, a considerable precision disadvantage associated with fewer trials. (The calculations can be made easily using the calculator at the following link: https://app-normal.shinyapps.io/N SingleSample EstimateMean KnownVariance, see Li et al., 2020.) Thus, there is scarcely room for deadwood measures that are irrelevant to the research goals.

It is not our argument that researchers should never measure demographic variables, only that they should have a theoretical reason for doing so in the first place. For example, if a researcher believes that the process by which perceptual load influences distractions works

differently for those who identify as females versus males (or for old versus young, or some other group difference), then it clearly makes sense to collect that information. In fact, if such demographics are theoretically relevant, researchers should include them in the main statistical analysis to formally assess their moderating effects. In contrast, if the variables are not theoretically relevant, then it makes little sense to use up valuable minutes in the experimental session measuring them.

A sensible objection might be that the researcher might not know if a demographic variable is theoretically relevant sans collection. However, even researchers who collect a demographic will not know if it is theoretically relevant unless the demographic is included in the main data analysis as a potential effect moderator. Thus, rather than demographics that JEP:HPP demands (or encourages) as a condition for sending a manuscript out to reviewers, a better strategy would be to demand only those demographics that are included in the main data analysis as potential moderators. Collecting data that are not analyzed—other than as a cursory report of percentages of participants in different categories to satisfy the JEP:HPP requirements—seems wasteful of valuable experimental time that could be better invested.

We reiterate that collecting demographics will not answer the question in the JEP:HPP statement about knowing "to whom to generalize findings" unless the demographics are included in the main data analysis as potential moderators. If the journal is serious about the importance of knowing "to whom to generalize findings," that seriousness demands that all demographic measures be included in the main analyses (as potential moderators) to address the issue. However, as the journal has not made this requirement, it is difficult to assess how serious JEP:HPP considers the issue to be. Perhaps the hope is that other researchers will experience generalization failures, note the difference in demographics from the original experiment, and

make appropriate attributions about the moderating demographics. But unless these other researchers formally test the hypothesized demographics as potential moderating variables, the mere inclusion of demographic information will be insufficient to address the issue.

Cases where researchers use demographics as moderating variables in their main analyses to obtain revealing insights are few and far between in the literature on perception and performance. More common is for researchers to use group level summary statistics to draw conclusions about individuals. However, a wealth of conceptual and empirical analyses shows that interindividual summary statistics often fail to generalize to the intraindividual level (e.g., Castro-Schilo, & Ferrer, 2013; Fisher, et al., 2018; Hamaker, Dolan, & Molenaar, 2005; Molenaar, 2004; 2008; Trafimow & Finlay, 1996; Trafimow, Kiekel, & Clason, 2004; Trafimow & Rice, 2008). Thus, the hope that interindividual summary statistics pertaining to demographics would shed light on intraindividual processes involved in perception or performance is optimistic at best, and fanciful at worst. Demographics are not process variables that provide insights into perceptual or performance mechanisms; they are merely proxies that often mislead.

For example, based on a demographic difference between Blacks and Whites on IQ (or related) scores, researchers across decades have concluded that Blacks are genetically inferior to Whites with respect to intelligence (see Herrnstein & Murray, 1994 for a famous example). Such inferences have persisted despite a devastating debunking by Allport (1954). Researchers have used (and continue to use) demographic effects to invalidly justify prior biases, including those against minorities. The mingling of demographic summary data with IQ (or related) scores exemplifies the potential dangers of demographics-based theorizing.

Hence, we see two philosophically justifiable choices. JEP:HPP should either (1) require all demographics be included in the main analyses as potential moderating variables or (2)

reconsider their position that it is crucial to determine to whom findings generalize and drop the demographics reporting requirement. Choice 1 would constitute a sea change in the field, so we see Choice 2 as obviously preferable.

Ethical issues

JEP:HPP states the following:

> "Authors are encouraged to **justify their sample demographics** in the discussion section. If Western, educated, industrialized, rich, and democratic (WEIRD) or all-White samples are used, authors should justify their samples and describe their sample inclusion efforts…" (boldfacing in the original)

All else being equal, we applaud efforts to be more inclusive, but all else may not be equal. Many (perhaps most) experiments in JEP:HPP use undergraduate students enrolled in introductory psychology; i.e., typical 'convenience' samples. The implication of the boldfacing in the JEP:HPP statement is that a convenience justification is insufficient. If researchers are required to obtain inconvenient samples, this would likely require financial resources that many researchers might not have (e.g., those at smaller institutions, or institutions with a teaching rather than a research focus). Thus, the JEP:HPP policy inadvertently increases the advantage that richer researchers have over poorer ones, or the advantage that researchers with grant money have over researchers without grant money. Ironically, many of the researchers who are poorer or devoid of grant money are likely non-White (Lauer & Roychowdhury, 2021; Odedina & Stern, 2021; Taffe & Gilpin, 2021). Thus, the net result could be doubly deleterious; only researchers with substantial resources would be able to publish in the journal and proportionately fewer of these would be non-White, the very people the policy is likely intended to aid. As the ultimate justification for the existence of scientific journals is to improve science and thereby better the human condition, reducing the number of researchers able to participate, whether

White or non-White, reduces competition for publication, to the harm of everyone that might otherwise benefit from the science.

Another ethical issue stems from a concern for participant welfare. Researchers are encouraged to obtain "nativity or immigration history." However, if a participant is in a country without necessary legal documentation, then asking this question is likely to cause undue anxiety. Certainly, researchers could include a "refuse to respond" option, but that is unlikely to prevent anxiety experienced by the participant; the participant might have concerns about the ramifications of endorsing that option, or they might feel as if the researchers are unwelcoming to undocumented individuals (simply by virtue of having asked that question), or a variety of other scenarios. One might argue that—for the sake of such anxiety not impacting behavioral data collection—such questions should come at the end of a study. Now imagine a training study that requires a participant to attend multiple sessions. Even if those questions are asked at the completion of the first training session, an undocumented participant might be made to feel uncomfortable and opt not to come back for further studies. In this instance, the policy of asking this question would end up excluding people from diverse groups rather than encouraging them.

A similar set of problems potentially plague other items on the list, such as "clinical diagnoses and comorbidities (as appropriate)," "sexual orientation," and "socioeconomic status." It is one thing to collect sensitive information when it is theoretically relevant, and where potential benefit to science could be argued to balance potential harm to participants. It is quite another thing to cause participants anxiety by bringing up sensitive topics, or to remind them that they are poor or uneducated (by collecting socioeconomic data), when these are theoretically irrelevant and the potential benefit to science is extremely hypothetical. In our view, a more scientifically and ethically justifiable way to promote diversity would be if researchers were to

combine experimental and individual differences approaches to form or test theories or models that integrate group and individual level effects.

Finally, the General Data Protection Regulation (GDPR) that operates in Europe and the UK states the following: "The GDPR requires you to be clear about the purposes for which you collect personal data, to only collect the minimum amount of personal data you need for those purposes…" Hearkening back to an earlier point, if personal data are theoretically relevant, they should feature in the main data analyses; otherwise, the researcher is violating the GDPR. The journal policy potentially places researchers from Europe and the UK at a significant disadvantage; complying with the GDPR may be inconsistent with complying with the journal policy.

Conclusion

In conclusion, the JEP:HPP requirements are indefensible on both philosophy of science and ethical grounds. A reversion from generalizing theory to generalizing findings constitutes a backward step for psychological science, and collecting personal data that are theoretically irrelevant is unethical. We hope and expect that journal editors will carefully consider the present arguments and avoid demographics requirements that do more harm than good.

References

Allport, G. W. (1954). *The nature of prejudice.* Addison-Wesley.

Castro-Schilo, L., & Ferrer, E. (2013). Comparison of nomothetic versus idiographic-oriented methods for making predictions about distal outcomes from time series data. *Multivariate Behavioral Research*, *48*(2), 175–207. https://doi.org/10.1080/00273171.2012.736042

Duhem, P. (1954). *The aim and structure of physical theory* (P. Wiener, Trans.). New York, NY: Atheneum. (Original work published 1914)

Fisher, A. J., Medaglia, J. D., & Jeronimus, B. F. (2018). Lack of group-to-individual generalizability is a threat to human subjects research. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(27), E6106–E6115. https://doi.org/10.1073/pnas.1711978115

Hamaker E (2012) Why researchers should think "within-person": A paradigmatic rationale. *Handbook of Research Methods for Studying Daily Life* (The Guilford Press, New York), pp 43–61.

Hempel, C. G. (1965). *Aspects of scientific explanation and other essays in the philosophy of science*. New York, NY: The Free Press.

Herrnstein, R. J., & Murray, C. (1994). *The Bell Curve: Intelligence and class structure in American life.* New York: Free Press.

Lakatos, I. (1978). *The methodology of scientific research programmes: Vol. 1. Philosophical papers*. Cambridge, UK: Cambridge University Press.

Lauer, M. S., & Roychowdhury, D. (2021). Inequalities in the distribution of National Institutes of Health research project grant funding. *eLife*, 10:e71712. https://doi.org/10.7554/eLife.71712

Lavie, N. (1995). Perceptual load as a necessary condition for selective attention. *Journal of Experimental Psychology: Human Perception and Performance, 21(3)*, 451-468.

Li, H., Trafimow, D., Wang, T., Wang, C., & Hu, L. (2020). User-friendly computer programs so econometricians can run the a priori procedure. *Frontiers in Management and Business*. *1*(1), 2-6. doi: 10.25082/FMB.2020.01.002

Meehl, P. E. (1990). Appraising and amending theories: The strategy of Lakatosian defense and two principles that warrant using it. *Psychological Inquiry*, *1*(2), 108–141. download;jsessionid=E2E98557E78260208125135C1C3CFF4F (psu.edu)

Molenaar, P. C. M. (2004) A manifesto on psychology as idiographic science: Bringing the person back into scientific psychology, this time forever. *Measurement: Interdisciplinary Research and Perspectives*, *2*(4), 201-218, doi: 10.1207/s15366359mea0204_1

Molenaar P. C. M. (2008). On the implications of the classical ergodic theorems: Analysis of developmental processes has to focus on intra-individual variation. *Developmental Psychobiology*, *50*(1), 60–69. https://doi.org/10.1002/dev.20262

Odedina, F. T., Stern, M. C. (2021). Role of funders in addressing the continued lack of diversity in science and medicine. *Nature Medicine*, *27,* 1859–1861. https://doi.org/10.1038/s41591-021-01555-8

Popper, K.R. (1963). *Conjectures and refutations*. London: Routledge.

Popper, K.R. (1972). *Objective knowledge*. Oxford, UK: Oxford University Press.

Quine, W. V. O. (1952). *The dogmas of empiricism*. Reprinted from "A logical point of view," Cambridge, MA: Harvard University Press.

Taffe, M. A., Gilpin, N. W. (2021). Racial inequity in grant funding from the US National Institutes of Health. *Elife*, 10:e65697. doi: 10.7554/eLife.65697

Trafimow, D. (2009). The theory of reasoned action: A case study of falsification in psychology. *Theory & Psychology, 19*(4), 501–518. doi: 10.1177/0959354309336319

Trafimow, D. (2012). The role of auxiliary assumptions for the validity of manipulations and measures. *Theory & Psychology*, *22*(4), 486-498. doi: 10.1177/0959354311429996

Trafimow, D. (2023). A new way to think about internal and external validity. *Perspectives on Psychological Science*.

Trafimow, D., & Finlay, K. A. (1996). The importance of subjective norms for a minority of people: Between-subjects and within-subjects analyses. *Personality and Social Psychology Bulletin, 22,* 820-828. doi: 10.1177/0146167296228005

Trafimow, D., Kiekel, P. A., & Clason, D. (2004). The simultaneous consideration of between-participants and within-participants analyses in research on predictors of behaviors: The issue of dependence. *European Journal of Social Psychology, 34,* 703-711. doi: 10.1002/ejsp.225

Trafimow, D., & Rice, S. (2008). Potential performance theory (PPT): A general theory of task performance applied to morality. *Psychological Review, 115*(2), 447-462. doi: 10.1037/0033-295X.115.2.447

Whewell, W. (1840). *Philosophy of the inductive sciences*. London, UK: Parker.