



ARTICLE

Hasty Generalizations Are Pervasive in Experimental Philosophy: A Systematic Analysis

Uwe Peters^{1,2*}  and Olivier Lemeire³ 

¹Leverhulme Centre for the Future of Intelligence, University of Cambridge, Cambridge, UK, ²Center for Science and Thought, University of Bonn, Bonn, Germany and ³Centre for Logic and Philosophy of Science, Katholieke Universiteit Leuven, Leuven, Belgium

Corresponding author: Uwe Peters; Email: u.peters@uu.nl

(Received 29 April 2023; revised 07 July 2023; accepted 10 August 2023; first published online 23 August 2023)

Abstract

Scientists may sometimes generalize from their samples to broader populations when they have not yet sufficiently supported this generalization. Do such hasty generalizations also occur in experimental philosophy? To check, we analyzed 171 experimental philosophy studies published between 2017 and 2023. We found that most studies tested only Western populations but generalized beyond them without justification. There was also no evidence that studies with broader conclusions had larger, more diverse samples, but they nonetheless had higher citation impact. Our analyses reveal important methodological limitations of many experimental philosophy studies and suggest that philosophical training may not protect against hasty generalizations.

1. Introduction

The tendency for hasty and unfounded generalization seems to be hardwired into the human brain. (Fogelin 2005, 7)

When conducting experiments, scientists aim to generalize from their study samples to larger populations (Little 1993). Such inductive inferences are an important way of maximizing scientific knowledge. However, although these generalizations are often adequately supported by the data from the sample, they may go wrong. Scientists may generalize their results to a larger population when their sample is too small, when it is not sufficiently representative, or when they have not considered whether the sample and target population are in relevant respects similar so as to warrant the generalization. This inferential error, whereby scientists go beyond or jump ahead of the evidential support, has been called a “hasty generalization” (Hurley 1997).

*Main author.

Hasty generalizations may be common in science. Studies found that in psychology, generalizations of results from Western, educated, industrialized, rich, and democratic (WEIRD) samples to non-WEIRD populations were pervasive but not justified by the researchers (e.g., via citing relevant demographic homogeneity between the populations) (Rad, Martingano, and Ginges 2018; DeJesus et al. 2019). To what extent are *philosophers* in their research susceptible to such generalizations when they conduct empirical studies?

The question is topical. In experimental philosophy (x-phi), philosophers routinely draw inductive inferences based on studies of people's intuitions (Knobe and Nichols 2017). Moreover, it has been argued that these x-phi studies are often limited in their generalizability because they sample mostly only people from WEIRD populations (Machery 2023). If there were no demographic (including cultural) variations in philosophical intuitions, this focus on WEIRD populations may not threaten the generalizability of study results. However, the extent of such invariance remains a matter of debate. Some analyses found significant demographic differences in judgments about philosophical cases (Stich and Machery 2023). Others report that many x-phi findings are cross-culturally robust (Knobe 2021). It is currently unclear whether or when experimental philosophers can expect convergence or variance in philosophical intuitions across populations (Machery 2023). If they neglect demographic variance or invariance, experimental philosophers may produce hasty generalizations in their studies.

That said, there is also the “widespread belief that majoring in philosophy is a superior way for a student to develop critical thinking skills” (Weinberg 2015). It might therefore be that experimental philosophers are especially apt at avoiding hasty generalizations. Indeed, research found that many x-phi studies were more replicable (Cova et al. 2021), contained fewer statistical reporting inconsistencies (Colombo et al. 2018), and were less affected by common questionable research practices (e.g., *p*-hacking) than psychology studies (Stuart, Colaço, and Machery 2019). Consequently, it has been suggested that experimental philosophers may be “more sensitive to certain methodological questions, such as what counts as strong evidence for a given claim” (Cova et al. 2021, 31). If so, they may be immune to hasty generalizations. This could have significant implications beyond experimental philosophy by raising the possibility that philosophical training may help tackle the recently reported pervasive hasty generalizations among behavioral scientists (DeJesus et al. 2019; Peters, Krauss, and Braganza 2022). Examining whether experimental philosophers are indeed immune to such generalizations is also important because hasty extrapolations across populations may obscure philosophically relevant variations between people. This can undermine scientific efforts to explore how demographically different populations respond to philosophical cases.

However, to date, no systematic analysis has been conducted to investigate and provide evidence on how broadly experimental philosophers extrapolate their findings. To change this, we analyzed an extensive corpus of x-phi articles ($N = 171$) published between 2017 and 2023 in eight leading journals publishing x-phi. We found that most articles tested only WEIRD populations but generalized the study results much beyond them without justifying their broad generalization scope. There was also no evidence of a correlation between broader conclusions and larger, more diverse samples. Furthermore, x-phi studies with broader conclusions had higher

citation impact despite being less evidentially supported than studies with narrower conclusions. These findings suggest that hasty generalizations are pervasive in many x -phi studies and that there are significant methodological weaknesses in sampling, extrapolating, and reporting practices in these studies.

To clarify, although we provide evidence that hasty generalizations are common in many x -phi studies, we do not mean to criticize any particular philosopher for them. Rather, our aim is to be constructive and raise awareness among experimental philosophers of a common tendency to overly broadly generalize results so that this tendency can be better controlled moving forward.

We begin by further specifying hasty generalizations. We then introduce our corpus-analysis methodology, present our analysis results, discuss the implications, and rebut some potential objections.

2. Background: Generalizations versus hasty generalizations

When scientists conduct tests on whether people have a certain feature, they commonly cannot test the whole population (e.g., due to resource limitations) but select a sample of it and then generalize. To ensure external validity (i.e., that study results hold across people, stimuli, times, etc.), scientists need to select a sample that is representative of the target population. To do so, they frequently use random sampling from a target population to give individuals from different groups an equal chance of being selected, thus (ideally) creating a subset of the population that accurately reflects the larger group's characteristics (age, gender, etc.) (Andrade 2018).

However, given the range of individual differences (hair color, toe size, etc.), samples are never 100 percent representative (Rothman, Gallacher, and Hatch 2013). Researchers thus need to distinguish relevant from irrelevant differences when selecting their sample and making generalizations (Reichenbach 1951). Generalizations that (implicitly or explicitly) overlook group differences in, say, hair color can be unproblematic, as these are typically irrelevant variations. Which differences are relevant for a given study outcome and generalization may depend on many factors, including on what is being tested (e.g., when testing the usability of a new smartphone, the results may apply only to a particular age group). Researchers therefore need to reflect on variations between individuals, test material, and other study aspects, because when relevant differences are overlooked and generalized over, these generalizations are too quick, that is, hasty, because factors influencing the study's external validity are then ignored.

In some cases, the relevant differences might not be known before conducting a study, making it challenging to factor them into the sampling. Sometimes it might also be unclear how to distinguish outcome-relevant from irrelevant differences. The boundary between warranted and hasty generalizations is thus not always clear-cut (Walton 1999). However, this does not mean that the two cannot be distinguished. Scientists have methods available to select samples that, by the shared epistemic standards within a given scientific field, count as sufficiently large (e.g., use power analyses) and sufficiently representative (e.g., use stratified randomization) for a given generalization. Relatedly, in the behavioral sciences, it is frequently criticized that while many studies have only WEIRD samples, they often generalize their results

to all humans (Thalmayer, Toscanelli, and Arnett 2021), even though it has been shown that, in a wide range of behavior and cognition, WEIRD people are outliers compared to the rest of humanity (Henrich, Heine, and Norenzayan 2010). Such generalizations are thus widely regarded as hasty (DeJesus et al. 2019; Peters, Krauss, and Braganza 2022), and to counteract them, some science journals require researchers to specify the main features of their samples that may limit study result generalizability (Appelbaum 2018). Hence, unless scientists provide evidence in their articles that they have considered and justified the extent to which their samples are relevantly similar to the target populations, extrapolations from their samples to those populations are typically viewed as too quick.

There are, then, different ways of checking whether a scientific generalization is hasty. To assess whether a study's sample is too small or not representative enough, we might examine whether its size is based on a power analysis or whether randomized sampling occurred, respectively. Alternatively, we may consult the discussion and limitations section of the article to see if the researchers considered whether their sample and target population were relevantly similar, justified any similarity assumption, or accounted for potential variation effects. Furthermore, broader conclusions generally require larger, more representative samples to be adequately supported (Asiamah, Mensah, and Oteng-Abayie 2017).¹ To check for hasty generalizations in a field, we may thus examine whether broader generalizations in that field (e.g., about people as such) correlate with larger, more diverse samples than more restricted generalizations (e.g., about people in a given country). If no evidence of such correlation emerges and the researchers have not considered potentially relevant individual differences, there is reason to suspect hasty generalizations.

While such generalizations have been detected in fields like psychology (DeJesus et al. 2019) and artificial intelligence (Peters and Carman 2023), it might be that due to their training, philosophers are less prone to them. If so, then philosophical training may help make scientific inferences less vulnerable to these errors. So, are experimental philosophers immune to hasty generalizations?

3. A systematic analysis of x-phi research

To investigate to what extent (if at all) hasty generalizations can be found in x-phi, we first divided this question into the following five, more specific research questions (RQs) related to sampling, extrapolation, and study impact:

- RQ1.** Do experimental philosophers predominantly sample only WEIRD populations?
- RQ2.** Do they restrict their study conclusions to their samples and study populations (i.e., the subsets of the target population² that are available for study and from which the samples are drawn) or extrapolate beyond them?

¹ We are setting aside Bayesian approaches.

² The target population is the large set of individuals in the world to which researchers may wish to generalize their study results (e.g., all philosophers). The study population is the part of the target population from which researchers can (depending on availability, resources, etc.) recruit for a study (e.g., US philosophers). The study sample comprises the individuals selected from the study population.

- RQ3.** Do experimental philosophers in their articles consider whether their samples and the populations to which results are generalized are in relevant respects similar to warrant the generalization?
- RQ4.** Do articles with broader conclusions have larger or more diverse samples, and are these conclusions correlated with larger or more diverse samples?
- RQ5.** Is the scope of experimental philosophers' conclusions related to the impact of their study such that broader conclusions correlate with higher impact?

3.1 Methodology

To answer these questions and avoid selection bias, we conducted a systematic literature review of x-phi articles.³ To identify articles for review, we focused on philosophical journals (i.e., journals with philosophers as editors or “philosophy” in their “aims and scope”), because even though x-phi studies also appear in some psychology and cognitive science journals, we were interested in how philosophers would generalize their study results in articles peer reviewed by other philosophers. X-phi articles accepted by psychology and cognitive science journals will have undergone peer review by psychologists (e.g., journal editors). This can affect philosophers' reporting practices in these articles, blurring our insight into how they would generalize their results independently of psychologists' evaluations.

Because scientific databases (e.g., Scopus) do not monitor all relevant philosophy journals, we adopted an approach by Polonioli et al. (2021), who combined the quantitative ranking provided by the h-index with established informal polls, such as the *Leiter Report* journal ranking, to form a list of twenty journals frequently publishing x-phi. From these journals, we focused on those that Polonioli et al. found to have published four or more x-phi articles over three years. This resulted in eight journals: *Philosophical Psychology*, *Review of Philosophy and Psychology*, *Synthese*, *Mind and Language*,⁴ *Philosophical Studies*, *Nous*, *Philosophy and Phenomenological Research*, and the *Journal of Consciousness Studies*.

3.1.1 Selection criteria

From the eight journals, we included any article published between January 2017 and January 2023 (including online-first articles) with at least one quantitative x-phi study and at least one philosopher as author or coauthor. We focused on quantitative studies because we were interested in generalizations and qualitative studies often aim not to produce generalizations but to provide detailed insights into personal experiences (Polit and Beck 2010). We focused on articles with at least one philosopher as author because we wanted to examine philosophers' generalizations, and although nonphilosophers among an article's authors may also produce study generalizations, research ethics guidelines specify that every author is responsible for all content of a jointly written article (Wager and Kleinert 2011). We excluded articles

³ We used a protocol adapted from Peters and Carman (2023).

⁴ *Mind and Language* is interdisciplinary but mentions philosophy in its scope description, and the editors are predominantly philosophers.

that covered only simulations, modeling, corpus analyses, or replications. Using these criteria, we (two researchers) independently read the titles and abstracts of all articles published in the specified journals and time. 171 articles met the criteria and were selected for full-text analysis.

3.1.2 Data extraction

We extracted journal name, article title, and publication year to collect data on an article's impact. Following others, we operationalized impact as Google Scholar citation count (Li and Zhu 2023). We also extracted final sample size and participants' country or region (e.g., Europe). Based on the participants' country or region, we coded an article as "WEIRD," "non-WEIRD," or "mixed" using the WEIRD/non-WEIRD categorizations proposed by Klein et al. (2018). We additionally coded articles (yes/no) on whether they compared different demographic (cultural, gender, expertise, etc.) groups and reported findings of demographic variance or invariance in philosophical judgments. Relatedly, we coded articles (yes/no) on whether the authors considered if their samples and the population(s) to which results were generalized were relevantly similar or whether variations (e.g., in demographics or stimuli) might limit generalizability.

Finally, we extracted information on an article's scope of conclusion. Researchers may use qualifiers or past tense to indicate that their findings are specific to the sample, their study population (e.g., US philosophers), or a particular context, time, or culture. Articles containing only result claims with such specifying features, minority quantifiers (e.g., "many laypeople"), or hedging terms ("may," "to some extent," etc.) in the abstract, results, discussion, or conclusion sections were coded as "restricted." Alternatively, in these sections, researchers may make claims that are not scope limited in these ways but that instead suggest that the study results apply beyond the study population to people, philosophers, and so on in general, concern majorities of them ("most philosophers"), or hold across all contexts, times, or cultures (e.g., by describing findings as pertaining to folk psychology as such). Articles with at least one such broad result claim were coded as "unrestricted" (for examples, see figure 1). We also applied this label when an article contained some restricted claims in addition to unrestricted ones, because articles usually undergo many revisions when authors can qualify their broader claims. If that does not happen, there is reason to believe that the authors consider their broader generalizations warranted, making the "unrestricted" label apt. Within this category of claims, we further coded for *generics*, that is, generalizing sentences with a noun phrase that refers without a quantifier and describes the members of a kind as such (e.g., "Ks do F ," "a K tends to F ," or "Ks generally reason like F " vs. "most Ks do F " or "66 percent of Ks tend to F ") (Krifka et al. 1995).

After coding the data, we calculated the interrater agreement between our classifications (Cohen's kappa). It was consistently between substantial and almost perfect ($\kappa = 0.72$, 95 percent CI [0.61, 0.83] to $\kappa = 0.85$, 95 percent CI [0.77, 0.93]) (Landis and Koch 1977). We additionally asked two project-naive researchers to independently classify a random 25 percent of the data for the scope of the conclusion variable (our most complex variable) using our predefined instructions. Agreement between their and our ratings was $\kappa = 0.74$, 95 percent CI [0.50, 0.98] and $\kappa = 0.81$, 95 percent CI [0.60, 1.06], respectively. Disagreements were resolved by discussion. If

needed, the ratings were updated before the data were analyzed ($\alpha = 0.05$). All our materials and data are accessible on an Open Science Framework (OSF) platform.⁵

3.2 Results

From our final sample ($N = 171$), most x-phi articles (71.9 percent, $n = 123$) were published between 2020 and 2023 (table A1). The highest proportion appeared in *Philosophical Psychology* (31.5 percent, $n = 54$), followed by *Review of Philosophy and Psychology* (24.6 percent, $n = 42$), *Synthese* (21.6 percent, $n = 37$), *Mind and Language* (9.4 percent, $n = 16$), and *Philosophical Studies* (7.6 percent, $n = 13$). *Nous, Philosophy and Phenomenological Research*, and the *Journal of Consciousness Studies* had the lowest numbers ($n = 2-4$).

3.2.1 RQ1

Do experimental philosophers predominantly sample only WEIRD populations?

A significant proportion of articles (30.4 percent, $n = 52$) did not report any specific details on participants' country or region, precluding a WEIRD/non-WEIRD categorization. Across the remaining 119 articles, study participants came from fifty-four countries or regions. The three most frequent ones were the United States (69.7 percent, $n = 83$), the United Kingdom (19.3 percent, $n = 23$), and Germany (7.6 percent, $n = 9$). Importantly, 82.4 percent ($n = 98$) of the 119 articles contained studies that sampled only WEIRD populations. 10 percent ($n = 12$) sampled mixed populations, and 7.6 percent ($n = 9$) sampled only non-WEIRD populations.

3.2.2 RQ2

Do experimental philosophers restrict their study conclusions to their samples and study populations or extrapolate beyond them?

Researchers can limit the scope of their conclusions by using past tense (e.g., “we found that laypeople judged”) or quantifiers (e.g., “many US philosophers believe”) or by referring to study participants only (e.g., “respondents thought”). However, of the 171 reviewed articles, 69.6 percent ($n = 119$) contained at least one unrestricted claim,⁶ that is, a conclusion that extended results beyond the study population to people (e.g., philosophers) as such, to majorities of them, to folk psychology, to the human mind in general, or across culture and time. Figure 1 provides ten examples. Moreover, in the 119 unrestricted articles, we found a total of 646 unrestricted claims, of which 94.7 percent ($n = 612$) were generics, that is, claims that did not describe particular individuals but concerned the members of a kind as such.⁷

Broad conclusions of the kind outlined in figure 1 may be justified;⁸ they are not necessarily hasty generalizations. Experimental philosophers may have considered relevant auxiliary assumptions about their samples' demographic features before

⁵ <https://osf.io/xfdb7/>.

⁶ Because we counted claims with hedging modals like “may” as restricted even when they contained generics, if anything, our results may underestimate the pervasiveness of overly broad claims in x-phi studies.

⁷ Interrater agreement was $\kappa = 0.81$, 95 percent CI [0.74, 0.87].

⁸ Unrestricted claims sometimes included sentences of the form “We provide evidence that people do X.” Although these claims have different truth conditions than claims like “We believe that people do X” or “People do X,” we grouped them together as unrestricted conclusions because all three types of claims contain broad conclusions about people as such.

-
1. “The results of our experiments clearly show that a large majority of people distinguish hard cases from easy ones and reveal response patterns that are predicted by the philosophical consensus.” (6)
 2. “Overall, the results clearly demonstrate that folk psychology views belief as voluntary.” (9)
 3. “Our first experiment shows that physicists are reliable when making judgements in thought experiments in physics.” (13)
 4. “Our results also provide evidence that people take the passage of time to be a function of subjective experiences.” (35)
 5. “Our findings suggest philosophers are better at deploying concepts than laypeople but are susceptible to the linguistic salience bias to a similar extent and at similar points.” (42)
 6. “Instead, we find that laypeople are willing to count both a multiply realized property and its realizers as causes.” (60)
 7. “We show that ordinary people think that morality is important for psychological continuity and that this judgment is related to subsequent perceptions of moral duties.” (101)
 8. “Our results show that people believe that science (abstractly) is, and scientific statements (concretely) are, a matter of objectivity.” (147)
 9. “A first lesson we can draw from our results is simply that people do not conflate a meaningful life with a happy life.” (153)
 10. “The results presented in this paper show that people report anger, sadness, and fear in the absence of bodily feelings.” (161)
-

Figure 1. Examples of unrestricted conclusions found in X-phi articles. The number in brackets indicates the number of the article on our OSF spreadsheet (<https://osf.io/xfdb7/>).

extrapolating in these broad ways. However, if the authors of these claims arrived at their generalizations after reflecting on and discounting potential demographic or other variation that might limit generalizability, then their articles should contain signs of such reflection. For in leaving potential assumptions of demographic invariance implicit and in not supporting the view that generalizability concerns about individual variation can be set aside, key parts of a full justification of these broad conclusions would be missing, and the objection that the authors generalized hastily can gain traction.

3.2.3 RQ3

Do experimental philosophers in their articles consider whether their samples and the populations to which results are generalized are in relevant respects similar to warrant the generalization?

In 60.8 percent ($n = 104$) of all articles, philosophers did not do so; that is, there was no reflection on the appropriateness of generalizing from the sample beyond the study population to a broader group. Yet, of these articles, 74 percent ($n = 77$) nevertheless contained unrestricted conclusions, that is, claims whose scopes extended to people, philosophers, and so on in general. Moreover, in a phi-coefficient test, we could not find any evidence of a correlation between articles with indications of reflection on potential generalizability concerns related to individual variation and articles drawing restricted versus unrestricted conclusions.⁹ Yet, if broad conclusions beyond study populations were based on researchers' reflection, one would expect such evidence.

⁹ $\Phi = -0.120, p = 0.115$.

Granted, even if articles did not contain considerations on relevant similarities between samples and the populations to which the authors generalized, the authors might still have carefully reflected on the matter. However, because making the basis for one's generalizations explicit is important to fully support them, offering the relevant reflection in the articles would have increased the articles' methodological quality. It is thus not clear why, if they did reflect on potential generalization-limiting factors, the authors did not mention such considerations. Only in 39.2 percent ($n = 67$) of all articles did this happen. Intriguingly, however, 62.7 percent ($n = 42$) of these articles still contained unrestricted conclusions. That is, in these articles, philosophers noted factors that would limit the generalizability of their studies but nonetheless extrapolated beyond their study populations without justifying this broad extrapolation. To illustrate the point without singling out particular researchers, following is one anonymized example:

We only collected data from an American sample, so we can't generalize based on our findings Despite these limitations, we think we have advanced the debate concerning natural compatibilism by providing new evidence that people find free will and responsibility to be incompatible with determinism. (102, 991)¹⁰

If a study samples only Americans and finds that they think that p , this will provide evidence for the claim that *some* people think that p (Americans). However, it does not also provide sufficient evidence for the claim that people in general think this. It might be that in all study-relevant respects, Americans and the rest of the world are similar, warranting the broader claim. However, the authors do not support this extrapolation in their article but only acknowledge that the data came from an American sample.

One might argue that while evidence that Americans think that p does not justify concluding that people in general think that p , it nonetheless incrementally supports the hypothesis that people think that p in the following sense: this broad hypothesis predicts that Americans (among others) think that p , and the evidence confirms this prediction; therefore the claim "we provide evidence that people think that p " is no longer entirely unsupported if the authors provide evidence that Americans think that p .

However, even if the generalized claim is incrementally supported in that way, this incremental support is insufficient to make the broad scope of the generalization adequate. To see this, consider biologists who study the color of a population of ravens, finding them to be black and claiming to thereby have provided support for holding that "birds are black." Clearly, without justification that their sample of ravens is representative in color of birds in general, this generalization would be viewed as hasty, even though the biologists' findings do support "birds are black" in the incremental sense outlined earlier. Hence this kind of incremental support is not enough to make such generalized conclusions adequate in scope. The point equally applies to the example of the preceding quotation and suggests that the extrapolation from Americans to people in general remains hasty even if the notion of incremental

¹⁰ The first number in parentheses indicates the number of the paper on our OSF spreadsheet (<https://osf.io/xfdb7/>); the second number is the page number.

support is invoked. Relatedly, just as the biologists would have adequately limited their generalization by concluding that “most ravens are black,” the authors of the study on Americans could have more adequately limited their generalization by stating that “we provide evidence that *Americans* think that p ” or “we provide evidence that *some* people think that p .”

3.2.4 RQ4

Do articles with broader conclusions have larger or more diverse samples, and are these conclusions correlated with larger or more diverse samples?

Even if experimental philosophers did not explicitly justify that their samples and the populations to which they generalized are relevantly similar, their broad conclusions might still be warranted. If so, one would expect χ -phi articles with unrestricted conclusions (which refer to people beyond the population sampled) to have larger, more diverse samples than articles with restricted conclusions (which refer to study participants or specific study populations). For, in classical statistics, broader conclusions generally require larger, more representative samples to be adequately supported (Asiamah, Mensah, and Oteng-Abayie 2017). We therefore first analyzed the sample sizes within both groups of articles. Figure 2 presents the total distribution of sample sizes by group. It shows that a greater number of unrestricted articles had smaller samples compared to the restricted articles. The largest samples were in fact in articles with restricted conclusions. This is the opposite of what one would expect if broader generalizations were aligned with larger samples.

To test statistically for sample size differences between both groups of articles, we treated conclusion scope (unrestricted vs. restricted) as a binary variable and a study’s sample size (final participant n) as a scale variable and conducted a Mann-Whitney U -test (data normality was violated). We did not find evidence of a significant difference in sample size between unrestricted and restricted articles.¹¹ In a rank-biserial correlation test, we also found no evidence of any significant link between articles with unrestricted conclusions and larger samples.¹²

However, even if they do not have larger samples, unrestricted articles might still have more culturally diverse samples, potentially warranting broader claims. To assess this, we related the scope of the conclusion variable to the sample country/region variable. Excluding the articles that did not report specific details about their samples’ country/region ($n = 52$), and focusing only on the remaining *unrestricted* articles ($n = 79$), we found that 91.1 percent ($n = 72$) of them had only either WEIRD ($n = 66$) or non-WEIRD ($n = 6$) samples. That is, of all the unrestricted articles with country/region details, only 8.9% ($n = 7$) had mixed (WEIRD and non-WEIRD) individuals in their samples. These findings suggest that articles with broader conclusions did not have more diverse, more widely representative samples that could potentially support broader claims.

We also statistically examined whether unrestricted articles had more diverse samples. Using a Mann-Whitney U -test (data normality was again violated), we found no evidence of a significant difference in mean ranks on the number of countries/

¹¹ Unrestricted articles, $n = 119$, mean rank = 84.11, vs. restricted articles, $n = 52$, mean rank = 90.34; $U = 2,868.50$, $p = 0.449$.

¹² $r_{rb}(169) = -0.06$, 95 percent CI $[-0.211, 0.097]$, $p = 0.451$.

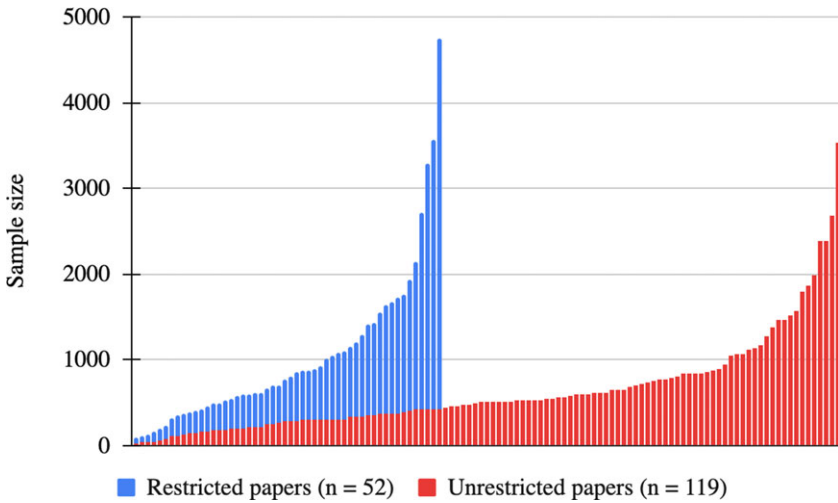


Figure 2. Full distribution of sample sizes within each group of x-phi articles.

regions mentioned in unrestricted versus restricted articles.¹³ We also did not find evidence of a statistically significant correlation between scope of conclusion and number of countries/regions.¹⁴ In sum, we found no evidence that unrestricted versus restricted articles had or were correlated with larger or more diverse samples.¹⁵

3.2.5 RQ5

Is the scope of experimental philosophers' conclusions related to the impact of their studies such that broader conclusions correlate with higher impact?

Using citation count as a proxy for impact, time effects on publishing may confound the results, as older articles will have had more time to accrue citations than newer ones. Citation count data thus need to be normalized. One previously used normalization method (Li and Zhu 2023) is calculating the *relative citation rate* (RCR):

$$\text{RCR} = \frac{\text{Observed citation count (OCC)}}{\text{Expected citation count (ECC)}}.$$

OCC represents a given article's raw total citations. ECC captures an article's expected citations in the year it was published. For instance, in our sample, forty-four articles were published in 2021, receiving 303 total citations until data collection. Therefore, the ECC for any article published in 2021 is 6.9. If an article published in 2021 has been cited fourteen times so far, its RCR will be 2. By controlling for the number of years an article has been published, this normalized citation rate allows for comparing an article's impact across the time span we investigated.

¹³ Number of countries/regions mentioned in unrestricted articles, $n = 119$, mean rank = 83.15, vs. restricted articles, $n = 52$, mean rank = 92.52; $U = 2,755.00$, $p = 0.212$.

¹⁴ $r_{\text{th}}(169) = -0.10$, 95 percent CI $[-0.247, 0.060]$, $p = 0.213$.

¹⁵ These two trends remained when we excluded articles without information on country/region.

There is another challenge, however. Many articles are published online much earlier than in print. We therefore recorded each article's online publication date and used it for analysis. Whereas our print publication years began in 2017, the online publication dates ranged from 2014 to 2023. After calculating the RCR for each article falling within that period, we conducted a Mann-Whitney U -test (data normality was not met) to examine whether there was a mean rank difference in citations between unrestricted and restricted articles. Overall, unrestricted articles had a significantly higher citation count ($n = 119$, mean rank = 91.68) than restricted articles ($n = 52$, mean rank = 73.00; $U = 2,418.00$, $z = -2.27$, $p = 0.023$). A subsequent rank-biserial correlation test revealed a (weak) positive correlation between articles with unrestricted conclusions and a higher citation count ($r_{rb}[169] = 0.174$, 95 percent CI [0.021, 0.320], $p = 0.022$).

To enrich our analysis, we also related the numbers of unrestricted claims and generics in each article to the impact variable, calculating Spearman's rho. The positive correlation increased both in strength and statistical significance for the impact and the number of unrestricted claims variables ($r_s[169] = 0.224$, 95 percent CI [0.072, 0.365], $p = 0.003$) and (even more so) for the impact and the number of generics variables ($r_s[169] = 0.231$, 95 percent CI [0.080, 0.372], $p = 0.002$).

3.2.6 Comparative studies

The results thus far suggest that in many of the reviewed articles, philosophers generalized their results from WEIRD samples to non-WEIRD populations without testing the latter, comparing both, or otherwise justifying these broad extrapolations. In fact, of the 171 articles we examined, only 20.5 percent ($n = 35$) reported studies that compared demographic (cultural, gender, etc.) groups on philosophically relevant judgments. From these studies, 57.1 percent ($n = 20$) found differences in such judgments, 37.1 percent ($n = 13$) found invariance across demographic groups, and 5.7 percent ($n = 2$) mentioned both kinds of results for different factors. While we have until now focused on generalizations in which relevant demographic (e.g., WEIRD vs. non-WEIRD) variations may be overlooked, hasty generalizations may also occur when researchers overlook relevant demographic *invariance*. For instance, based on their studies, philosophers may conclude that ethicists believe X whereas laypeople do not do so, thus postulating a variation between these groups even when their sample is too small, when it is not representative enough, or when relevant similarities between the samples and target populations were not considered.

To explore whether this happened too, we focused on the thirty-three articles that reported either demographic variance or invariance. We first examined whether one of the two groups of articles contained significantly more unrestricted articles. There was no evidence that unrestricted articles were less common in either group.¹⁶ We then analyzed the articles reporting demographic variance ($n = 20$) and found that 40 percent ($n = 8$) did not contain any indication that the authors considered whether their samples and the population to which they generalized were in relevant respects similar, suggesting that hasty generalizations occurred. Unrestricted articles had larger samples ($n = 5$, mean rank = 15.40) than restricted articles ($n = 15$, mean

¹⁶ $\chi^2(1, N = 33) = 1.587, p = 0.208$.

rank = 8.87; $U = 13.00$, $z = -2.139$, $p = 0.032$), but there was no evidence that they had more diverse samples ($p = 0.501$).

4. General discussion

Our analyses provide novel insights into methodological limitations of many x-phi studies. To make them explicit, we revisit our five main findings.

4.1 *Missing information on participants' demographic backgrounds*

More than 30 percent of the x-phi articles we reviewed did not mention information about their samples' country or region, precluding an evaluation of the cross-cultural generalizability of results. This proportion is higher than that observed in psychology, where research on recent articles found that approximately 11 percent lacked such information (Rad, Martingano, and Ginges 2018).

4.2 *WEIRD sampling*

Although it is well known that behavioral scientists sample mainly only WEIRD populations, our study provides the first large-scale quantitative evidence of this phenomenon in x-phi research published in philosophical journals. Eighty-two percent of the reviewed articles (with relevant information) sampled only WEIRD populations. This number is high. But the problem might be worse in psychology, where WEIRD sampling was found in 94 percent of studies (Rad, Martingano, and Ginges 2018). There is an ongoing controversy about the extent to which philosophical judgments are affected by demographic variation. However, at present, we cannot assume that non-WEIRD populations share all the same intuitions, thoughts, and responses as do WEIRD people, who constitute only 12 percent of the world population (Henrich, Heine, and Norenzayan 2010). If experimental philosophers aim to discover philosophically relevant features of human cognition in general, then, given the low number of comparative studies we found, most of the reviewed articles may only scratch the surface of the matter.

4.3 *Sensitivity to individual variation, unrestricted claims, and generics*

Although philosophers primarily sampling from WEIRD populations might not necessarily be problematic if they adequately justify or limit their subsequent generalizations, our results indicate that in most (61 percent) of the 171 reviewed x-phi articles, the authors did not even consider that the generalizability of their results could be affected by individual variation between people. There was no justification in these articles for extrapolating from the samples (commonly WEIRD groups) beyond the study populations to larger (commonly WEIRD and non-WEIRD) groups. Yet, in the vast majority of cases (74 percent), researchers still did so by producing unrestricted claims. Broader claims may be supported if larger, more diverse samples are tested. But most (91 percent) of the unrestricted articles (with country/region details) had sampled only either WEIRD or non-WEIRD individuals. This suggests that these articles did not have sufficient support for their broad conclusions because supporting conclusions that pertain to people, folk psychology, and so on in general requires testing both WEIRD and non-WEIRD populations or providing auxiliary assumptions about demographic

invariance. This did not happen in these articles. We also could not find any evidence that broader conclusions were correlated with larger, more diverse samples, and there was no difference in these respects between articles reporting findings of demographic variance and articles reporting findings of demographic invariance. That is, we found that hasty generalizations were common on both sides of the current debate in x -phi on demographic variance in philosophical intuitions.

To be sure, such generalizations have also been found in psychology (DeJesus et al. 2019) and artificial intelligence articles (Peters and Carman 2023). But their prevalence in experimental philosophy is remarkable given that philosophers are trained in logic, including inductive and informal logic (Weinberg 2015), and are thought to be “more sensitive to . . . what counts as strong evidence for a given claim” (Cova et al. 2021, 31).

However, generalizing sentences without a quantifier in the noun phrase, that is, *generics* (e.g., “introverts like X ,” “people think that p ”), appear to be significantly more common in scientific generalizations in, for instance, psychology, where some studies found them in 89 percent of articles (e.g., DeJesus et al. 2019),¹⁷ than in x -phi studies, where we found them in, overall, no more than 70 percent (i.e., 119/171) of articles. Still, focusing only on the x -phi articles with unrestricted conclusions, almost all (95 percent) of these conclusions were generics.

Using generics in science and x -phi may have benefits. They can convey (1) that a relationship between a kind and a property is robust and can be expected to persist (Ritchie 2019), (2) that a property is characteristic of the kind (Leslie 2007), or (3) that the property is caused by a particular type of mechanism (Vasilyeva and Lombrozo 2020). They may also be more effective than more qualified claims in initiating social change, provoking reflection, and guiding people’s behavior, as they can simplify complex phenomena.¹⁸

Nevertheless, compared to using precisely quantified language about a population, using generics to communicate scientific results may also create significant epistemic problems. Generics about ‘people’, ‘philosophers’, or ‘the folk’ gloss over differences between the members of these categories, which may encourage researchers toward overgeneralizations. Moreover, in contrast to precisely quantified generalizations (e.g., “66 percent of K s believe F ”), generics communicate only a vague prevalence level, making them inherently harder to scientifically test and inaccurate when the facts are not vague (Peters 2023). Relatedly, generics allow for exceptions and can be used to convey different levels of a property’s prevalence: whereas “ravens are black” conveys that almost all ravens are black, “mosquitos carry malaria” is true, even though fewer than 10 percent of all mosquitos carry malaria. Many other generics convey property prevalence levels in between (Tessler and Goodman 2019). This can lead to miscommunication, as people need more background information to determine what a generic conveys than to determine what a precisely quantified generalization conveys. The communicative benefits of using generics in science (e.g., conveying more complex content than just how many individuals instantiate a property) may sometimes outweigh the drawbacks. However, given the pervasiveness

¹⁷ However, DeJesus et al. (2019) also included, for instance, sentences like “ K s may do F ” as (hedged) generics. We excluded them, as they strike us as less problematic.

¹⁸ That is why we used a generic in our article’s title.

of generics in the reviewed x-phi articles, caution is warranted about their potentially high epistemic costs in the field.

4.4 Lack of awareness of hasty generalizations

The prevalence of hasty generalizations in the reviewed x-phi articles raises the question whether the philosophers in our sample consciously generalized in these ways. Some of the broad conclusions we encountered may have been chosen deliberately to boost a study's perceived importance and perform better in academic competition and selection.

However, there is reason to believe that unintentional processes also contribute to the phenomenon. This is supported by the fact that most of the articles (63 percent) that indicated awareness that demographic or other variations might threaten the generalizability of their results still contained hasty generalizations. In these articles, researchers acknowledged the limitations of their studies (e.g., having only an American sample) but nonetheless concluded that they had provided evidence for a claim about people as such, without providing further justification. That some philosophers drew these inferences despite noting generalizability limitations suggests that the resulting generalizations may have been unintentional and based on an automatic extrapolation tendency or "generalization bias" that facilitates generalizations even when they are not warranted (Peters, Krauss, and Braganza 2022).

4.5 Impact

The automatic tendency just mentioned may interact with social factors, such as publication impact. We found that unrestricted articles had, on average, higher impact. This is perhaps unsurprising. Broader conclusions attract more attention because they purport to hold for more cases. When broader conclusions are evidentially sufficiently supported, their higher impact is epistemically beneficial, as a key goal of science is exactly to produce warranted generalizations that enable explanations and reliable predictions (Kitcher 1989).

However, as noted, of the x-phi studies with country/region information and with unrestricted conclusions, more than 90 percent were not well supported. This is because these studies only sampled either WEIRD or non-WEIRD populations and offered no evidence of relevant similarity between the two to support extrapolations across them. Yet, overall, the impact of x-phi articles with these conclusions was higher than that of articles with more qualified, narrower conclusions that, by being more restricted, were in these circumstances (e.g., of WEIRD sampling) better aligned with the evidential support. This suggests that, overall, x-phi articles with less evidentially supported conclusions performed better in terms of impact than articles with more supported conclusions. Reliable belief formation in the field may suffer if less evidentially supported claims spread more easily and receive more uptake. Moreover, since researchers need to compete for impact, if hasty generalizations yield higher impact, overly broad claims may accumulate over time in academic outputs, driving a "natural selection of bad science" (Smaldino and McElreath 2016, 2). In these conditions, it can become adaptive for experimental philosophers to

proliferate hasty generalizations and develop precisely the kind of automatic extrapolation tendency mentioned earlier.

5. Objections

Some of our results rely on interpreting statements like “people believe that X,” “folk psychology views Y as F,” or “a large majority of people distinguish Z” (see figure 1) as broad claims whose scopes extend to people, folk psychology, and so on in general. One might object that when read in context, these sentences express generalizations with a restricted scope referring only to study participants or the population sampled (e.g., WEIRD folk). Indeed, one might argue that because almost all (95 percent) of the x-phi conclusions we viewed as unrestricted claims contained generics, and generics allow for (in some cases, numerous) exceptions, our claim that these conclusions are hasty generalizations is itself too quick.

However, there are three reasons to believe that the sentences we classified as unrestricted conclusions did not express narrower claims even when read in their proper contexts. First, recall that during data collection, two researchers independently distinguished articles with unrestricted conclusions from those with only restricted conclusions. Two other researchers who were naive to our project and RQs did the same for 25 percent of the relevant data. Crucially, there was strong interrater agreement on classifications regarding unrestricted versus restricted conclusions among all four researchers (consistently between $\kappa = 0.72$ and 0.85). If the claims we viewed as extrapolations to people, folk psychology, and so on in general were usually interpreted narrowly, this consistent agreement across independent classifiers should not appear.

Second, if the broad generalizations we found in x-phi studies referred narrowly only to WEIRD people, study participants, and so on, one would expect there to be some convention among researchers in the field that these generalizations should be understood as restricted. However, this does not seem to be the case, as in 39.2 percent (67/171) of the articles, philosophers felt the need to include *explicit clarifications* that their results may have limited generalizability (e.g., to WEIRD individuals). If people in the field already assumed that generic conclusions are relativized to WEIRD samples, study participants, and so on, such clarifications would be redundant.

Finally, the broad generalizations with a generic noun phrase that we found either lacked any quantification (e.g., “people believe that X”) or contained only a vague adverbial quantifier (e.g., “the folk generally think X”). Philosophers of language have shown that, unlike generalizations with quantifiers like “every” or “no,” generalizations with a generic noun phrase do not allow for contextual scope restriction (von Fintel 1994). For example, when visiting a zoo and finding that all the lions in the zoo are albinos and hence white, it would be felicitous to claim that “every lion is white” but not that “lions are white” or “lions are generally white.” This is because the scope of the quantifier “every” can be contextually limited to a contextually relevant subset (the lions in this zoo), but the generic or the adverbial quantifier “generally” cannot. Therefore, the generic generalizations we encountered in many x-phi articles also cannot be used to refer only to a specific subset of contextually relevant individuals

(e.g., WEIRD people). Generics do not allow for this type of contextual restriction; they are used to describe a kind as such.

But because generics allow for exceptions, one might insist that the truth of minority generics (e.g., “mosquitos carry malaria”) shows that generics *can* be used to talk about a subset of individuals. We grant that generics like “people believe that X” may be nonuniversal in both scope and prevalence level (e.g., more than 70 percent). However, even then, these kinds of statements in the reviewed x-phi articles would still gloss over variation and purport to extend across the entire human population (including WEIRD and non-WEIRD populations), even though, as noted, most authors did not show that their samples and this much broader population are relevantly similar. Thus, these statements remain hasty generalizations. Moreover, if the authors had intended these broad claims to refer only to WEIRD people or to have a limited scope, it is unclear why they did not use less ambiguous terms to prevent misunderstanding. This was feasible and did happen in more than 30 percent of all articles. There are therefore good grounds to think that statements like “people believe that X” or “the folk generally think X” that we (and two author-independent researchers) interpreted as referring to people, folk psychology, and so on in general did have such a broad scope.

6. Limitations

To avoid hasty generalizations ourselves, we provide three “generality constraint statements” (Simons, Shoda, and Lindsay 2017) concerning our own study. First, we focused only on eight philosophical journals that publish x-phi. However, we followed a systematic selection method and adopted a list of journals that was also adopted by other researchers reviewing x-phi studies, including journals that very frequently publish x-phi. Our sample of articles and their reporting practices should thus be representative of a wide range of current x-phi. Second, we used nationality/region as a proxy for sample diversity and citation count as a metric for impact. These are simplifications that limit generalizability. Future research with more granular operationalizations is desirable. Finally, we collected our data manually, not automatically. However, to mitigate potential human error, we analyzed articles independently, cross-checked the coding, had author-independent raters classify data subsets, and calculated interrater agreement. Moreover, we have carefully documented all unrestricted claims and generics (sample sizes, article details, etc.) that we quantified here on a spreadsheet that is publicly available.¹⁹ Our results can be verified with this data set.

7. Conclusion and recommendations

We started by asking how susceptible experimental philosophers are to hasty generalizations, that is, generalizations from samples to larger populations when the samples are too small, when they are not representative enough, or when the researchers have not justified the assumption that their samples and the larger populations are relevantly similar for extrapolations from one to the other. We divided this single question into five more specific ones and conducted a systematic

¹⁹ <https://osf.io/xfdb7/>.

analysis of x-phi studies to answer them. Our results are the first quantitative evidence that hasty generalizations are widespread in many x-phi articles. Most articles in our sample tested only WEIRD populations but generalized their results beyond them without justifying such generalizations. There was also no evidence that broader conclusions were linked to larger, more diverse samples. Even many philosophers who indicated awareness that individual variations between people may have influenced their studies' generalizability still produced hasty generalizations, suggesting that an unintentionally operating generalization bias may have been involved. Finally, we found that many x-phi studies with broader conclusions also had higher impact, despite being less evidentially supported than studies with more qualified conclusions. Philosophical training may therefore be limited in its efficacy to guard against hasty generalizations and their proliferation.

To tackle them, we recommend that journals that publish x-phi articles ask authors to provide constraints on generality statements in their articles, that is, statements that specify the intended target population, the limits of a given study's generalizability, and the basis for believing that the chosen sample, materials, and procedures are sufficiently representative to support extensions of results from study participants to broader populations. Moreover, philosophers should consider using quantifiers, qualifiers (e.g., "may"), frequencies, or past tense when describing their results. To illustrate different forms of rephrasing, in the appendix (figure A1), we present restricted versions of unrestricted conclusions that we found in our review. We hope our data help draw attention to the methodological problems outlined in this article and encourage experimental philosophers to adopt mitigation strategies to reduce hasty generalizations.

Acknowledgments. This article has greatly benefited from comments by Joshua Knobe, Edouard Machery, Shen-yi Liao, Andreas De Block, Alex Krauss, Jan Sprenger, and anonymous reviewers of this journal. Many thanks also to Mary Carman and Charlotte Gauvry for cross-checking our classifications.

Author Contribution. UP conceived and designed the study, collected the data, did all data analyses, developed the main arguments, and revised, and edited the paper. OL assisted with the data collection, argumentation, revising, and editing of the paper.

References

- Andrade, Chittaranjan. 2018. "Internal, External, and Ecological Validity in Research Design, Conduct, and Evaluation." *Indian Journal of Psychological Medicine* 40 (5):498–99. https://doi.org/10.4103/IJPSYM.IJPSYM_334_18.
- Appelbaum, Mark, Harris Cooper, Rex B. Kline, Evan Mayo-Wilson, Arthur M. Nezu, and Stephen M. Rao. 2018. "Journal Article Reporting Standards for Quantitative Research in Psychology: The APA Publications and Communications Board Task Force Report." *American Psychologist* 73 (1):3–25. <https://doi.org/10.1037/amp0000191>.
- Asiamah, Nestor, H. Kofi Mensah, and E. Fosu Oteng-Abayie. 2017. "Do Larger Samples Really Lead to More Precise Estimates? A Simulation Study." *American Journal of Educational Research* 5 (1):9–17.
- Colombo, Matteo, Georgi Duev, Michèle B. Nuijten, and Jan Sprenger. 2018. "Statistical Reporting Inconsistencies in Experimental Philosophy." *PLoS ONE* 13 (4):e0194360. <https://doi.org/10.1371/journal.pone.0194360>.
- Cova, Florian, Brent Strickland, Angela Abatista, Aurelien Allard, James Andow, Mario Attie, James Beebe et al. 2021. "Estimating the Reproducibility of Experimental Philosophy." *Review of Philosophy and Psychology* 12 (1):45–48. <https://doi.org/10.1007/s13164-018-0407-2>.

- DeJesus, Jasmine M., Maureen A. Callanan, Graciela Solis, and Susan A. Gelman. 2019. "Generic Language in Scientific Communication." *Proceedings of the National Academy of Sciences of the United States of America* 116 (37):18370–77. <https://doi.org/10.1073/pnas.1817706116>.
- Fogelin, Robert J. 2005. "The Logic of Deep Disagreements." *Informal Logic* 25 (1):3–11. <https://doi.org/10.22329/il.v25i1.1040>.
- Henrich, Joseph, Steven J. Heine, and Ara Norenzayan. 2010. "The Weirdest People in the World?" *Behavioral and Brain Sciences* 33 (2–3):61–83. <https://doi.org/10.1017/S0140525X0999152X>.
- Hurley, Patrick J. 1997. *A Concise Introduction to Logic*. 6th ed. Belmont, CA: Wadsworth.
- Kitcher, Philip. 1989. "Explanatory Unification and the Causal Structure of the World." In *Scientific Explanation*, vol. 13, edited by Philip Kitcher and Wesley Salmon, 410–506. Minneapolis: University of Minnesota Press.
- Klein, Richard A., Michelangelo Vianello, Fred Hasselman, Byron G. Adams, Reginald B. Adams Jr., Sinan Alper, Mark Aveyard et al. 2018. "Many Labs 2: Investigating Variation in Replicability across Samples and Settings." *Advances in Methods and Practices in Psychological Science* 1 (4):443–90. <https://doi.org/10.1177/2515245918810225>.
- Knobe, Joshua. 2021. "Philosophical Intuitions Are Surprisingly Stable across both Demographic Groups and Situations." *Filozofia Nauki* 29 (2):11–76. <https://doi.org/10.14394/filnau.2021.0007>.
- Knobe, Joshua, and Shaun Nichols. 2017. "Experimental Philosophy." In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta. <https://plato.stanford.edu/archives/win2017/entries/experimental-philosophy/>.
- Krifka, Manfred, Francis Pelletier, Gregory Carlson, Alice Ter Meulen, Gennaro Chierchia, and Godehard Link. 1995. "Genericity: An Introduction." In *The Generic Book*, edited by Gregory Carlson and Francis Pelletier, 1–125. Chicago: Chicago University Press.
- Landis, J. Richard, and Gary G. Koch. 1977. "The Measurement of Observer Agreement for Categorical Data." *Biometrics* 33 (1):159–74. <https://doi.org/10.2307/2529310>.
- Leslie, Sarah-Jane. 2007. "Generics and the Structure of the Mind." *Philosophical Perspectives* 21 (1):375–403. <https://doi.org/10.1111/j.1520-8583.2007.00138.x>.
- Li, Jincui, and Xiaozhen Zhu. 2023. "Twenty Years of Experimental Philosophy Research." *Metaphilosophy* 54 (1):29–53. <https://doi.org/10.1111/meta.12602>.
- Little, Daniel. 1993. "On the Scope and Limits of Generalizations in the Social Sciences." *Synthese* 97:183–207. <https://doi.org/10.1007/BF01064114>.
- Machery, Edouard. 2023. "Why Variation Matters to Philosophy." *Res Philosophica* 100 (1):1–22. <https://doi.org/10.11612/resphil.2264>.
- Peters, Uwe. 2023. "Science Communication and the Problematic Impact of Descriptive Norms." *British Journal for the Philosophy of Science* 74(3):713–738. <https://doi.org/10.1086/715001>
- Peters, Uwe, and Mary Carman. 2023. "Unjustified Sample Sizes and Generalizations in Explainable AI Research: Principles for More Inclusive User Studies." *IEEE Intelligent Systems*. <https://arxiv.org/pdf/2305.09477.pdf>.
- Peters, Uwe, Alexander Krauss, and Oliver Braganza. 2022. "Generalization Bias in Science." *Cognitive Science* 46 (9):e13188. <https://doi.org/10.1111/cogs.13188>.
- Polit, Denise F., and Cheril Tatano Beck. 2010. "Generalization in Quantitative and Qualitative Research: Myths and Strategies." *International Journal of Nursing Studies* 47 (11):1451–58. <https://doi.org/10.1016/j.ijnurstu.2010.06.004>.
- Polonioli, Andrea, Mariana Vega-Mendoza, Brittany Blankinship, and David Carmel. 2021. "Reporting in X-phi: Current Standards and Recommendations for Future Practice." *Review of Philosophy and Psychology* 12 (1):49–73. <https://doi.org/10.1007/s13164-018-0414-3>.
- Rad, Mostafa Salari, Alison Jane Martingano, and Jeremy Ginges. 2018. "Toward a Psychology of *Homo sapiens*: Making Psychological Science More Representative of the Human Population." *Proceedings of the National Academy of Sciences of the United States of America* 115 (45):11401–5. <https://doi.org/10.1073/pnas.1721165115>.
- Reichenbach, Hans. 1951. *The Rise of Scientific Philosophy*. Bognor Regis, UK: University of California Press. <https://doi.org/10.1525/9780520341760>.
- Ritchie, Katherine. 2019. "Should We Use Racial and Gender Generics?" *Thought* 8 (1):33–41. <https://doi.org/10.1002/tht3.402>.

- Rothman, Kenneth J., John E. J. Gallacher, and Elizabeth E. Hatch. 2013. "Why Representativeness Should Be Avoided." *International Journal of Epidemiology* 42 (4):1012–14. <https://doi.org/10.1093/ije/dys223>.
- Simons, Daniel J., Yuichi Shoda, and D. Stephen Lindsay. 2017. "Constraints on Generality (COG): A Proposed Addition to All Empirical Papers." *Perspectives on Psychological Science* 12 (6):1123–28. <https://doi.org/10.1177/1745691617708630>.
- Smaldino, Paul E., and Richard McElreath. 2016. "The Natural Selection of Bad Science." *Royal Society Open Science* 3 (9):160384. <https://doi.org/10.1098/rsos.160384>.
- Stich, Stephen P., and Edouard Machery. 2023. "Demographic Differences in Philosophical Intuition: A Reply to Joshua Knobe." *Review of Philosophy and Psychology* 14:401–34. <https://doi.org/10.1007/s13164-021-00609-7>.
- Stuart, Michael T., David Colaço, and Edouard Machery. 2019. "P-Curving x-Phi: Does Experimental Philosophy Have Evidential Value?" *Analysis* 79 (4):669–84. <https://doi.org/10.1093/analys/anz007>. <https://doi.org/10.1093/analys/anz007>
- Thalmayer, Amber Gayle, Cecilia Toscanelli, and Jeffrey Jensen Arnett. 2021. "The Neglected 95% Revisited: Is American Psychology Becoming Less American?" *American Psychologist* 76 (1):116. <https://doi.org/10.1037/amp0000622>.
- Tessler, M. H., and N. D. Goodman. 2019. "The Language of Generalization." *Psychological Review* 126(3):395–436. <https://doi.org/10.1037/rev0000142>
- Vasilyeva, Nadya, and Tania Lombrozo. 2020. "Structural Thinking about Social Categories: Evidence from Formal Explanations, Generics, and Generalization." *Cognition* 204:104383. <https://doi.org/10.1016/j.cognition.2020.104383>.
- von Fintel, Kai. 1994. "Restrictions on Quantifier Domains." PhD diss., University of Massachusetts at Amherst.
- Wager, Elizabeth, and Sabine Kleinert. 2011. "Responsible Research Publication: International Standards for Authors." In *Promoting Research Integrity in a Global Environment*, edited by Tony Mayer and Nicholas H. Steneck, 309–16. Singapore: Imperial College Press/World Science.
- Walton, Douglas. 1999. "Rethinking the Fallacy of Hasty Generalization." *Argumentation* 13:161–82. <https://doi.org/10.1023/A:1026497207240>.
- Weinberg, Justin. 2015. "Does Philosophy Improve Critical Thinking?" *Daily Nous*, October 22. <https://dailynous.com/2015/10/22/does-philosophy-improve-critical-thinking/>.

Appendix

Table A1. Number of x-phi articles published in the eight selected journals by year

Year	Number
2017	15
2018	14
2019	19
2020	23
2021	44
2022	30
2023	26
Total	171

-
1. *Unrestricted:* “The results of our experiments clearly show that a large majority of people distinguish hard cases from easy ones and reveal response patterns that are predicted by the philosophical consensus.” (6)
Restricted: “The results of our experiments clearly show that a large majority of **the US target population** distinguish hard cases from easy ones and reveal response patterns that are predicted by the philosophical consensus.”

 2. *Unrestricted:* “Overall, the results clearly demonstrate that folk psychology views belief as voluntary.” (9)
Restricted: “Overall, the results clearly demonstrate that **study participants** viewed belief as voluntary.”

 3. *Unrestricted:* “Our first experiment shows that physicists are reliable when making judgements in thought experiments in physics.” (13)
Restricted: “Our first experiment found that physicists **were** reliable when making judgements in thought experiments in physics.”

 4. *Unrestricted:* “Our results also provide evidence that people take the passage of time to be a function of subjective experiences.” (35)
Restricted: “Our results also provide evidence that, **in some contexts, many people in the United States** (Europe, etc.) **may** take the passage of time to be a function of subjective experiences.”

 5. *Unrestricted:* “Our findings suggest philosophers are better at deploying concepts than laypeople but are susceptible to the linguistic salience bias to a similar extent and at similar points.” (42)
Restricted: “Our findings suggest the philosophers **in our sample were** better at deploying concepts than laypeople but **were** susceptible to the linguistic salience bias to a similar extent and at similar points.”
-

Figure A1. Unrestricted conclusions and restricted reformulations. Restricting components are in boldface.