Psychology is a Property of Persons, Not Averages or Distributions: Confronting the Group-to-Person Generalizability Problem in Experimental Psychology

Journal:	*Advances in Methods and Practices in Psychological Science*
Manuscript ID	AMPPS-22-0036.R2
Manuscript Type:	Empirical Article
Date Submitted by the Author:	n/a
Complete List of Authors:	McManus, Ryan; Boston College, Psychology Young, Liane; Boston College, Psychology Sweetman, Joseph; University of Exeter, Psychology
Substance Keywords:	cognition
Method and Stats :	variability, hypothesis testing, repeated measures < Design
Additional Keywords:	

AMPPS Questions about Transparency Practices - AMPPS-22-0036.R2

Registered Report

Q: Is your manuscript a Stage-1 Registered Report?

A: No

If yes, provide a link to the project location where the manuscript, materials, and data will eventually be stored if the study is provisionally accepted.

CUST_TRANSPARENCY_RR_TEXT :No data available.

Empirical Work

Q: Does your paper present the results of new studies or analyses of data from human participants or from animals?

A: Yes

If Yes, provide the name of the institution that granted ethical approval and the protocol number (e.g., e.g., Protocol #12345 approved by the University of Illinois IRB). If no ethical approval was required, give a brief explanation of why not.

Boston College IRB - Protocol 12.064

If your paper presents empirical work with human participants, please indicate whether it adhered to the Declaration of Helsinki.

A: Yes - 2013 Seventh Revision

If your paper presents new empirical work, does it include the following statement: "We report how we determined our sample size, all data exclusions, all manipulations, and all measures in the study"?

A: Yes

If your paper presents new empirical work and (a) you did not include this statement in your manuscript, (b) you included a modified version of this statement, or (c) any part of this statement is untrue, please explain why.

CUST_TRANSPARENCY_STATEMENT_TEXT :No data available.

Available Data

Q: Does your paper rely on new or previously unpublished empirical data from your lab?

A: Yes

Does your paper analyze data from pre-existing datasets or data made available by other researchers?

A: Yes

If you answered Yes to either of these questions, provide a URL (either public or view-only) where the data can be accessed by the editors and reviewers.

All URLs are contained on our OSF page, which is available here: https://osf.io/xyse4/

If your paper relies on existing data that are available via third parties, please indicate who controls access to those data and how other researchers can access them in the same way you have.

Some of our paper relies on existing data, all of which is available on OSF.

If needed, add any additional explanation about the data used in your paper.

CUST_TRANSPARENCY_TEXT_DATA_OTHER :No data available.

Available Materials

Q: Have you made available any and all materials necessary to reproduce your experiments, analyses, or other paper contents?

A: CUST_TRANSPARENCY_MATERIALS :No data available.

If No, please explain which materials are unavailable and explain why they are not available.

CUST_TRANSPARENCY_MATERIALS_TEXT :No data available.

Q: Does your paper rely on any materials, code, or other resources that are new to this project (i.e., they were developed or created as part of the research reported in this paper)?

A: Yes

If you answered Yes, provide a URL (either public or view-only) where reviewers and editors can view those materials and resources. Enter “Not Applicable” if your paper does not rely on any such materials.

All materials are contained on our OSF page, which is available here: https://osf.io/xyse4/

Q: Does your paper rely on any materials, code, or other resources that are in the public domain or previously made available by you or by other researchers?

A: No

If Yes, please indicate who controls access to those materials and how other researchers can access them in the same way you have. Enter “Not Applicable” if you are not relying on data from other researchers or third parties.

CUST_TRANSPARENCY_EXISTING_MATERIALS_TEXT :No data available.

Preregistration

Q: Were any of the studies or analyses reported in the manuscript preregistered?

A: Yes â€“ all studies were preregistered

If only a subset of the reported studies were preregistered, indicate which ones were and which ones were not. For any studies that were not preregistered, please indicate why not. If your manuscript does not report new studies, please enter, “Not applicable.”

CUST_TRANSPARENCY_PREREG_STUDIES :No data available.

Please provide a URL for the main project page where reviewers and editors can access the preregistration documentation (leave blank if there are no studies or none was preregistered). This may be an anonymous view-only link for the review process.

CUST_TRANSPARENCY_PREREG_URL :No data available.

Which aspects of your project were preregistered? (check all that apply. Note that if your preregistration includes a complete analysis script that handles coding of measures, missing data, exclusions, analyses, etc., you could check multiple boxes on this list based on preregistering that script.)

Theoretical hypotheses . Tasks/measures used for confirmatory hypothesis tests. Data exclusion criteria and procedures. Data analysis plan. Planned interpretation for different patterns of results (could be part of the â€œTheoretical hypothesesâ€☐ if each hypothesis states how different patterns of results would support or disconfirm it).. Target sample size

When did you complete the preregistration:

Prior to any data collection for the study (not including pilot testing of procedures)

Did you make any changes to the preregistered procedures when completing your study?

A: Yes

If you made changes to your preregistered plans at any stage of the process of completing your research, list all of those changes below.

For our statistical cognition studies, one of our reviewers pointed out that we should use one-tailed rather than two-tailed binomial tests. Therefore, we have made this change and footnoted that it deviated from our original pre-registration.

Such changes must also be documented in the manuscript itself. Are all of the changes specified above fully reported in the manuscript text?

Yes

If you have other comments or explanations about your preregistration that not covered by the questions above, enter them here (leave blank if no you have no comments/notes or if your paper has no preregistered studies):

CUST_TRANSPARENCY_PREREG_OTHER :No data available.

Q: Authors were asked to select items from the list below to indicate what was preregistered:

- Not Applicable - no preregistered studies
- Theoretical hypotheses
- Tasks/measures used for confirmatory hypothesis tests
- Tasks/measures used for exploratory hypothesis tests
- Tasks/measures that were collected for other purposes
- Data collection stopping rules
- Data source(s) (for preregistration of analyses of pre-existing data)
- Data coding procedures (e.g., how measures would be coded and scored)
- Data exclusion criteria and procedures
- Procedures for handling missing data
- Procedures to handle failures of quality control
- Data analysis plan
- Data analysis scripts/code (e.g., full R scripts for analysis)
- Planned interpretation for different patterns of results (could be part of the "Theoretical hypotheses" if each hypothesis states how different patterns of results would support or disconfirm it).
- Target sample size

The author checked the following boxes:

A: Theoretical hypotheses . Tasks/measures used for confirmatory hypothesis tests. Data exclusion criteria and procedures. Data analysis plan. Planned interpretation for different patterns of results (could be part of the â€œTheoretical hypothesesâ€☐ if each hypothesis states how different patterns of results would support or disconfirm it).. Target sample size

Psychology is a Property of Persons, Not Averages or Distributions: Confronting the Group-to-Person Generalizability Problem in Experimental Psychology

Ryan M. McManus[1*]

Liane Young[1]

Joseph Sweetman[2]

[1]Department of Psychology and Neuroscience, Boston College, Boston, MA, USA

[2]Department of Psychology, University of Exeter, Exeter, Devon, UK

* Corresponding author email:

mcmanurd@bc.edu

Acknowledgements: We would like to thank Stefano Anzellotti, Hiram Brownell, Richard Morey, Ehri Ryu, and Jordan Theriault for helpful feedback at the beginning of this project. We would like to thank Adam Bear, Tony Chen, Isaac Handley-Miner, Robin Ince, Minjae Kim, Aditi Kodipady, Gordon Kraft-Todd, Matthew Leitao, Shangzan (Sunny) Liu, Michael (Mookie) Manalili, Julia Marshall, Joshua Rottman, and Abraham Rutchick for helpful conversations at various stages of this project, as well as providing feedback on an early draft of the manuscript. We thank Nathan Liang and Sunny Liu (again) for investigating and diagnosing coding/output issues in R during the revision process. Finally, we thank James W. Grice, an anonymous reviewer, and Katie Corker, who provided invaluable feedback during the review process.

Abstract

When experimental psychologists make a claim (e.g., "Participants judged X as morally worse than Y"), how many participants are represented? Such claims are often based exclusively on group-level analyses; here, psychologists often fail to report, or perhaps even investigate, how many participants judged X as morally worse than Y. More troubling, group-level analyses do not necessarily generalize to the person-level: "the group-to-person generalizability problem." We first argue for the necessity of designing experiments that allow investigation of whether claims represent most participants. Second, we survey researchers (and laypeople), finding that most interpret claims based on group-level effects as being intended to represent most participants in a study. Importantly, most believe this ought to be the case if a claim is used to support a general, person-level psychological theory. Third, building on prior approaches, we document claims in the experimental psychology literature, derived from sets of typical group-level analyses, that describe only a (sometimes tiny) minority of participants. Fourth, we reason through an example from our own research to illustrate this group-to-person generalizability problem. Additionally, we demonstrate how claims from sets of simulated group-level effects can emerge without a single participant's responses matching these patterns. Fifth, we conduct four experiments that rule out several methodology-based noise explanations of the problem. Finally, we propose a set of simple and flexible options to help researchers confront the group-to-person generalizability problem in their own work.

Psychology is a Property of Persons, Not Averages or Distributions: Confronting the Group-to-Person Generalizability Problem in Experimental Psychology

Francis Galton attended the 1906 "West of England Fat Stock and Poultry Exhibition" where attendees, hoping to win a prize, estimated an ox's weight. Galton calculated that the crowd's average estimate was 1,197 pounds, a perfect match to the ox's true weight (Galton, 1907; Wallis, 2014). In this case, we might reasonably say that "people judged the ox's weight perfectly." Though this impressive example suggests the "wisdom of crowds" (Surowiecki, 2005), it is worth noting the considerable variability in person-to-person estimates, ranging below 1,000 pounds to above 1,400 pounds. In fact, the person-level data reveals that only one person guessed the correct weight of 1,197 pounds (Wallis, 2014). Consequently, we might question whether "people judged the ox's weight perfectly" in truth describes what happened, as the group-level average represented only one person. Due to the ubiquity of aggregation approaches in experimental psychology, this "group-to-person generalizability problem" may hinder progress and understanding. Psychologists average sets of person-level responses—largely ignoring person-to-person variability—and then use these averages to make claims about the mind. However, if psychology aims to understand the mind as a property of *persons*—to uncover the uniqueness or universality of certain psychological processes—person-level responses ought to be the explananda.

In this paper, we argue that although experimental psychologists often strive to describe person-level phenomena, they sometimes fail to do so. First, we make a data-free argument for closely matching experimental designs and analytic methods to precise research questions. Second, we survey laypeople and psychology researchers to understand what is inferred about person-level phenomena from group-level analyses. Third, we document instances in published literature where a person-level analytic approach yields different conclusions than typical group-level approaches. Fourth, in a tutorial, we show readers how this can occur, and how to describe person-level patterns in their own data. Additionally, we demonstrate how claims from sets of simulated group-level effects can describe zero persons. Fifth, we conduct four pre-registered experiments to rule out several methodology-based explanations of group-to-person generalizability failures. Finally, we propose a set of simple and flexible design and analytic strategies (ranging from descriptive to inferential) to address the group-to-person generalizability problem.

Psychology as the Study of Person-Level (Not Group-Level) Properties

Psychology is often defined as "the study of the mind and behavior." Therefore, its essential goals are describing cognitive functions and uncovering their antecedents and consequences. We contend that researchers intend to apply these goals to the study of persons, as psychological processes are properties of minds, and each mind resides inside a single person. To strengthen this argument, we ask readers to engage in a thought exercise. Recall your most recent meeting with collaborators where you discussed hypotheses and experimental designs to test them. At any point in that meeting, did you reason about possible patterns in a way that reflected how *persons* may respond to different stimuli, or did you exclusively reason in a way that reflected how different stimuli would affect *averages or locations of distributions*? Furthermore, given the seeming frequency with which studied phenomena are described as applying to people generally, we also contend that many experimental psychologists intend to uncover processes, regularities, and mechanisms that describe a *majority* of persons (i.e., "general psychological laws"; Hamaker, 2012). Therefore, what follows are the most important takeaways from this paper:

1. Psychologists sometimes fail to design experiments that permit descriptive or inferential investigation of person-level hypotheses.

2. Even when appropriate experimental designs are used, psychologists often report *only* their group-level analyses and interpret them *as if* they support or falsify person-level hypotheses.

Because it is possible for the above statements to be misinterpreted or overgeneralized, we first communicate what we mean by "person-level,", and we then clarify our position on designing studies to test person-level hypotheses.

Examining "Person-Level" Hypotheses

A "person-level" hypothesis is one that predicts some effect(s) on an outcome measure for a single person (e.g., the direction and magnitude of an effect for person X). To test it one can employ within-person or "single-subject" analysis as seen in (relatively high-trial) neuroimaging designs (Friston, et al., 1994) or "intensive" sampling in longitudinal designs (e.g., Kurz, et al., 2019). If the goal is to know how many participants show a predicted effect, a "pervasiveness" proportion can be obtained (Speelman & McGann, 2020). By pervasiveness, we mean the choosing of one possible person-level pattern and investigating, descriptively, "How many persons match this pattern?" Randomization tests can examine whether the pervasiveness of the effect(s) in the sample is unrelated to experimental condition – i.e., emerges more than "physical chance" (Grice, 2021; Grice et al., 2020). Finally, we can combine pervasiveness and within-person approaches to estimate the prevalence of person-level effects in the population (see Allefeld, et al., 2016; Donhauser, et al., 2018; Ince, et al., 2022; Ince, et al., 2021) and test against a "global null hypothesis" (no effect in any subject in the population) or a "majority null hypothesis" (the effect is in less than, or equal to, half the population), if one is intending to test or make a *general* psychological claim about most people in the population.

Within-Subjects (vs. Between-Subjects) Designs for Testing Person-Level Hypotheses

Between-subjects experiments do not permit tests of person-level hypotheses (Speelman & McGann, 2020; Whitsett & Shoda, 2014). These common designs make it impossible to ask the simple question, "How many people's responses match the pattern(s) indicated by the mean difference(s) between conditions?" (see Speelman & McGann, 2020), and they prohibit examination of unfolding person-level processes (e.g., Brandt & Morgan, 2022; Fisher, et al., 2018; Moeller, 2022). For example, consider the following research question: "Is Coca-Cola tastier than Mountain Dew?" To assess this, the leading soda cognition lab designs an experiment which randomly assigns half of participants to rate the tastiness of Coca-Cola, and the remaining participants to rate Mountain Dew in the same way. An independent-samples t-test suggests that the average tastiness judgment is higher for Coca-Cola. However, a rival soda cognition lab also attempts to answer this question, instead using a within-subjects design and finding an average tastiness difference in the opposite direction. Assuming the within-subjects effect generalizes to the person-level (i.e., most people judged Mountain Dew as tastier than Coca-Cola), which of these designs better answers the question, "Is Coca-Cola tastier than Mountain Dew?" If *tastier* implies a comparison of at least two taste-able stimuli, we suggest that the within-subjects design is superior. Moreover, there are many plausible non-substantive mechanisms for the between-subjects results (e.g., the participants who rated Coca-Cola as extremely tasty may have been implicitly comparing it to Pepsi instead of Mountain Dew, an unlikely problem in the within-subjects design).

To illustrate this possibility in a different domain, Birnbaum (1999) had participants judge the largeness of numbers on a 10-point scale ranging from *very very small* to *very very large.* He showed that "People judge 9 as larger than 221" can be inferred from a between-subjects design, as 9 invokes a context of 2-digit numbers whereas 221 invokes a context of 3-digit numbers. We argue (and it was indeed Birnbaum's point) that no serious experimentalist would interpret these results to suggest that people would judge 9 as larger than 221 if they explicitly compared the numbers (and we again note that "judge… as larger than…" implies a comparison). If Birnbaum were to use his data to argue that this finding reflected true numerical cognition, it would be easy to criticize because we all believe that there is a truth of the matter (i.e., most [if not all] people believe 9 is smaller than 221), and that there are better and worse ways of verifying it. In many psychological experiments, however, measures of interest do not have clear numerical translations that map onto often-used Likert-type scales (e.g., angriness, agreement, etc.), making it more difficult to identify the problem raised by Birnbaum. Additionally, unlike Birnbaum's numerical cognition example where we know the truth of the matter, the point of many psychological experiments is to infer the truth of the matter from the data (e.g., "face A is judged as angrier than face B"). This means that it is unknown how often between-subjects results are taken to reflect within-subject phenomena when the between-subjects results are truly akin to Birnbaum's findings. If some non-trivial proportion of between-subjects experiments in psychology are designed with the intention to reveal a psychological process or its outcome, this problem may be pervasive.

Clarifying the Problem

We are not suggesting that between-subjects designs are never useful. These designs may be preferable when within-subjects designs are practically infeasible or impossible. For example, many intervention(-like) research questions may be best answered with between-subjects designs (e.g., see our SOM's experiments). Additionally, hypotheses about population(-like) differences require at least one between-subjects factor, such as testing whether psychopaths show different experimental effects than non-psychopaths. Finally, between-subjects designs are unproblematic when the research goal is to provide generalization evidence (e.g., finding similar effects across instructions/measures; see Yarkoni, 2020).

We note, however, that between-subjects designs cannot conclusively provide person-level evidence of an experimental effect, just as group-level correlations among variables cannot provide evidence of person-level correlations among those variables (see Fisher et al., 2018). For example, in our own recent moral cognition research, we assessed moral character judgments to test their sensitivity to social relationship information in the context of helping behavior (McManus et al., 2021). Among other variations, participants in our experiments were given two scenarios: one in which someone helps a total stranger, and another in which someone helps a distant family member. Standard group-level analyses suggested that participants–*on average*–judged agents who helped strangers as more morally good than agents who helped family members, presumably because people believe that there is less obligation to help strangers. Importantly, this was tested using a within-subjects design. Therefore, although it was not reported, our design permitted investigation of the question, "How many people's responses match the pattern indicated by the difference between conditions?" A between-subjects design would have disallowed such investigation.

Importantly, using within-subjects designs does not automatically prevent group-to-person generalizability inference errors from occurring. Researchers can still commit ecological or ergodic fallacies (Kuppens & Pollet, 2014; Speelman & McGann, 2020), due to special instances of Simpson's paradox—when group-level patterns poorly represent lower-level units constituting the group (Simpson, 1951; Kievit, Frankenhuis, Waldorp, & Borsboom, 2013; also see Hamaker, 2012, for an illustrative example on the relation between typing speed and mistake frequency). To reiterate, even when psychologists deploy appropriate experimental designs, they often, if not always, only report their group-level analyses, leaving it unclear whether their group-level findings generalize to the person-level.

Overall, we are suggesting that, if a research hypothesis or theory is a person-level one, and the goal of a study is to make a general claim (Hamaker, 2012), then researchers ought to choose appropriate designs and analytical procedures that allow themselves (and readers) to answer the question, "What proportion of people in the sample (or population) show the effect(s) indicated by the mean difference(s) between conditions?" However, it could be argued that most psychology researchers (and lay readers of the psychology literature) do not expect published claims to be representative of most people, nor may they believe it is important evidence for evaluating the validity of a psychological theory, so long as typically reported group-level effects corroborate predictions.

Empirically Assessing Laypeople's and Researchers' Inferences

We have argued that because of the ubiquity of typical group-level statistical tests (e.g., t-tests), there may be a group-to-person generalizability problem in psychology (i.e., when claims derived from typical group-level tests fail to describe most participants in the sample or the population). However, there is obvious subjectivity involved when deciding what should count as sufficient person-level evidence for a claim. Moreover, perhaps readers of psychology research (laypeople and psychology researchers themselves) do not interpret authors as intending to make claims that represent most participants. We therefore set out to answer two questions empirically:

1. Do a majority of people who read psychology research believe that authors intend to communicate claims as representing most participants in their data?
2. Do a majority of people who read psychology research believe that claims ought to represent most participants when the authors use their data to claim support for a general theory of person-level psychology (i.e., a theory/model of processes occurring within individual minds/brains)?

To answer these questions, we surveyed laypeople and researchers by presenting modified excerpts of "results" and "general discussion" sections from publications that contain the group-to-person generalizability problem. We report how we determined our sample sizes, all data exclusions, all manipulations, and all measures.

Method

Participants

All laypeople were U.S. residents recruited and compensated via CloudResearch's "approved participants" list. Participants from McManus et al. (2021) were unable to access the current study. Additionally, participants from our methods experiments could not participate. Researchers were affiliated with the Society for Personality and Social Psychology (SPSP), recruited via SPSP's Open Forum listserv and compensated with Amazon gift cards. Participants who did not complete the entire study were not included in our final analyses. As pre-registered (https://osf.io/6qay8 and https://osf.io/nucbf), we aimed to collect at least 642 analyzable laypeople and 280 analyzable researchers. In total, we were able to collect 705 and 256 unique responses, respectively. After applying the pre-registered exclusion criterion (failing a comprehension check), this resulted in $N_{Laypeople}$=588 (gender: 309 female, 273 male, 6 non-binary; ethnicity: 457 White, 68 Black, 5 American Indian, 41 Asian; 1 Pacific Islander; 16 other; M_{Age} = 38.69, SD_{Age} = 11.29) and $N_{Researchers}$=244 (165 female, 68 male, 8 non-binary, 3 other; ethnicity: 158 White, 3 Black, 1 American Indian, 55 Asian; 17 other, 9 Biracial; 1 Multiracial; M_{Age} = 33.09, SD_{Age} = 11.34). Although we did not pre-register a stopping rule, we decided not to resample due to still having high statistical power for our focal hypothesis tests (see *Statistical Power & Hypotheses)*.

Design

Participants were randomly assigned to one of two conditions. Half of participants learned about a simple effect comparison, whereas the other half of participants learned about a more complex, two-way interaction effect. We note that we used both simple and complex effect examples to test the generality of our hypotheses. That is, had we only conducted the study using one effect type, we could have capitalized on our hypothesis only being true of a specific effect type. This is why our pre-registration refers to our design as "observational," even though we randomly assigned participants to one effect type; we never intended to (nor did we) explicitly compare the simple effect data to the complex effect data.

Materials and Procedure

At the beginning of the study, all participants were informed that they would be answering questions about a moral cognition experiment. For the simple effect condition, participants learned about a two-condition comparison from the supplemental materials of Law, Campbell, & Gaesser (2021). For the complex effect condition, participants learned about a crossover interaction effect from McManus et al. (2021).

Participants first read text communicating results in typical journal article format (with means, SDs, t-values, p-values, within-subject standardized effect sizes for comparisons of interest [d_z], and a barplot; see OSF for full materials). After learning the results, they then read text that simulated how data-based claims are made in a general discussion section (e.g., "People judged fictional agents who helped a stranger as more morally good than fictional agents who helped a cousin, but they judged fictional agents who helped a stranger instead of a cousin as less morally good than fictional agents who helped a cousin instead of a stranger").

After learning about the claim, participants were then asked to respond to a series of true-false questions about what the reported results suggested. However, these questions were not of primary interest (see OSF for Rmarkdown results). Participants were then again shown the claim in general discussion format, and asked "By *people,* approximately what percentage of the study's participants do you think the researchers mean?" We call this measure the "empirical proportion estimate." Responses ranged from 0-100% on a sliding scale, with the starting position (0, 50, 100) counterbalanced across participants. This measure allows categorization of responses into two categories: less than a simple majority (50% or less), and equal to or greater than a simple majority (51% or more). To move on to the next page, participants had to at least click on the slider, meaning that the slider's starting value would have been recorded as the participant's response. As can be seen in Figure 1, however, these exact starting values were very infrequent, suggesting that participants indeed engaged with the task.

Next, participants learned about a (fictional) general, person-level theory that the authors had developed pre-study. Participants were then asked to respond to a series of true-false questions about how the reported results informed the theory (see OSF). Participants were again shown the claim in general discussion format and told that, later in the paper, the authors used their study's results to claim support for their theory. Participants were then asked, "In order for the study's results to support the researchers' theory/model, approximately what percentage of the study's participants do you think need to respond in the way described by [the general discussion's language]?" We call this measure the "theoretical proportion estimate." Responses were measured identically to the empirical estimate. Finally, participants could write an open-ended response to communicate anything that they were unable to communicate thus far. After the main task, participants answered several demographic questions.

Statistical Power

As pre-registered, we aimed for at least 321 participants per condition for the laypeople sample, and 140 participants per condition for the researcher sample. The pre-registered laypeople sample size yielded 95% power to detect a 10-point proportion difference from 50% (e.g., 60%) using a two-tailed binomial test and assuming an alpha level = 0.05, the focal test to examine whether a majority of empirical/theoretical proportion estimates reflect inferences being made about a majority of a study's participants. As explained in our pre-registrations, we planned the researcher sample based on the results of the laypeople sample. For the researcher sample, the pre-registered sample size yielded 95% power to detect a 15-point proportion difference from 50% using identical test specifications as the laypeople sample.

In the laypeople sample, applying the pre-registered exclusion criterion (i.e., missing a comprehension check question) led to N_{Simple}=303 and $N_{Complex}$=285. In the researcher sample, we were unable to successfully recruit our entire desired sample size. After one attempt to get more responses (via reposting to SPSP's Open Forum listserv), we decided to close the survey once incoming responses completely stalled, which occurred after two weeks. Applying the same exclusion criterion led to N_{Simple}=123 and $N_{Complex}$=121. We did not resample for either population because sensitivity analyses revealed that we still had more than 90% power to detect our pre-registered minimal effect sizes.

Hypotheses

1) Empirical Proportion: The majority of laypeople and researchers (i.e., 51% or more) will believe authors' claims are intended to describe at least a simple majority (i.e., 51% or more) of their study's participants.

2) Theoretical Proportion: The majority of people will believe at least a simple majority of a study's participants ought to be described by the authors' claims in order for the results to support a general theory of person-level psychology.

Results

Empirical Proportion Estimate

The majority of laypeople believed authors intended to describe at least a simple majority of their study's participants, for both simple (81%) and complex (88%) effects. The majority of researchers agreed for both simple (73%) and complex (80%) effects (see Table 1 for additional descriptive statistics and Tables 2-3 for inferential statistics). Strikingly, as shown in Figure 1, there is no discernible pattern as a function of being relatively inexperienced (e.g., layperson or undergraduate) and relatively experienced with academic research (e.g., professor). Moreover, even though most people's judgments were above 50%, judgments ranged from nearly 0% to 100%. This suggests a lack of generality in inferences across persons, additional evidence in favor of the importance of investigating person-level responses.

Theoretical Proportion Estimate

The majority of laypeople believed that at least a simple majority of a study's participants ought to be described by authors' claims for the results to support a person-level psychological theory, for both simple (93%) and complex (92%) effects. The majority of researchers agreed for both simple and (80%) and complex (90%) effects (see Table 1 for additional descriptive statistics and Tables 2-3 for inferential statistics). As shown in Figure 1, again, there is no discernible pattern as a function of research experience[1].

Table 1. Descriptive Statistics for Empirical and Theoretical Estimates (Split by Population)

Estimate	Effect Type	Population	Mean (SD)	Median	Range
Empirical					
	Simple	*Laypeople* (N = 303)	62.17 (18.08)	62	5 – 100
		Researchers (N = 123)	61.24 (20.40)	60	0 – 100
	Complex	*Laypeople* (N = 285)	68.56 (15.96)	62	0 – 100
		Researchers (N = 121)	63.20 (18.37)	65	0 – 100
Theoretical					
	Simple	*Laypeople* (N = 303)	65.77 (15.12)	65	10 – 100
		Researchers (N = 123)	64.10 (19.93)	65	0 – 100
	Complex	*Laypeople* (N = 285)	69.80 (14.96)	74	10 – 100
		Researchers (N = 121)	67.89 (16.69)	71	10 – 100

Note: In the researcher sample, for the empirical estimates, a small minority used the open-ended question to *correctly* communicate that inferences about percentages cannot be derived from average differences (n = 17). Therefore, some of the empirical estimates were not true beliefs, as the researchers simply had no other option but to respond. To conduct the most stringent test of our hypothesis, we recoded all of the hypothesis-consistent slider responses (n = 6) as being hypothesis-inconsistent. We did not remove any of the 17 responses to ensure that, even accounting for some researchers understanding the problem, a majority still responded in a hypothesis-consistent way. This resulted in similar proportions for both simple (70%) and complex (79%) effects. Similarly, for the theoretical estimates, some people communicated that there were other features that matter for establishing that a claim provides evidence for the validity of a theory (e.g., showing an effect across diverse samples, under multiple conditions, across stimulus sets, etc.). However, we did not recode any of these responses as being hypothesis-inconsistent, because implicit in these responses is part of the point we intend to make: To have evidence for a general theory, psychologists must show an effect's prevalence (across samples, situations, time, and importantly, across *persons*).

Table 2. Empirical Estimate Tests within Each Effect Type (split by Population)

Effect Type	Population	Proportion	*p-value*
Simple	*Laypeople*	81% [77% - 100%]	< .001
	Researchers	73% [67% - 100%]	< .001
Complex	*Laypeople*	88% [86% - 100%]	< .001
	Researchers	80% [75% - 100%]	< .001

Note: Proportions of laypeople/researchers who indicated that the empirical proportion of the study's participants who matched the claim was at least a simple majority. Brackets underneath proportions indicate 90% CIs for the proportion estimate. P-values were computed via one-tailed binomial tests against 0.50.[2]

Table 3. Theoretical Estimate Tests within Each Effect Type (split by Population)

Effect Type	Population	Proportion	*p-value*
Simple	*Laypeople*	93% [91% - 100%]	< .001
	Researchers	80% [75% - 100%]	< .001
Complex	*Laypeople*	92% [90% - 100%]	< .001
	Researchers	90% [86% - 100%]	< .001

Note: Proportions of laypeople/researchers who indicated that the proportion of the study's participants who needed to match the claim was at least a simple majority if the results were to be used to support a person-level psychological theory. Brackets underneath proportions indicate 90% CIs for the proportion estimate. P-values were computed via one-tailed binomial tests against 0.50.

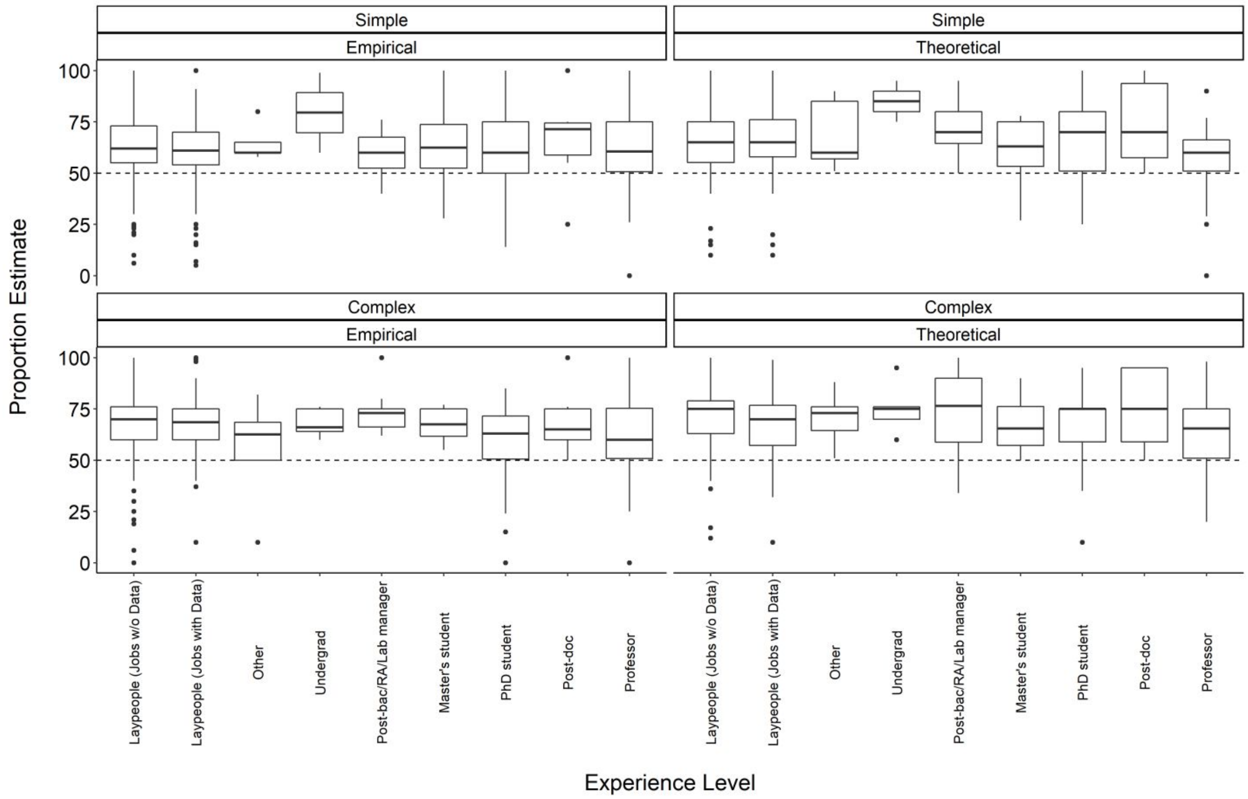

Figure 1. Boxplots of empirical/theoretical proportion estimates by effect type (simple versus complex), and by participants' level of experience. Note that "Other" refers to people involved in academic research in some way (via SPSP) but who indicated that they have never held an academic position. Histogram versions of these figures are available our OSF page under "Statistical Cognition Studies."

Discussion

Overall, our data suggests that most laypeople and researchers interpret claims as being intended to describe most participants. Moreover, they believe this ought to be the case if the data are used to support a general theory of person-level psychology. These findings are problematic when considering how analyses are typically conducted and reported. First, if most researchers (and the public) interpret results of group-level tests as representing most sampled participants (and therefore most people in the population), it is unknown how often this interpretation is incorrect, as person-level statistics are rarely (if ever) reported in published articles. Second, if a criterion for a claim to be able to properly support a theory or model is that it represents most sampled participants (and therefore most people in the population), then there are multitudes of psychological claims in the published literature that have not yet been properly tested, as aggregation approaches (e.g., averaging across different participants' responses) are ubiquitous in experimental psychology. The rest of this paper focuses on documenting and explaining published and simulated instances in which within-subjects group-level effects fail to describe most sampled persons – the group-to-person generalizability problem.

Group-to-Person Generalizability Problems in the Wild

We examined open data from psychological research over the past five years (2016–2021), looking for the group-to-person generalizability problem. Due to the larger reform movements in psychology, publications from this era should be relatively more rigorous than prior eras (e.g., larger samples, better statistical inferences). Our investigation was not systematic in the sense that we can say, "X% of publications contain the group-to-person generalizability problem." Rather, using a person-level approach, we re-analyzed open data with the goal of finding five instances of the problem from moral cognition—as we ourselves are moral psychologists—and five instances from social cognition generally (e.g., on race, gender, humor, etc., see Table 4). Even though we investigated examples from social cognition in particular, this problem is not limited to social cognition, as others have identified pitfalls of averaging across persons in somewhat lower-level research on judgment and decision-making (Liew et al., 2016) and face perception (Grice et al., 2020).

To accomplish person-level analysis, we adopted “pervasiveness” or “persons-as-effect-sizes” approaches (see Grice et al., 2020; Speelman & McGann, 2020). Put simply, we created variables in each dataset that distinguished participants based on whether their response patterns supported the reported group-level patterns. If a participant’s responses had at least *some* distance between experimental conditions (e.g., 1-point on a Likert/sliding scale in a one-trial per condition design) and were directionally consistent with a group-level pattern, then that participant was categorized as supporting group-to-person generalizability. An important nuance is that all investigated claims are based on *sets* of group-level tests (e.g., multiple paired t-tests). We therefore extended extant person-level approaches to accommodate such claims. Specifically, we categorized participants as supporting generalizability if their full set of responses matched the full set of group-level patterns. For example, if a 2x2 interaction pattern underlaid the claim, we counted person-level responses as supporting generalizability if a participant’s simple effects’ directions and differential magnitudes reflected the group-level pattern. But the ordering of all four condition averages was not accounted for, as this is not typically relevant to the interpretation of statistical interactions. A minimal difference in the predicted direction could be seen as a liberal threshold for examining the group-to-person generalizability problem. Readers can imagine (and if they wish, investigate) what these analyses look like under stricter constraints (see our OSF page: https://osf.io/xyse4/).

For each claim, we used the descriptive sample proportion as a proxy for the proportion of people in the population who would be expected to show the group-level patterns. If the sample proportion was equal to or lower than 0.50, then we considered the claim unsupported at the person-level. We chose this 0.50 value because most claims in psychology articles do not use language that suggests an experimental effect is one that describes only a subset of participants. This means that, at least by implication, effects are being communicated as applying to *most* participants. Moreover, our statistical cognition studies revealed that most laypeople and researchers infer reported effects as applying to more than 50% of participants. As Table 4 shows, proportions of participants favoring generalizability varied across publications but was low overall (3%-50%, with most proportions ranging between 20%-40%). Critically, this occurred across a variety of dependent variables (e.g., sliding scales, Likert scales, reaction times, error rates) and pattern types (crossover interactions, attenuation interactions, ordinal patterns, conjunctive differences), suggesting that this problem is not constrained to specific designs or measures.

Table 4. Quotes, relevant tests, and person-level proportions for instances of the group-to-person generalizability problem

Publication	Exact Quote(s)	Group-Level Test(s)	Person-Level Proportions
McManus, Mason, & Young (2021)	"On the one hand, people judged agents who helped a stranger as more morally good than agents who helped a family member. On the other hand, people judged agents who helped a stranger instead of a family member as less morally good than agents who helped a family member instead of a stranger."	Experiments 1a-b -2 x 2 interactions -Set of paired t-tests -See Figure 2	E1a: **31%** (62 / 203) E1b: **29%** (59 / 203)
Law, Campbell, & Gaesser (2021)	"People consistently view socially distant altruism as less morally acceptable as the person not receiving help becomes closer to the agent helping."	Experiments 1 & 4 -Set of paired t-tests -See Figures 1 & 7b (Country vs Town vs Friend vs Family)	E1**: 3%** (3 / 97) E4: **8%** (30 / 397)
Fowler, Law, & Gaesser (2021)	"The results showed that moral judgments of empathy are biased toward preferring more empathy for a socially close over a socially distant individual. Despite this bias in moral judgments, however, people consistently judged feeling equal empathy as the most morally right perspective."	Experiment 2 -Set of paired t-tests -See Figure 3 (More For Distant vs More For Close vs Equal)	**32%** (97 / 304)
Soter, Berg, Gelman, & Kross (2021)	"Participants said they should protect close others more than distant others. However, the effect of relationship was consistently weaker for "should" judgments than "would" judgments, revealing that people show *relatively less* partiality in their judgments of what is morally right, compared to judgments of how they would act."	Experiment 2 -2 x 2 interaction -Simple comparisons -See Figure 2	**29%** (104 / 356)
Rottman & Young (2019)	"In three studies, adult participants judged the moral wrongness of harm and purity transgressions that varied in frequency (e.g., occasionally vs. regularly) or magnitude (e.g., small vs large) with the same sets of modifiers or the same quantities (e.g., a single drop vs. a teaspoon) repeated across content domains. All studies found that evaluations of purity violations were considerably less sensitive to variations in scope than evaluations of harms, yielding robust statistical interactions between domain and dosage."	Experiments 1-3 -2x2 interactions -Simple comparisons -See Figures 1-3	E1: **29%** (51 / 177) E2: **46%** (37 / 81) E3: **22%** (37 / 168)

Deska et al. (2020)	"We also observed an interaction between target race and target gender for life hardship. As with social pain, it was clear that participants generally agreed that Black targets experience greater life hardship than White targets; however, this seemed to be especially true for male targets."	Experiment 4 -2x2 interaction -Simple comparisons	**50%** (66 / 131)
Stroessner et al. (2020)	"An association between a gender category and a shape would be revealed by faster categorization speeds following compatible (masculine-square and feminine-circle) compared with incompatible (masculine-circle and feminine-square) prime-target pairings." "Along with the results of Studies 3a–3c, these data demonstrate that gender categorization of basic squares and circles occurs without intention."	Experiments 2 & 4 -2x2 interaction -Sets of paired t-tests -See Figure 3	E2: **38%** (26 / 69) E4: **41%** (61 / 150)
Craig, Nelson, & Dixon (2019)	"We found that the presence of a beard increased the speed and accuracy with which participants recognized displays of anger but not happiness." "In Experiment 1, facial hair facilitated recognition of anger, and the advantage in response times cannot be attributed to a shift toward responding "angry." Recognition of facial expressions of happiness, which are positive and nonthreatening, was slowed by the presence of a beard in this task."	Experiment 1 -2x2 interactions -Sets of paired t-tests -See Figure 2	Speed: **45%** (99 / 219) Accuracy: **25%** (55 / 219) Both: **13%** (29 / 219)
Decelles, Adams, Lowe, & John (2021)	"Using a sample of working professionals, including fraud investigators and auditors, we found in Study 4 that an angry response to an accusation was interpreted as a sign of guilt, relative to remaining calm. Moreover, compared with remaining calm and with angrily denying an accusation, remaining silent was also perceived as a cue of guilt and therefore does not appear to be a viable solution for the accused to avoid the negative effects of anger."	Experiment 4 -Set of paired t-tests (Anger vs Calm & Silent vs Calm)	**38%** (52 / 136)
Thai, Borgella, & Sanchez (2019)	"Study 3 demonstrated that it was deemed most acceptable for a person to make jokes about a particular social group if they themselves were a part of that social group. This remained true for both minority-directed and majority-directed humor. This pattern emerged consistently for all three categories of humor studied, including race-based, sexual orientation-based, or gender-based humor."	Experiment 3 -2x2 interaction -Simple comparisons -See Figure 4 (Gender-based Jokes)	**45%** (31 / 70)

Note: Across publications, it was sometimes difficult to find specific claims which could be connected back to specific hypothesis tests. For some publications, there was not a specific, insulated claim which clearly referenced a specific hypothesis test (e.g., Stroessner et al., 2020), which is why some quoted sections are taken from multiple sections of the publication. In Law, Campbell, & Gaesser (2021), the verbal claim was not an accurate representation of the set of group-level patterns (some necessary group-level patterns did not emerge). However, re-analysis of their data was based on the claim rather than the group-level patterns.

At this point, an important objection may be raised. Some of the proportions in Table 4 are quite far from zero, meaning that it is likely that some of the documented group-level patterns are indeed the most common (i.e., modal) person-level pattern within their respective datasets. If this is generally true, then perhaps there is not a problem of group-to-person generalizability. For example, in our own prior research (McManus et al., 2021), the documented group-level patterns are the modal person-level patterns, at ~30% of participants, with the next most common patterns matching only ~13% of participants. Upon this person-level re-analysis, we could have argued, "Although the group-level patterns are not ones that *most* participants show, the most common person-level patterns mirror the group-level patterns. That is, if we were to randomly survey one new person from the population and asked to make a bet, we would (and should) bet on the documented group-level patterns being the pattern that the new person shows."

While we value this argument, it is important to consider whether this is what most psychologists are intending to achieve when conducting experiments and making claims. There are at least two possibilities. First, most psychologists may be interested in basic science and therefore attempting to document general psychological laws (e.g., Hamaker, 2012), regularities or mechanisms. Second, most psychologists may be interested in applied science and therefore answering questions about whether it is a good idea to get a certain intervention or enact a certain policy change (e.g., to help or appease the largest subset of people). These are obviously not mutually exclusive, and we see either of these options as worthy pursuits. However, because of what our statistical cognition studies revealed, and because we ourselves are more concerned with basic science, we focus the rest of this paper on group-to-person generalizability problems when the research goal is attempting to document general psychological laws, regularities, or mechanisms (though we still advocate for investigating person-level data in applied research so that the commonness of certain responses is known and disclosed). We next unpack an example from our own moral cognition research showing how the group-to-person generalizability problem can occur.

Tutorial for the Group-to-Person Generalizability Problem (McManus et al., 2021)

For relevant background, consider the two earlier moral cognition scenarios: someone helps an unrelated stranger, and someone helps their cousin. We predicted that agents who helped strangers should be judged as more morally good than agents who helped their cousin, due to stranger-helping agents lacking an obligation to help but doing so anyway. Now consider these two scenarios in a slightly different context: someone chooses to help an unrelated stranger *instead of* their cousin, and someone chooses to help their cousin instead of an unrelated stranger. We predicted the opposite pattern here, as stranger-helping agents would be violating their family obligation. These two contexts were described as “No Choice” and “Choice” contexts, respectively. Indeed, this interaction and context-based reversal of simple effects emerged at the group-level.

In the general discussion, we communicated this effect as follows: “On the one hand, people judged agents who helped a stranger as more morally good than agents who helped a family member. On the other hand, people judged agents who helped a stranger instead of a family member as less morally good than agents who helped a family member instead of a stranger.” As two of the three authors of the current paper were authors, we can say, honestly, that we intended to communicate this effect as applying to most people (i.e., as a general, causal regularity). Therefore, our claim is interesting, and arguably, accurate, if *and only if* the interaction describes most participants’ psychology. We next explain how readers can reason through and investigate this person-level prediction by using their typical ANOVA and t-test knowledge as scaffolding.

To investigate the above claim at the person-level, each simple effect and the interaction can be described by a set of directional patterns. The No Choice simple effect can be computed by subtracting the “helped a cousin” ratings from the “helped a stranger” ratings, whereas the Choice simple effect can be computed by subtracting the “helped a cousin instead of a stranger” ratings from the “helped a stranger instead of a cousin” ratings. An interaction effect can then be computed by subtracting the Choice effect from the No Choice effect (see Table 5 and Figure 2 for an example of 13 hypothetical participants who reflect all possible qualitative patterns, and Table 6 for example R code to create generalizable 2x2 person-level patterns and investigate their descriptive proportions). The person-level combination in Table 5 and Figure 2 which matches the published claim is pattern number 6 (i.e., the “Positive, Negative, Positive” pattern: No Choice simple effect, Choice simple effect, Interaction effect). Conversely, a person-level combination which does not match the published claim but can still be categorized as showing a “Positive” interaction value is pattern number 10 (i.e., the “Positive, Zero, Positive” pattern).

Table 5. Example hypothetical participants, showing all possible qualitative patterns in McManus et al. (2021)

Subj	NC_Stranger	NC_Cousin	C_Stranger	C_Cousin	NC_Diff	C_Diff	Intx	NC_Direction	C_Direction	Int_Direction
1	1	3	2	3	-2	-1	-1	Negative	Negative	Negative
2	2	3	1	3	-1	-2	1	Negative	Negative	Positive
3	2	3	2	3	-1	-1	0	Negative	Negative	Zero
4	2	3	2	1	-1	1	-2	Negative	Positive	Negative
5	2	3	2	2	-1	0	-1	Negative	Zero	Negative
6	**3**	**2**	**1**	**2**	**1**	**-1**	**2**	**Positive**	**Negative**	**Positive**
7	3	2	3	1	1	2	-1	Positive	Positive	Negative
8	3	1	3	2	2	1	1	Positive	Positive	Positive
9	3	2	3	2	1	1	0	Positive	Positive	Zero
10	3	2	2	2	1	0	1	Positive	Zero	Positive
11	3	3	1	2	0	-1	1	Zero	Negative	Positive
12	3	3	2	1	0	1	-1	Zero	Positive	Negative
13	3	3	2	2	0	0	0	Zero	Zero	Zero

Note: Each of these hypothetical person-level patterns constitute all possible combinations of two simple effects directions, leading to 13 possible interaction patterns. “NC” and “C,” denote No Choice and Choice, respectively, as communicated in McManus et al., (2021). Subject row 6 is bolded to highlight the pattern that matches the claimed effect. The first four non-subject columns are hypothetical raw scores in each within-subjects condition. The next two columns are hypothetical difference scores which constitute the simple effects of interest. Simple effects (NC_Diff and C_Diff) are calculated by subtracting “Cousin” scores from “Stranger” scores. The “Intx” column contains the interaction values which are computed by subtracting the second simple effect from the first simple effect. The last three columns are directional labels to communicate the full person-level pattern for each subject. For ease of calculation and communication, this table assumes that hypothetical participants used a simple three-point scale. In principle, the number of scale points are irrelevant so long as the scale has more than two points (otherwise, there could not be differential magnitudes of simple effects).

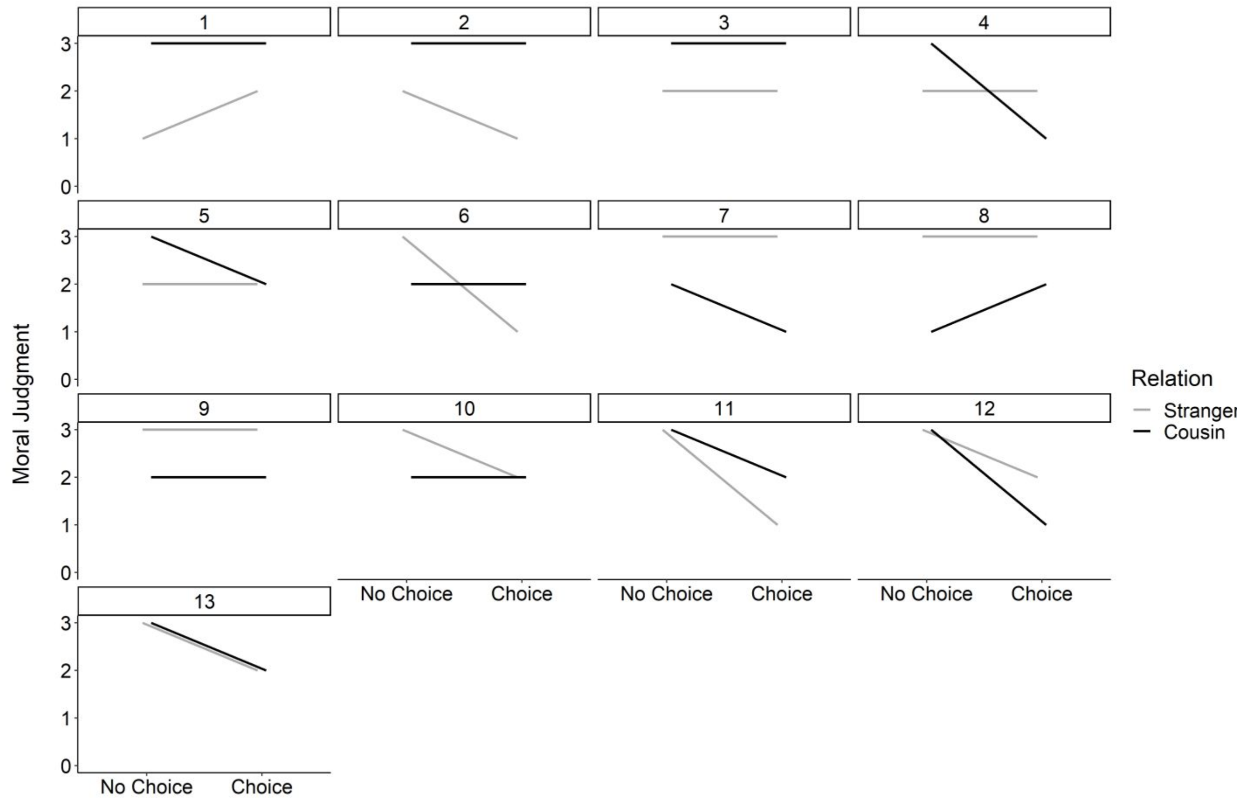

Figure 2. Visualization of Example Hypothetical Participants in McManus et al. (2021). If "Stranger" and "Cousin" lines are not parallel, then an interaction is implied. However, as documented in Table 5, there are multiple interaction patterns that do not match the hypothesized interaction pattern when considering the hypothesized simple effects. Only pattern number 6 is implied by the hypotheses (i.e., "People judge agents who help strangers as more morally good than agents who help a family member, but agents who help a stranger instead of a family member are judged as less morally good than agents who help a family member instead of a stranger").

Table 6. Instructions and Example R Code to Investigate Person-Level Patterns in a 2x2 Design

Step 1	Use wide-formatted data (i.e. 1 row per participant) to create simple effects of interest.	data_wide <- data_wide %>%
		  mutate(SimpleEff1 = A1 - A2) %>%
		  mutate(SimpleEff2 = B1 - B2)

Step 2	Create variables which constitute person-level pattern possibilities.	data_wide <- data_wide %>%
		  mutate(\`2x2_Pattern\` = case_when(
		    (SimpleEff1 == 0 & SimpleEff2 == 0) ~ "Zero, Zero, Zero",
		    (SimpleEff1 == 0 & SimpleEff2 < 0) ~ "Zero, Neg, Pos",
		    (SimpleEff1 == 0 & SimpleEff2 > 0) ~ "Zero, Pos, Neg",
		    (SimpleEff1 < 0 & SimpleEff2 == 0) ~ "Neg, Zero, Neg",
		    (SimpleEff1 < 0 & SimpleEff2 < 0 & SimpleEff1 == SimpleEff2) ~ "Neg, Neg, Zero",
		    (SimpleEff1 < 0 & SimpleEff2 > 0) ~ "Neg, Pos, Neg",
		    (SimpleEff1 < 0 & SimpleEff2 < 0 & SimpleEff1 > SimpleEff2) ~ "Neg, Neg, Pos",
		    (SimpleEff1 < 0 & SimpleEff2 < 0 & SimpleEff1 < SimpleEff2) ~ "Neg, Neg, Neg",
		    (SimpleEff1 > 0 & SimpleEff2 == 0) ~ "Pos, Zero, Pos",
		    (SimpleEff1 > 0 & SimpleEff2 < 0) ~ "Pos, Neg, Pos", # predicted effect
		    (SimpleEff1 > 0 & SimpleEff2 > 0 & SimpleEff1 == SimpleEff2) ~ "Pos, Pos, Zero",
		    (SimpleEff1 > 0 & SimpleEff2 > 0 & SimpleEff1 < SimpleEff2) ~ "Pos, Pos, Neg",
		    (SimpleEff1 > 0 & SimpleEff2 > 0 & SimpleEff1 > SimpleEff2) ~ "Pos, Pos, Pos"))

Step 3	Create person-level tabled data and investigate frequencies of all person-level patterns.	plvl_table <- data_wide %>%
		  group_by(\`2x2_Pattern\`) %>%
		  summarize(freq = n())

Note: The above R code was created using functions from the “tidyverse” package. In Step 2, all text-based patterns reflect the direction of the first simple effect, the second simple effect, and the interaction (e.g., “Zero, Zero, Zero”), in that order.

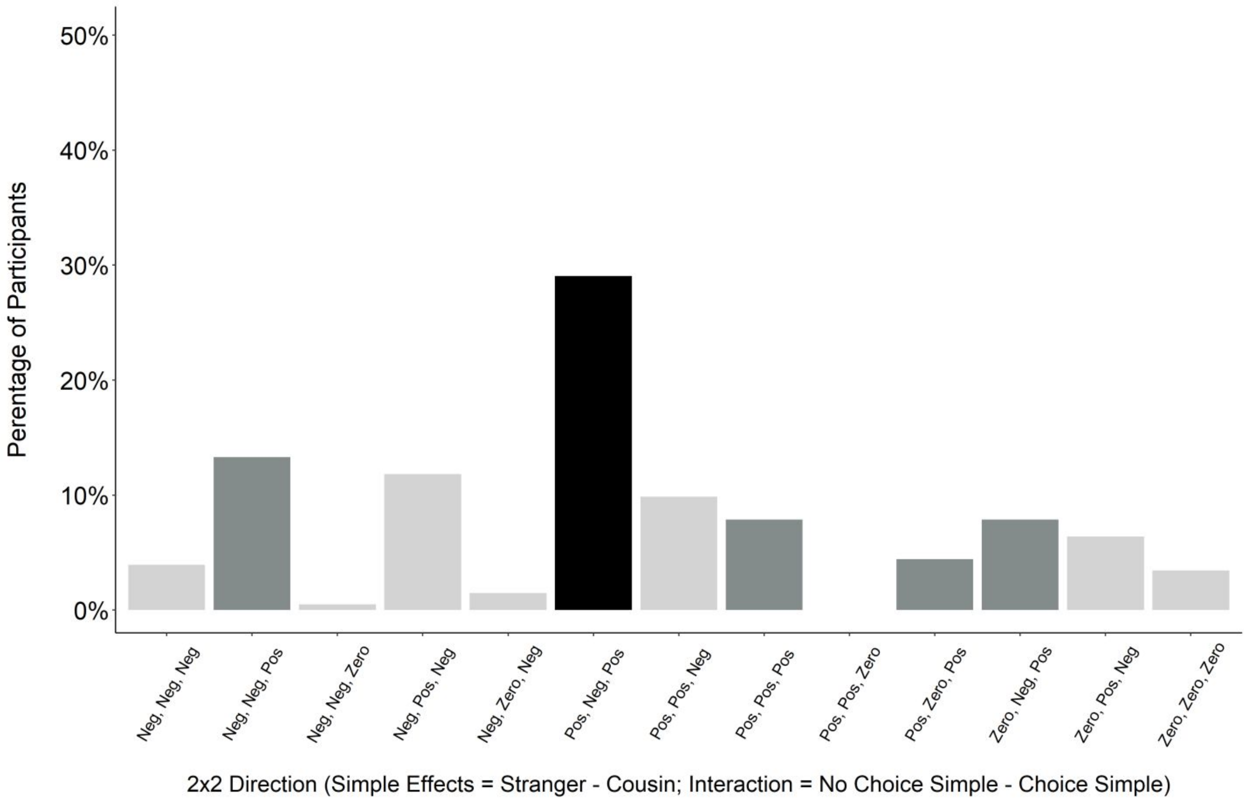

Figure 3. Empirical Person-Level Patterns from McManus, Mason, & Young (2021). Pattern descriptions (e.g., Pos, Neg, Pos) communicate the No Choice difference, Choice difference, and Interaction difference, respectively. The black bar represents the claimed group-level patterns. Dark grey bars represent patterns which also yielded a positive interaction value and therefore contributed to the group-level interaction pattern's emergence. It is noteworthy that this claimed pattern was not even the modal pattern in much of our earlier research (McManus, Kleiman-Weiner, & Young, 2020); however, because we consider our 2021 experiments as better designed, we report only their person-level patterns here.

As shown in Figure 3, ~ 30% of our participants showed the full set of group-level effects. How can this happen? Consider first the crossover interaction. This interaction is typically tested for using a 2x2 repeated-measures ANOVA, as we did. Importantly, the interaction can be assessed using t-tests, which can help to explain the discrepancy. To use the t-test methods, the analyst first creates difference score variables by subtracting the second response from the first response within each simple effect of interest. The paired-samples t-test method is completed by conducting a t-test on the two difference scores. The one-sample t-test method involves an extra step, creating a third difference score variable—the interaction score—by subtracting the second simple effect's difference score from the first simple effect's difference score. The one-sample t-test method is completed by conducting a t-test (against zero) on the interaction scores. If either t-test returns a below-alpha p-value, then an interaction effect exists. Importantly, in this context, the p-value from both t-test methods would be identical to one another and to the p-value of the ANOVA's interaction F-test, as all methods are testing for a difference in differences (see SOM for a demonstration).

Why does this matter? As shown in Table 5 and Figure 2-3, there are five patterns which yield a positive interaction value, only one of which is the claimed pattern[3]. This is problematic considering that the interaction test is simply assessing whether the interaction scores' average differs from zero, nothing more. Therefore, it is possible that more participants had a positive interaction value constituted by the "incorrect" set of simple effects than had a positive interaction value constituted by the "correct" set of simple effects. Indeed, more than 60% of our sample had a positive interaction value that contributed to the group-level interaction test (see Figure 3).

Now consider the opposite-signed simple effects. It is an obvious but crucial point that a person-level claim about the full interaction pattern requires that participants show *both* simple effects. However, what seems non-obvious is that *sets* of typical inferential tests cannot provide this evidence. Because the units of analysis for a single paired-samples t-test are the person-level difference scores, two separate paired-samples t-tests cannot connect units across analyses (and as has already been established, the connection of units via the interaction test has its own problems). The only way to ensure that a particular proportion of participants show both simple effects is to first count how many show each individual pattern. Tabulations of within-person differences showed that the first simple effect described 51% of participants, whereas the second simple effect described 55% of participants. Consequently, the *maximum* proportion of participants who could have shown both patterns was 51%. As established, however, fewer than 30% of participants showed both patterns.

Given this re-analysis and explanation, we suggest that the goal of a psychological experiment should not be to explain a large proportion of variance (e.g., as is often reported in an ANOVA/regression context), but to instead explain a large proportion of persons, as psychology is a property of persons, not averages or distributions. Once this is recognized, psychologists can instead focus on developing and testing causal models which attempt to explain the underlying data generation process happening at the person-level (e.g., Grice, 2015; Grice et al., 2017).

The Problem Worsens (and is Difficult to Fix)!

We believe that we have provided compelling reasoning that person-level hypotheses (common in experimental psychology) should be tested using pervasiveness approaches—tabulating the proportion of participants whose responses match predictions (Grice et al., 2020; Speelman & McGann, 2020). To provide further supporting evidence, we generated hypothetical datasets in which sets of group-level analyses are extremely poor representations of person-level psychology. In these three datasets (each with N = 100), we created 2x2 crossover interactions, 2x2 attenuation interactions, and three-level ordinal effects, all of which yield group-level effects (and survive non-parametric tests) but with none of the participants' scores showing *all* of the relevant effects! For example, in the attenuation interaction dataset (i.e., when two same-direction simple effects emerge that are statistically different in magnitude), even though the interaction and two simple effects emerged at the group-level, not a single participant's scores matched all three effects (see Figure 4, and our SOM for additional examples). We also note that if these existence proofs indeed occurred in the real world, they would void any argument about the usefulness of modal patterns. Although we are unaware of such real-world instances, the theoretical possibility of group-level patterns being perfectly unrepresentative of persons should warrant caution[4].

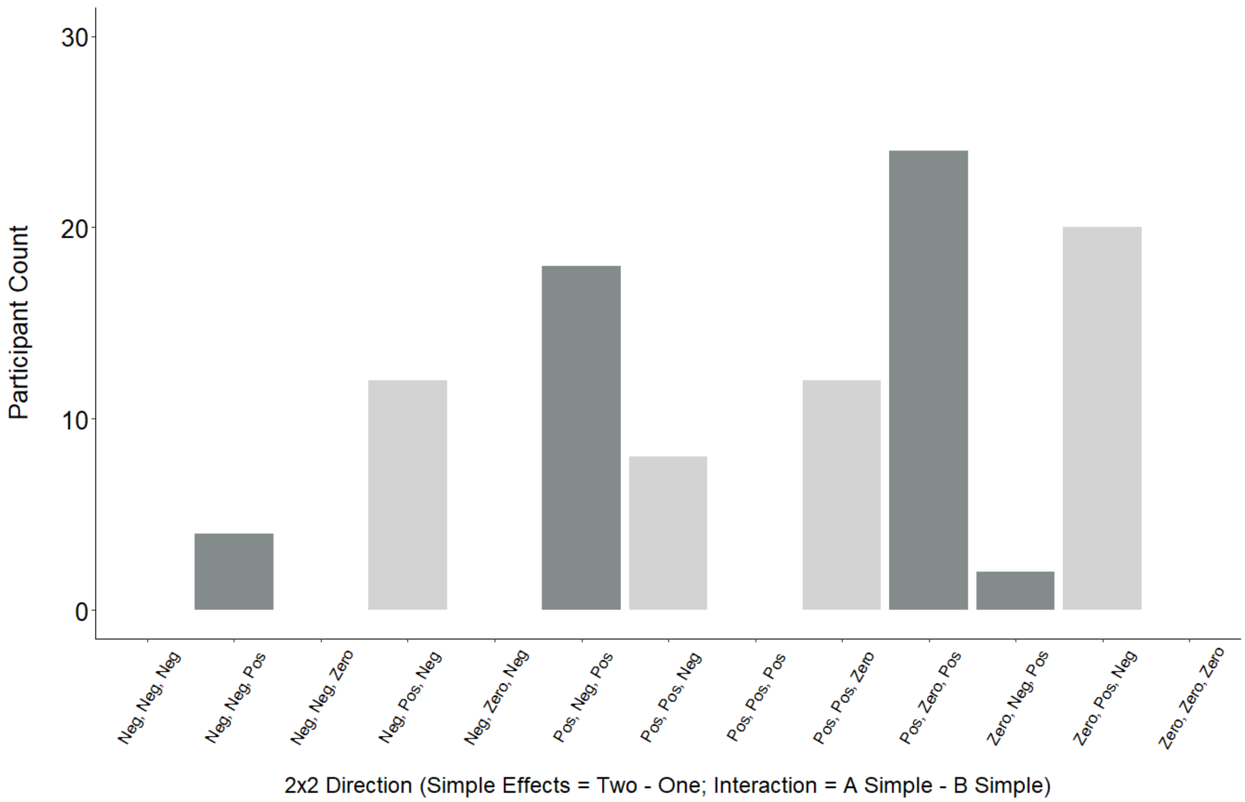

Figure 4. Person-level patterns for A2-A1 and B2-B1 simple effects, and their interaction. Pattern descriptions (e.g., Pos, Neg, Pos) communicate the A difference, B difference, and Interaction difference, respectively. The absent black bar represents the claimed group-level attenuation interaction pattern (i.e., "Pos, Pos, Pos," which describes zero participants here). Dark grey bars represent patterns also yielding an interaction value that contributed to the group-level interaction pattern. See SOM for group-level test statistics and additional examples.

Despite the low proportions found in published research (sometimes as little as 3%; see Table 4), and the existence proofs of group-level patterns being perfectly unrepresentative of persons, it could be argued that most discrepancies between group-level and person-level analyses are due to low measurement reliability and measurement error that can be remedied by appropriate improvements in experimental design. That is, most experiments may not be correctly designed to minimize measurement error and maximize measurement reliability. If strategies to increase reliability and reduce measurement noise were adopted, then group-level patterns may better represent person-level patterns.

As an example, consider the problem of sequential stimulus presentation in typical judgment paradigms. When participants are presented with many stimuli, they are typically presented with one stimulus at a time, after which a judgment is measured. This sequential procedure continues until participants see and respond to all stimuli. This procedure can induce measurement noise in the following way. Some participants might not have judged an early stimulus with the extreme response option if they knew that they would perceive a later stimulus as more extreme; consequently, false ties between stimuli might emerge when participants truly wish to judge them differently. Additionally, this same procedure can lead to some participants forgetting how they made judgments of earlier stimuli, leading to false differences between stimuli that they wished to judge similarly. Therefore, if this kind of noise occurs in typical judgment paradigms (and it is systematically reducing the number of participants who respond in a manner consistent with the predicted group-level effects), participants who have the ability to see all stimuli before making their judgments may be more likely to match the predicted effect.

To address this, using our moral cognition paradigm described in the tutorial above (McManus et al., 2021), we conducted four pre-registered experiments (all similar in spirit to the above description) that systematically varied methodological features hypothesized as reliability and measurement error-related causes of the group-to-person generalizability problem. Across these experiments, we replicated our original group-level effects, as well as the low proportions of participants represented by them (17%-27%). However, none of our experiments was successful in explaining the problem and therefore better aligning person-level and group-level patterns (see Table 7 for a summary of the experiments' logic and results, and SOM for full details). All four experiments were pre-registered at the following links: https://osf.io/wfz3b, https://osf.io/7utrg, https://osf.io/8x69c, and https://osf.io/fcbxe.

Table 7. Underlying Logic and Results for Methodology-Based Experiments (see SOM for full details)

Manipulation	Underlying Logic	Results
Absence/Presence of Calibration Trials	**Problem 1**: If participants do not engage in calibration trials or get feedback about their scale use, then different participants may have different interpretations of identical points along the scale. **Problem 2:** If participants do not engage in calibration trials which are designed to elicit responses along the entire range of the scale, then, when the main task starts, some participants may use extreme ends of the scale for the first stimulus they see, disallowing them from distinguishing between the first stimulus and a later stimulus which they truly wish to judge as more extreme. **Solution:** Before the main experimental task, give participants calibration trials and normative feedback about how most other people use the scale. **Hypothesis:** If the group- versus person-level discrepancy is due to noise of this kind, then participants in an experimental condition (i.e., those who engage in pre-task calibration trials) should be more likely to show the person-level response pattern that matches the group-level pattern, compared to participants in a control condition (i.e., those who do not engage in pre-task calibration trials).	**N per Condition** *N*Control: 658 *N*Experimental: 589 **Predicted Interaction** Control: 24% Experimental: 27% **Eq of Proportions Test** $\chi^2 = 1.17$, $p = .280$ **Hypothesis Decision** Unsupported
Inability/Ability to Respond to Stimuli Simultaneously	**Problem 1:** If participants cannot consider all stimuli simultaneously, then some participants may fail to distinguish between stimuli that they truly wish to distinguish between. **Problem 2:** If participants cannot consider all stimuli simultaneously (and they instead encounter stimuli sequentially), then some participants may use the extreme end of a scale for an early stimulus and be unable to distinguish between it and a later stimulus which they believe is more extreme. **Solution:** Give participants the opportunity to see all stimuli before making any judgments. Then, re-present the important details of all stimuli simultaneously, requesting that participants make any single judgment while considering how they would make their other judgments. **Hypothesis:** If the group- versus person-level discrepancy is due to noise of this kind, then participants in an experimental condition (i.e., those who can see all stimuli and make judgments simultaneously) should	**N per Condition** *N*Control: 628 *N*Experimental: 609 **Predicted Interaction** Control: 24% Experimental: 19% **Eq of Proportions Test** $\chi^2 = 4.65$, $p = .031$

	be more likely to show the person-level response pattern that matches the group-level pattern, compared to participants in a control condition (i.e., those who see stimuli and make judgments sequentially).	**Hypothesis Decision** Unsupported (Wrong direction)
Absence/Presence of Matched Stimuli	**Problem:** If participants respond to stimuli which differ in content across experimental conditions (even if all stimuli variants appear in each condition across the entire sample), then some participants may attend to non-experimental features of stimuli when responding. **Solution:** Give participants matched-in-content stimuli across experimental conditions, varying only the experimental features of interest. **Hypothesis:** If the group- versus person-level discrepancy is due to noise of this kind, then participants in an experimental condition (i.e., those who see perfectly matched stimuli) should be more likely to show the person-level response pattern that matches the group-level pattern, compared to participants in a control condition (i.e., those who see different-in-content stimuli).	**N per Condition** *N*Control: 638 *N*Experimental: 641 **Predicted Interaction** Control: 24% Experimental: 17% **Eq of Proportions Test** $\chi^2 = 10.94$, $p < .001$ **Hypothesis Decision** Unsupported (Wrong Direction)
Inability/Ability to "Opt Out" of using Measures/Scales	**Problem:** If participants do not have the opportunity to "opt out" of using a measurement scale, then some participants' responses may not reflect the construct of interest in exactly the way that researchers intend. For example, participants may not believe a measurement scale captures how they think; therefore, they may actively transform the scale or respond completely randomly. **Solution:** Give participants the ability to opt out of using a measurement scale. **Hypothesis:** If the group- versus person-level discrepancy is due to noise of this kind, then participants in an experimental condition (i.e., of those who have an opportunity to opt out, those who do not) should be more likely to show the person-level response pattern that matches the group-level pattern, compared to participants in a control condition (i.e., those who cannot opt out).	**N per Condition** *N*Control: 746 *N*Experimental: 691 **Predicted Interaction** Control: 22% Experimental: 23% **Eq of Proportions Test** $\chi^2 = 0.09$, $p = .779$ **Hypothesis Decision** Unsupported

Recommendations for Confronting the Group-to-Person Generalizability Problem

Given the group-to-person generalizability problem, what should experimental psychologists do? In this section we propose three easy-to-implement analytic strategies to aid in making person-level claims (see Table 8 for pros and cons of each, and Figure 5 for a simple decision flowchart). Scripts for each strategy are provided at our OSF page: https://osf.io/xyse4/.

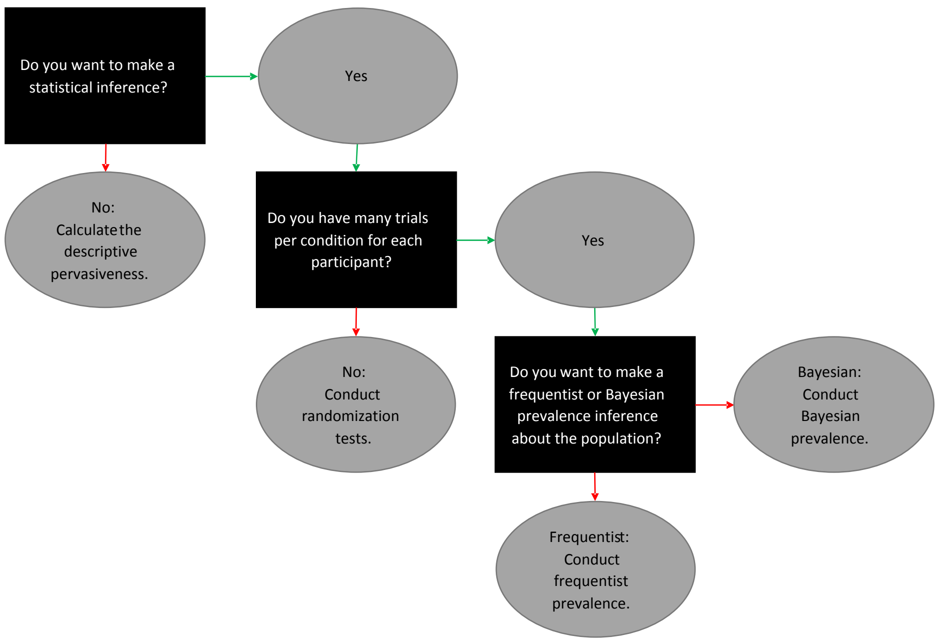

Figure 5. Decision flowchart for investigating proportions. Black boxes represent questions that researchers need to answer, whereas grey ovals represent possible decisions. Red arrows from black boxes to grey ovals indicate that there are no more decisions to be made, but green arrows indicate that there is at least another question and therefore decision to be made.

To further investigate the proportion of people showing predicted effects, researchers can engage in various analytic strategies. First (see the first black box of Figure 5), it must be decided whether a statistical inference is desired. If not, researchers can simply calculate and report the sample proportion's descriptive pervasiveness (see Table 4 and SOM). If, however, researchers want to make a statistical inference, then their next step will depend on whether they have many trials per condition for each participant (see the second black box of Figure 5). If not, researchers can conduct randomization tests, which test whether the predicted effect(s) in the sample is unrelated to experimental condition – i.e., emerges more than “physical chance” (Grice, 2021; Grice et al., 2020; see SOM for an example, as well as an explanation of what constitutes physical chance). This approach has the attractive property that it does not rely on assumptions about populations. Importantly, this approach does not allow an inference from the sample to the population.

If, however, researchers have many trials per condition for each participant, then they can make a population prevalence inference. The prevalence approach combines pervasiveness and within-person approaches to estimate the prevalence of person-level effects in the population (see Allefeld, Görgen, & Haynes, 2016; Donhauser, Florin, & Baillet, 2018; Ince, Kay, & Schyns, 2022; Ince, Paton, Kay, & Schyns, 2021). This is achieved by first conducting typical group-level tests within each person (controlling the false positive rate at the person-level), and second by estimating (using results from the first step) the most likely proportion of people in the population who would show the predicted pattern of effects. Unlike the other approaches (i.e., descriptive pervasiveness and randomization tests), the first step of prevalence approaches test whether qualitative differences between conditions are truly non-zero, assuming measurement error averages out within each person. Importantly, without many trials per condition for each participant, researchers will not be able to make inferences about the population prevalence of their effect, as they would have to assume that (rather than test whether) each person’s pattern reflects true non-zero effects. Prevalence approaches also allow calculation of within-person standardized effects sizes and intervals (see Table 8). This approach allows researchers to test against a “global null hypothesis” of no effect in any subject in the population (H_0: $\theta = \theta_0$ vs. H_1: $\theta \neq \theta_0$; where θ denotes the person-level population proportion and θ_0 a population proportion of 0 or "chance"). The more conservative (and intuitive) "majority null hypothesis" (the effect is in less than, or equal to, half the population; H_0: ≤ .5 vs. H_1: θ > .5) is what we recommend testing if one is intending to make a *general* psychological claim about most people in the population.

Here, researchers can decide whether they desire a frequentist or Bayesian approach to population prevalence (see the third black box of Figure 5), as prevalence inference can be conducted in both the frequentist (see Allefeld, Görgen, & Haynes, 2016; Donhauser, Florin, & Baillet, 2018) and Bayesian (see Ince, Kay, & Schyns, 2022, and Ince, Paton, Kay, & Schyns, 2021, and see SOM for an example) frameworks. In addition to the population prevalence estimate and its precision, the posterior in Bayesian prevalence estimation can be used to compute the probability or log odds that the population proportion is greater than the majority null hypothesis or any theoretically meaningful null hypothesis one deems sufficient for making *general* psychological claims. Because of the advantages of the prevalence approach, we recommend that researchers, if able, begin to adopt high-trial within-subjects designs. When this is not possible, we hope the arguments and options provided here still give researchers the motivation and tools to confront group-to-person generalizability in their own areas of interest. For a walkthrough of how researchers adopting this approach might think through their next experimental design, see our SOM for a detailed summary of how we believe this approach could be applied to our own area of research (McManus et al., 2021).

Table 8. Easy-to-implement analytic strategies to aid in making person-level prevalence claims

Analytic Method	Pros	Cons
Bayesian Prevalence Estimation	• Tests whether qualitative differences between conditions are truly non-zero, assuming measurement error averages out within each person • Allows calculation of person-level standardized effects sizes and intervals • Allows prevalence inferences from samples to populations • Allows calculation of posterior probabilities for specific population prevalence values	• Requires as many observations within each person as typical group-level methods require across persons (holding expected effect sizes constant) • Cannot be applied to all prior (e.g., low-trial) studies • Partially relies on NHST assumptions (for first step)
Frequentist Prevalence Testing	• Tests whether qualitative differences between conditions are truly non-zero, assuming measurement error averages out within each person • Allows calculation of person-level standardized effects sizes and intervals • Allows prevalence inferences from samples to populations	• Requires as many observations within each person as typical group-level methods require across persons (holding expected effect sizes constant) • Cannot be applied to all prior (e.g., low-trial) studies • Fully relies on NHST assumptions • Does not allow calculation of posterior probabilities for specific population prevalence values
Randomization Tests (against physical chance)	• No requirement for total number of observations within persons • Can be applied to all prior (even low-trial) studies • Does not rely on NHST assumptions • Rules out physical chance as an explanation of the sample's proportion	• Assumes qualitative differences between conditions are truly non-zero and error-free • Does not allow calculation of person-level standardized effect sizes and intervals • Does not allow prevalence inferences from samples to populations
Descriptive Pervasiveness	• No requirement for total number of observations within persons • Can be applied to all prior (even low-trial) studies • Does not rely on NHST assumptions	• Assumes qualitative differences between conditions are truly non-zero and error-free • Does not allow calculation of person-level standardized effect sizes and intervals • Does not allow prevalence inferences from samples to populations • Does not rule out physical chance as an explanation of the sample's proportion

General Discussion

Drawing on recent pervasiveness and persons-as-effect-sizes approaches (Grice et al., 2020; Speelman & McGann, 2020), we showed that most laypeople and social psychology researchers interpret psychologists as intending to make claims that represent a majority of their studies' participants. Moreover, most laypeople and researchers believe that this ought to be the case if psychologists are using results to claim support for a general, person-level psychological theory. This paper also documents instances of psychological claims, derived from typical sets of group-level statistical tests, that upon re-analysis are quite poor representations of person-level psychology. As far as we are aware, our work is the first to show that group-level effects in factorial experiments cannot provide the person-level evidence that psychologists likely desire, and that it is possible to have sets of group-level effects that fail to match the response patterns of any single person (see Figure 4 and our SOM). The current research also experimentally tested multiple method-based noise explanations for this group-to-person generalizability problem in a moral judgment paradigm, with obvious remedies proving unsuccessful. Finally, three easy-to-implement analytic strategies were outlined to help researchers confront the group-to-person generalizability problem in their own work and area of interest.

Overall, our research is consistent with recent critiques put forth, in which some researchers (e.g., Richters, 2021; Speelman & McGann, 2020) have argued that there is a pervasive mismatch between psychological theorizing and the analytic procedures used for testing it—typical theorizing occurs at the person-level but analytic procedures operate at the group-level. Over the past decade, much effort has gone toward correcting, and promoting better, statistical inferences (e.g., Lakens, 2021), but relatively fewer reform efforts have been aimed at appropriate psychological (i.e., scientific) inference (e.g., Moeller et al., *preprint;* Navarro, 2019;

Liew, Howe, & Little, 2016) and development of explanatory formal theory (e.g., van Rooij & Baggio, 2021). The current research suggests that even if theorizing indeed improves, inference can still go wrong if familiar group-level statistical methods are privileged over person-level approaches. Put simply, psychologists seem to have put the statistical cart ahead of the psychological horse. This problem, however, should not be judged as just another instance of "psychology in crisis." Instead, this is an opportunity to put past, current, and future research through more stringent tests—to better ground our psychological claims, and the theories they support or challenge, in *persons*.

Potential Objections, Limitations, and Future Directions

In the approach we used throughout this paper (re-analysis of ours and others' data, SOM experiments included), we used any one participant's responses to create a variable that indicated a qualitative directional (e.g., positive) difference between conditions, assuming that this feature was error-free. However, especially in cases when this variable was created from single scores in each condition, it is a fair objection that this qualitative difference cannot be assumed as error-free. The reported proportion estimates may be (extremely) higher or lower depending on how much measurement error played a role in single- and few-trial designs. This problem could be compounded in our own prior research by the fact that we often used many-pointed slider scales to measure constructs of interest. Therefore, it is possible that many participants who we counted as "hypothesis-inconsistent" were indeed "hypothesis-consistent," but our many-pointed sliding measure made it possible to make very small, wrong-direction distinctions between conditions when a participant's intention was to indicate a small, correct-direction distinction. To combat these two problems in future research, we recommend one analytic and one design-based approach.

First, when possible, we suggest using prevalence approaches. We argue that the first step of these approaches combat within-person measurement error in the same way that typical group-level approaches combat across-person measurement error. With large sample sizes, typical group-level approaches (e.g., t-tests) allow near-accurate estimation of population-level mean differences because measurement error is assumed to average out across persons. The first step of these prevalence approaches requires collecting enough person-level data to conduct typical group-level tests *within* each person's data. Therefore, with a large enough trial set, a t-test (or randomization test), for example, can be conducted to compare response scores across conditions within each person; as the logic goes for across-person measurement error, here, measurement error should average out within each person's set of high-N trials. Second, because the scale-point issue remains as another source of error, we also recommend a design-based approach. Specifically, when feasible, researchers could present stimuli/measures that require relative responses (e.g., "Which face is angrier?" with scales ranging from *Face A is much angrier* to *Face B is much angrier*). This might allow researchers to have more confidence in any one trial's difference being a true difference (or non-difference). The number of scale points here likely matters as well, with many-pointed (unmarked and/or sliding) measures likely increasing the number of true non-differences being recorded as small directional differences. This design-based approach should alleviate concerns about scale-based error, but more targeted research is necessary to fully support this recommendation[5].

Another, unrelated objection is that there are other sources of measurement noise accounting for the group-to-person generalizability problem, beyond those tested here (see SOM). For example, some participants are distracted, leading to frequencies of person-level patterns which do not represent the "true" frequencies. First, consistent with our experimental results, there is no reason to believe, if such noise was reduced, that most person-level patterns would conveniently shift to the group-level pattern. Second, as our tutorial and hypothetical datasets show, there are simple non-method explanations for how group-level patterns can be (even perfectly) unrepresentative of persons. Therefore, rather than assuming that there are solvable methodological issues underlying the problem, it should be conceded that person-level patterns cannot be inferred from group-level analyses (see Hamaker, 2012) and therefore the analytic approaches outlined here should be adopted.

One constraint of the pervasiveness and prevalence-based person-level approaches outlined here is that they ignore magnitude information (e.g., the within-person effect size). However, magnitude information can be incorporated into all of these approaches. Researchers can choose an "imprecision value" (Grice et al., 2020), allowing only certain magnitudes to support a qualitative pervasiveness pattern. Additionally, researchers can plot frequencies of qualitative patterns by different imprecision values, allowing discernment between participants who show small versus large effects (see Speelman & McGann, 2020, Figure 4). Similarly, prevalence approaches can consider the prevalence of different effect sizes in the population (Ince, Paton, Kay, & Schyns, 2021).

Relatedly, there are other (potentially better) methods for evaluating person-level effects in high-repetition studies that also yield magnitude information, such as person-level effect sizes and confidence intervals (see e.g., Kurz, Johnson, Kellum, & Willer, 2019, and for incorporating measurement error in N =1 designs specifically, see Schuurman, Houtveen, & Hamaker, 2015). While there are a broad range of powerful, albeit less familiar and technically more challenging, person-level approaches available (for a useful introduction, see Gates, Chow, S. & Molenaar, 2023), we believe the relative strengths of the pervasiveness and prevalence approaches are clear: they require very little statistical knowledge, are easy to implement and interpret (see SOM), and therefore, easy to communicate. We additionally note that prevalence approaches will require drastic changes in data collection practices for some subdisciplines of experimental psychology, as within-person statistical tests would be subject to the same issues that have pervaded the replicability movement (e.g., number of observations and therefore statistical precision/power).

Another limitation of this research is that we used only one moral judgment paradigm to test method-based noise explanations for the group-to-person generalizability problem. Additionally, much research in moral cognition—including our current experiments (see SOM)—utilizes on-the-fly measurement practices (see Flake & Fried, 2020). Future research is needed to determine whether method manipulations fail to remedy the problem in other paradigms and areas of psychology with better measurement practices. However, as shown earlier, there are obvious non-method (and non-measurement) explanations for the problem. Therefore, a person-level approach should still be used in disciplines with better measurement standards to ensure group-to-person generalizability.

We argue that adoption of high-trial per condition experimental designs will allow for better approaches to measurement reliability. For example, researchers with high-trial data can estimate permutation-based split-half reliability, something not possible with single-trial per condition designs (for details, see Parsons, Kruijt, & Fox, 2019). Moreover, high-trial designs also lend themselves to adopting statistical approaches that are aimed at addressing other features of researchers' generalization intentions. For instance, in addition to generalizing from group-to-person, researchers often intend to generalize across other experimental features such as stimuli (Yarkoni, 2020). Future research would do well to examine the relationship between these different forms of generalizability and measurement. As researchers following the various crises in psychological science, we find it exciting that high-trial approaches (along with the appropriate analytic techniques) may offer us a single way of beginning to address many of these challenges.

Finally, we did not assess the ubiquity of the group-to-person generalizability problem. We simply documented (and replicated) existence proofs. We expect the complexity of the experimental designs employed and the phenomenon under investigation will be important in determining the ubiquity of group-to-person generalizability problems. For example, when experiments have factors with more than two levels, or multiple factors, the problem should be more likely to occur because the number of possible person-level patterns explodes as design complexity increases. In contrast, simple binary choice designs common to developmental and comparative psychology may suffer less from the group-to-person generalizability problem. Intuitively the problem seems more likely in higher-level areas like social cognition compared to lower-level areas of inquiry like perception. Presumably this is due to basic shared physiological and neural perceptual mechanisms whereas higher-level cognition may be influenced more by individual differences (e.g., values and knowledge). Additionally, social psychologists in particular are often interested in phenomena that participants do not have introspective access to or are motivated to conceal, leading to the overuse of between-subjects designs rather than the creative use of within-subjects designs (see our SOM for an explanation of how we believe our suggested analysis and measurement approaches could alleviate two typical concerns about the use of within-subjects designs). Therefore, any subdisciplines which habitually rely on between-subjects designs to make inferences about psychology may be especially prone to committing the error of assuming that group-level patterns generalize to the person-level. Ultimately, we suggest that the group-to-person generalizability problem is an issue for any area of psychological research that does not routinely test or model person-level data.

Conclusion

Psychologists often make claims about, and interpret others' claims as being about, person-level processes. Sometimes, however, these claims are made from experiments that disallow investigation of person-level phenomena. Even when such investigation is possible, these claims are typically derived from group-level patterns, interpreted *as if* they reveal truths concerning person-level, psychological phenomenon. The current work confirms and builds upon previous warnings that this practice can lead to serious errors in inference, as (sets of) group-level patterns need not reflect even a simple majority of people in the sample or population. Put simply, psychology is a property of persons, not averages or distributions. Therefore, we should make person-level design and analytic approaches customary in psychological science.

Footnotes

1. In the main text's studies' pre-registrations, we note that the hypothesis sections had many exploratory questions included. Because none of these questions were of primary interest, we do not report them here. However, interested readers can investigate these exploratory questions by referring to our associated RNotebook .html files on OSF.

2. We note that, during the review process, it was argued that due to the one-directional nature of our predictions, we should have used one-tailed tests rather than two-tailed tests. Therefore, the results reported in table format show statistics for one-tailed tests against 0.50, a slight deviation from our pre-registration. The same results hold with two-tailed tests.

3. If the predicted effect is a crossover interaction, this is a special case in which the third "interaction" column is not needed to categorize persons. For example, if a person's first simple effect is positive, and their second simple effect is negative, then that information is enough to categorize the person into the predicted pattern. However, this does not generalize to an attenuation interaction effect. In an attenuation interaction, two persons could have two similar simple effects categorizations (e.g., negative, negative), but differ in how those simple effects differ from one another (e.g., person A has a more negative first simple effect, whereas person B has a more negative second simple effect), leading to different interaction categorizations (negative versus positive).

4. We note that for sets of group-level effects to emerge, at least one or more persons must respond in a manner consistent with at least one of the constituent simple effects; however, as shown, it need not be true that a single person shows *all* constituent simple effects for the set of group-level patterns to emerge.

5. At first glance, this design-based recommendation may seem equivalent to our "simultaneous judgments" intervention (see Table 7, and SOM for full details). However, this recommendation serves a different goal than our intervention served. Specifically, the recommendation to use relative, non-sliding, fewer-pointed scales is to guard against potential error associated with non-relative, sliding, many-pointed scales, so that psychologists can be more confident that any one participant's distinction (or non-distinction) between stimuli is more likely to be a true distinction (or non-distinction). In contrast, our intervention served the purpose of testing whether it was possible to better align person-level patterns with group-level patterns by removing error associated with typical presentation order of stimuli in judgment paradigms.

References

Allefeld, C., Görgen, K., & Haynes, J. D. (2016). Valid population inference for information-based imaging: From the second-level t-test to prevalence inference. Neuroimage, 141, 378-392.

Birnbaum, M.H. (1999). How to show that 9 > 221: Collect judgments in a between-subjects design. *Psychological Methods, 4*(3), 243-249.

Brandt, M.J., & Morgan, G.S. (2022). Between-person methods provide limited insight about within-person belief systems. *Journal of Personality and Social Psychology.*

Craig, B.M., Nelson, N.L., & Dixson, B.J.W. (2019). Sexual selection, agnostic signaling, and the effect of beards on recognition of men's anger displays. *Psychological Science, 30*(5), 728-738.

Cumming, G. (2014). The new statistics: Why and how. *Psychological Science, 25*(1), 7-29.

Decelles, K.A., Adamas, G.S., Howe, H.S., & John, L.K. (2021). Anger damns the innocent. *Psychological Science, 32*(8), 1214-1226.

Deska, J.C., Kuntsman, J., Lloyd, P.E., Almaraz, S.M., Bernstein, M.J., Gonzales, J.P., & Hugenberg, K. (2020). Race-based biases in judgments of social pain. *Journal of Experimental Social Psychology, 88,* 103964.

Donhauser, P. W., Florin, E., & Baillet, S. (2018). Imaging of neural oscillations with embedded inferential and group prevalence statistics. PLoS computational biology, 14(2), e1005990.

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, 115*(27), E6106-E6115.

Flake, J.K., & Fried, E.I. (2020). Measurement schmeasurement: Questionable measurement practices and how to avoid them. *Advances in Methods and Practices in Psychological Science, 3*(4), 456-465.

Fowler, Z., Law, K.F., & Gaesser, B. (2021). Against empathy bias: The moral value of equitable empathy. *Psychological Science, 32*(5), 766-779.

Friston, K. J., Holmes, A. P., Worsley, K. J., Poline, J. P., Frith, C. D., & Frackowiak, R. S. (1994). Statistical parametric maps in functional imaging: a general linear approach. *Human brain mapping*, *2*(4), 189-210.

Galton, F. (1907). Vox populi. *Nature, 75,* 450-451.

Gates, K. M., Chow, S. M., & Molenaar, P. C. (2023). *Intensive longitudinal analysis of human processes*. CRC Press.

Grice, J.W. (2021). Drawing inferences from randomization tests. *Personality and Individual Differences, 179,* 110963.

Grice, J.W. (2015). From means and variances to patterns and persons. *Frontiers in Psychology, 6,* 1007.

Grice, J.W., Barrett, P., Cota, L., Felix, C., Taylor, Z., Garner, S., Medellin, E., & Vest, A. (2017). Four bad habits of modern psychologists. *Behavioral Sciences, 7*(3), 1-21.

Grice, J.W., Medellin, E., Jones, I., Horvath, S., McDaniel, H., O'lansen, C., & Baker, M. (2020). Persons as effect sizes. *Advances in Methods and Practices in Psychological Science, 3*(4), 443-455.

Hamaker, E. (2012). Why researchers should think "within-person": A paradigmatic rationale. In M.R. Mehl & T.S. Conner (Eds.). *Handbook of Research Methods for Studying Daily Life,* 43-61, NY, NY: Guilford.

Ince, R, A.A., Kay, J.W., & Schyns, P.G. (2021). Within-participant statistics for cognitive science. *Trends in Cognitive Sciences, 26*(8), 626-630.

Ince, R, A.A., Paton, A.T., Kay, J.W., & Schyns, P.G. (2021). Bayesian inference of population prevalence. *eLife, 10,* e62461.

Lakens, D. (2021). The practical alternative to the p-value is the correctly used p-value. *Perspectives on Psychological Science, 16*(3), 639-648.

Law, K.F., Campbell, D., & Gaesser, B. (2021). Biased benevolence: The perceived morality of effective altruism across social distance. *Personality and Social Psychological Bulletin, 48*(3), 426-444.

Liew, S.H., Howe, P.D.L., & Little, D.R. (2016). The appropriacy of averaging in the study of context effects. *Psychonomic Bulletin and Review, 23*(5), 1639-1646.

Kievit, R.A., Frankenhuis, W.E., Waldorp, L.J., & Borsboom, D. (2013). Simpson's paradox in psychological science: A practical guide. *Frontiers in Psychology, 4,* 513.

Kuppens, T. Pollet, T.V. (2014). Mind the level: Problems with two recent national-level analyses in psychology. *Frontiers in Psychology, 5,* 1110.

Kurz, A.S., Johnson, Y.L., Kellum, K.K., & Wilson, K.G. (2019). How can process-based researchers bridge the gap between individuals and groups? Discover the dynamic p-technique. *Journal of Contextual Behavioral Science, 13,* 60-65.

McManus, R.M., Mason, J.E., Young, L. (2021). Re-examining the role of family relationships in structuring perceived helping obligations, and their impact on moral evaluation. *Journal of Experimental Social Psychology, 96,* 104182.

McManus, R.M., Kleiman-Weiner, M., & Young, L. (2020). What we owe to family: The impact of special obligations on moral judgment. *Psychological Science, 31*(3), 227-242.

Moeller, J. (2022). Averting the next credibility crisis in psychological science. Within-person methods for personalized diagnostic and intervention. *Journal for Person-Oriented Research, 7*(2), 53-77.

Moeller, J. et al. (*preprint*). Generalizability crisis meets heterogeneity revolution: Determining under which boundary conditions findings replicate and generalize.

Navarro, D.J. (2019). Between the Devil and the Deep Blue Sea: Tensions between scientific judgment and statistical model selection. *Computational Brain and Behavior, 2*(1), 28-34.

Parsons, S., Kruijt, A. W., & Fox, E. (2019). Psychological science needs a standard practice of reporting the reliability of cognitive-behavioral measurements. *Advances in Methods and Practices in Psychological Science*, *2*(4), 378-395.

Quintana, D.S. (2021). Towards better hypothesis tests in oxytocin research: Evaluating the validity of auxiliary assumptions. *Psychoneuroendocrinology,* 105642.

Richters, J.E. (2021). Incredible utility: The lost causes and causal debris of psychological science. *Basic and Applied Social Psychology, 43*(6), 366-405.

Rottman, J., & Young, L. (2019). Specks of dirt and tons of pain: Dosage distinguishes impurity from harm. *Psychological Science, 30*(8), 1151-1160.

Schuurman, N.K., Houtveen, J.H., & Hamaker, E.L. (2015). Incorporating measurement error in n = 1 psychological autoregressive modeling. *Frontiers in Psychology, 6,* 1038.

Simpson, E.H. (1951). The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society. Series B (Methodological), 13*(2), 238-241.

Soter, L.K., Berg, M.K., Gelman, S.A., & Kross, E. (2021). What we would (but shouldn't) do for those we love: Universalism versus partiality in responding to others' moral transgressions. *Cognition, 217,* 104886.

Speelman, C.P., & McGann, M. (2020). Statements about the pervasiveness of behavior require data about the pervasiveness of behavior. *Frontiers in Psychology, 11,* 1-16.

Stroessner, S.J., Benitez, J., Perez, M.A., Wyman, A.B., Carpinella, C., Johnson, K.L. (2020). What's in a shape? Evidence of gender category associations with basic forms. *Journal of Experimental Social Psychology, 87,* 103915.

Surowiecki, J. (2005). *The wisdom of crowds.*

Thai, M., Borgella, A.M., & Sanchez, M.S. (2019). It's only funny if we say it: Disparagement humor is better if it originates from a member of the group being disparaged. *Journal of Experimental Social Psychology, 85,* 103838.

Van Rooij, I., & Baggio, G. (2021). Theory before the test. How to build high-verisimilitude explanatory theories in psychological science. *Perspectives on Psychological Science, 16*(4), 682-697.

Wallis, K.F. (2014). Revisiting Francis Galton's forecasting competition. *Statistical Science, 29*(3), 420-424.

Whitsett, D.D., & Shoda, Y. (2014). An approach to test for individual differences in the effects of situations without using moderator variables. *Journal of Experimental Social Psychology, 50*(1), 94-104.

Yarkoni, T. (2020). The generalizability crisis. *Behaviorial and Brain Sciences, 45,* E1.