Psychology is a Feature of Persons, Not Averages or Distributions: 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, 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 small) minority of participants. Third, 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 describe *zero* participants. Fourth, we conduct four experiments that rule out several methodology-based noise explanations of the problem. Fifth, we survey psychology 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. Finally, we propose that, if experimental psychologists are indeed interested in person-level psychology, then they should deploy different analytic strategies from those typically used. Overall, our approach offers a simple and flexible method to help researchers begin to engage with person-level analysis.

Introduction

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, this problem of group-to-person generalizability plagues modern-day experimental psychology. 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 and describe *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 argue for closely matching experimental designs and analytic methods to precise research questions. Second, we document instances in published literature in which a person-level analytic approach yields different conclusions than typical group-level approaches. Third, in a tutorial, we show readers how this can occur, and how to conduct person-level analyses on their own data. Additionally, we demonstrate how claims from sets of simulated group-level effects can describe zero persons.

Fourth, we conduct four pre-registered experiments to rule out several methodology-based explanations of group-to-person generalizability failures. Fifth, we survey laypeople and psychology researchers to understand what is inferred about person-level phenomena from group-level analyses. Finally, we argue that experimental psychologists, if interested in person-level psychology, ought to deploy different design and analytic strategies than those typically used.

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

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 it is minds that possess psychological processes, and minds reside in persons. To strengthen this argument, we ask readers to engage in a thought exercise. Recall your most recent meeting with collaborators in which 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 most experimental psychologists intend to uncover phenomena 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 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" hypothesis or result, and we then clarify our position on designing studies to test such hypotheses or yield such results.

What is a "Person-Level" Result?

Specifically, a "person-level" result is one which provides information about a pattern of responding or activity for a single person (e.g., an experimental effect direction and magnitude for person X). If there are many experimental trials or "intensive" sampling in a longitudinal design, then one can compute single-subject statistics (e.g., effect sizes, confidence intervals, factor structures, etc.) in order to make person-level inferences (e.g., Kurz, Johnson, Kellum, & Willer, 2019). Alternatively, once all person-level patterns are investigated and counted, a "pervasiveness" estimate can be obtained (Speelman & McGann, 2020). By "pervasiveness," we mean the choosing of one possible person-level pattern and investigating, "How many persons match this pattern?" Finally, if the chosen person-level pattern is the one that was predicted or claimed, then that person-level pattern's pervasiveness can be tested against a value considered to be theoretically meaningful (e.g., 50% if one is intending to test or make a *general* psychological claim). This pervasiveness approach has the advantage of not requiring a high level of statistical expertise or a large number of observations.

Within-Subjects (vs. Between-Subjects) Designs for 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 respond this way?" (see Speelman & McGann, 2020), and they prohibit examination of unfolding person-level processes (e.g., Brandt & Morgan, 2022; Fisher, Medaglia, & Jeronimus, 2018; Moeller, 2022). For example, consider the following research question: "Do people prefer Coca-Cola to Mountain Dew?" To assess this, the leading soda cognition lab designs an experiment which randomly assigns half of participants to rate their likelihood of buying Coca-Cola in the next month, and the remaining participants to rate Mountain Dew in the same way. They analyze the resulting data with an independent-samples t-test, finding that the average likelihood judgment is higher for the Coca-Cola condition. However, a rival soda cognition lab also attempts to answer this question, but they use a within-subjects design and find an average difference in the opposite direction. Assuming the within-subjects effect generalizes to the person-level, which of these designs better answers the question, "Do people prefer Coca-Cola to Mountain Dew?" If *prefer* implies a psychological comparison, we suggest that the within-subjects design reigns supreme. Moreover, there are many plausible non-substantive mechanisms for the between-subjects results (e.g., for participants who rated Coca-Cola, the counterfactual soda of Pepsi was referenced).

To further illustrate this, Birnbaum (1999) showed that "People judge 9 as larger than 221" can be derived 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 that no serious experimentalist would interpret these results as people truly believing 9 is larger than 221 (and we again note that "judge… as larger than…" implies a psychological comparison). Of course, Birnbaum's methods could have been better (e.g., measuring "largeness" on a true ratio scale), but we argue that

criticizing his results on methodological grounds is only possible because we all believe that there is a truth of the matter, with better and worse ways of verifying it. In most psychological experiments, however, the truth of the matter is *inferred* from the data, meaning 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 (see also Speelman & McGann, 2020). 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.

To clarify, we are not suggesting that between-subjects are never useful. These designs may be preferable when within-subjects designs are impossible or practically infeasible. For example, many intervention(-like) research questions may be best answered with between-subjects designs. Additionally, hypotheses about population(-like) differences require at least some component of the study to be a between-subjects factor, such as testing whether psychopaths show different experimental effects than non-psychopaths. Similarly, if the primary object of study is not the psychological process/outcome itself (e.g., the variability of responses as a function of some manipulated ability), a between-subjects design may be most feasible (see our SOM experiments). Finally, between-subjects designs are unproblematic when the research goal is to provide generalization evidence (e.g., finding similar effects across stimulus sets; 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). We argue that within-subjects designs at least offer researcher the opportunity to engage with person-

level inferences. 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, Mason, & Young, 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. 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 no 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 respond this way?" A between-subjects design would have disallowed such investigation.

However, having within-subjects designs does not automatically prevent 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.

Overall, we are suggesting that, if a theory or research question 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 analytic procedures. However, such careful matching does not always occur in practice. The rest of this paper focuses on instances in which within-subjects group-

level effects fail to describe the majority of sampled persons. From here on, we refer to this as the "group-to-person generalizability problem."

Group-to-Person Generalizability Problems in the Wild

We examined open data from the past five years in psychological research (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 problem." Rather, using a person-level analytic 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., race, gender, humor, etc.; see Table 1). Anecdotally, many seemingly person-level claims (not reported here) were made from between-subjects designs that fundamentally disallowed reanalysis. We also note that 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 cognitive research on visual perception of faces (Grice et al., 2020) and context effects in decision-making (Liew et al., 2016).

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) and were directionally consistent with a group-level pattern, then that participant was categorized as

supporting group-to-person generalizability. Because some measures were averages of multiple items, this meant that, in practice, we counted participants as supporting generalizability even if their experimental effects were near- but not exactly zero. Therefore, we used an extremely liberal threshold.

An important nuance is that all of the investigated claims are based on *sets* of group-level tests (e.g., multiple paired t-tests). We therefore extended the 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 underlied 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. Readers can imagine (and if they wish, investigate) what these analyses look like under stricter constraints (see our OSF page: https://osf.io/xyse4/).

We statistically defined claims as being unsupported at the person-level if we could not rule out the possibility that fewer than a simple majority of people would show the group-level pattern, conducting a binomial test against 0.50. If the 95% CIs contained a proportion of 0.50 or lower, then the claim was unsupported at the person-level. We chose this 0.50 proportion as the null value because most claims in psychology articles do not use language that suggests an experimental effect is one that describes only a (minority) subset of participants. This means that, at least by implication, the effect is being communicated as applying to *most* participants (something we test empirically in *An Important Objection*). Therefore, testing against this null value, too, is a liberal threshold for categorizing claims as favoring generalizability (i.e., a simple majority). As Table 1 shows, proportions of participants favoring generalizability varied across

these publications but was low overall (3%-50%, with most proportions ranging between 20%-40%). Importantly, 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). We note that authors of these studies may not believe that their claims are describing most of their participants (although see *An Important Objection*). However, descriptions of results are, at the least, ambiguous enough to warrant uncertainty. We next break down a moral cognition example showing how the group-to-person generalizability problem can occur.

Table 1. Quotes, relevant tests, and person-level statistics 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% [34%-37%]** (62 / 203) E1b: **29% [23%-36%]** (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% [1%-9%]** (3 / 97) E4: **8% [5%-11%]** (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% [27%-37%]** (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% [25%-34%]** (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% [22%-36%]** (51 / 177) E2: **46% [35%-57%]** (37 / 81) E3: **22% [16%-29%]** (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% [42%-59%]** (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%-50%]** (26 / 69) E4: **41% [33%-49%]** (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% [38%-52%]** (99 / 219) Accuracy: **25% [20%-31%]** (55 / 219) Both: **13% [9%-18%]** (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% [30%-47%]** (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% [33%-57%]** (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 places 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.

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

Here we demonstrate how researchers can understand and perform analyses to make person-level inferences in within-subjects experiments. 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 helps an unrelated stranger *instead of* their cousin, and someone helps 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 a majority of participants. 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 family member" ratings from the "helped a stranger" ratings, whereas the Choice simple effect can be computed by subtracting the "helped a family member instead of a stranger" ratings from the "helped a stranger instead of a family member ratings." An interaction effect can then be computed by subtracting the Choice effect from the No Choice effect (see Table 2 for an example of 13 hypothetical participants who reflect all possible qualitative patterns, and Table 3 for example R code to create generalizable 2x2 person-level patterns, investigate their frequencies, and conduct a binomial test). The person-level combination 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 2. Example hypothetical participants, showing all possible patterns in McManus et al. (2021)

Subj	NC_Stranger	NC_Cousin	C_Stranger	C_Cousin	NC_Diff	C_Diff	Interaction		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 "Interaction" 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). Moreover, as the number of scale points increases, there are more possible ways to make similar-in-direction distinctions that differ in magnitude. Therefore, this problem may be more pervasive as the number of possible response options increases. Importantly, these patterns do not consider other features of interaction patterns, such as the rank-ordering of all four conditions on the numerical response scale.

Table 3. 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	Investigate frequencies of all person-level patterns.	data_wide %>%
		  group_by(`2x2_Pattern`) %>%
		  summarize(freq = n())

Step 4	Test the predicted effect's frequency against a meaningful proportion value, using a binomial test.	binom.test(x = *Predicted Effect Freq*, n = *Total N*, p = *Prop Value*, alternative = "two.sided")

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.

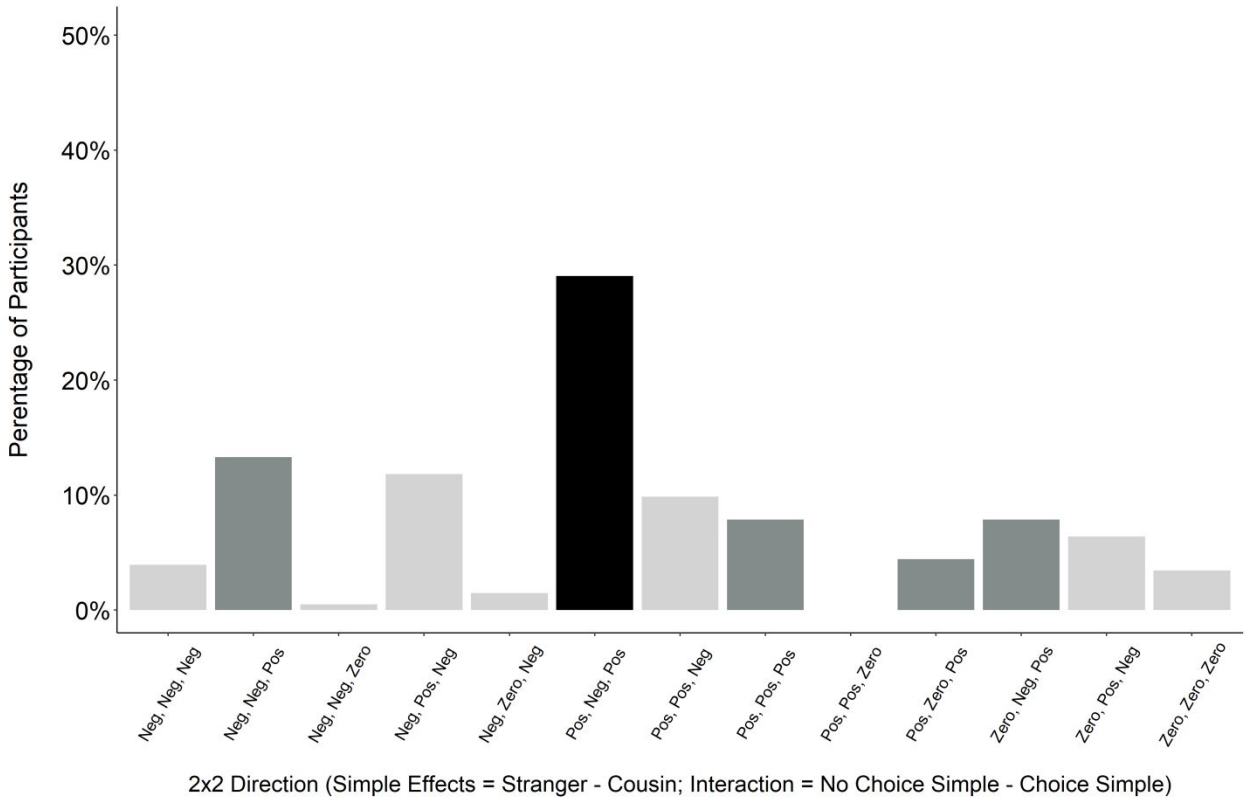

Figure 1. Person-Level Patterns from McManus, Mason, & Young (2021). Pattern descriptions (e.g., Pos, Neg, Pos) communicate the No Choice difference, Choice difference, and Interaction difference, respectively. The black bar represents the claimed group-level patterns. Dark grey bars represent patterns which also yielded a positive interaction value and therefore contributed to the group-level interaction pattern's emergence.

As shown in Figure 1, less than 30% of our participants show 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 2, there are five patterns which yield a positive interaction value, only one of which is the claimed pattern[1]. 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 1).

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. The only way to ensure that a particular proportion of participants show both group-level patterns 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 has been 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 psychological experiences occur within (not across) persons. Once this is realized, psychologists can instead focus on developing and testing causal models which attempt to explain the underlying data generation process (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 claims need to be tested using persons-as-effect-sizes or 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 cognition. In these datasets, 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 describe *zero* participants (see SOM). Although we are unaware of real-world instances, the theoretical possibility of group-level patterns being perfectly unrepresentative of persons should warrant caution[2].

Despite these existence proofs, it could be argued that most discrepancies between group-level and person-level analyses are due to methodological features of experiments which can be remedied. That is, most experiments may not be designed to minimize noise and therefore maximize the probability of persons exhibiting the group-level pattern. If such barriers could be

addressed, then group-level patterns may better represent person-level patterns. To address this, using our moral cognition paradigm described earlier, we conducted four pre-registered experiments which systematically varied methodological features hypothesized as partial, noise-inducing causes of the group-to-person generalizability problem.

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 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, e.g., differences between stimuli that they truly wish to judge similarly. Therefore, if this kind of noise is operant 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. Within four experiments of this spirit, we replicated our original group-level effects, as well as the low proportions of participants represented by them (17%-27%; see also Devezer et al. [2021] for a discussion on how replicability need not imply "true"). However, none of our experiments were successful in explaining the problem and therefore shifting person-level patterns to be better aligned with the group-level pattern (see Table 4 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 4. 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** $\chi^2 = 4.65$, $p = .031$ (Wrong direction)

	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
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

(Empirically Addressing) An Important Objection

We have argued that there is a group-to-person generalizability problem in psychology, documenting published instances of it, showing how it can occur, demonstrating its potential severity, and its resistance to obvious method-based remedies. However, there is obvious subjectivity involved when deciding what should count as person-level evidence for a claim. For example, many claims that we viewed as instances of the group-to-person generalizability problem (see Table 1) may seem unproblematic to other researchers. It could be argued that percentages of participants in the 20-40% range, who show the group-level patterns, are quite high. Moreover, perhaps readers of psychology research (laypeople and psychology researchers themselves) do not interpret authors as intending to make claims that represent at least a majority of 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 if 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 social psychology researchers by presenting modified excerpts of "results" and "general discussion" sections from publications that contain unrepresentative group-level patterns. 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 an effect from the supplemental materials of Law, Campbell, & Gaesser (2021). For the complex effect condition, participants learned about the 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 2, however, these 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 people (i.e., 51% or more) within each sample (laypeople and researchers) 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 within each sample 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[3] (see Table 5 for additional descriptive statistics and Table 6 for inferential statistics). Strikingly, as shown in Figure 2, 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).

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 5 for additional descriptive statistics and Table 7 for inferential statistics). As shown in Figure 2, there again is no discernible pattern as a function of research experience[4].

Table 5. Descriptive Statistics for Empirical and Theoretical Estimates (Split by Population)

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

Table 6. Empirical Estimate Tests within Each Effect Type (split by Population)

Effect Type	Population	Proportion	*p-value*
Simple	*Laypeople*	81% [76% - 85%]	< .001
	Researchers	73% [64% - 81%]	< .001
Complex	*Laypeople*	88% [84% - 92%]	< .001
	Researchers	80% [72% - 87%]	< .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 95% CIs for the proportion estimate. P-values were computed via binomial tests against 0.50.

Table 7. Theoretical Estimate Tests within Each Effect Type (split by Population)

Effect Type	Population	Proportion	*p-value*
Simple	*Laypeople*	93% [90% - 96%]	< .001
	Researchers	80% [72% - 87%]	< .001
Complex	*Laypeople*	92% [89% - 95%]	< .001
	Researchers	90% [83% - 95%]	< .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 95% CIs for the proportion estimate. P-values were computed via binomial tests against 0.50.

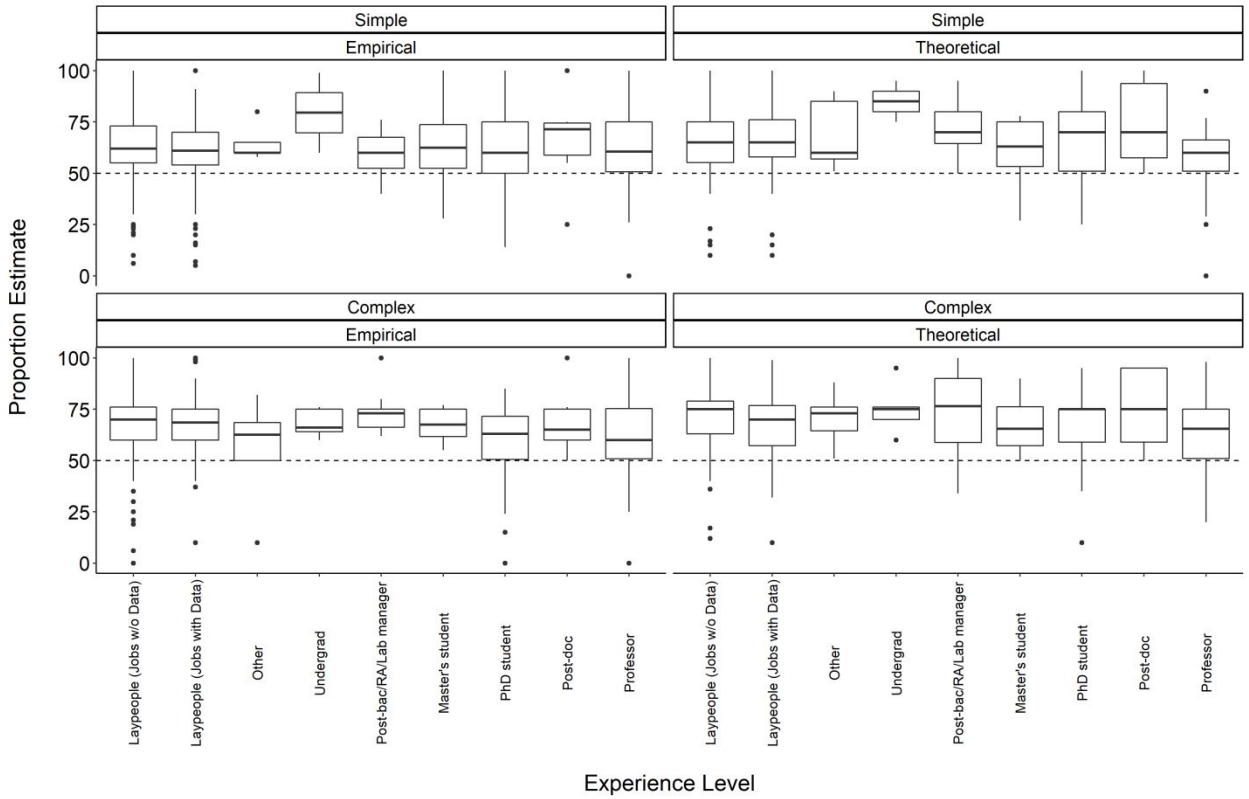

Figure 2. Boxplots of empirical/theoretical proportion estimates by effect type (simple versus complex), and by participants' level of experience. Note that "Other" refers to people involved in academic research in some way (via SPSP) but who indicated that they have never held an academic position.

General Discussion

Drawing on recent "pervasiveness" and "persons-as-effect-sizes" approaches (Grice et al., 2020; Speelman & McGann, 2020), the current work 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 showing that group-level effects in multi-factor experiments cannot provide the person-level evidence that psychologists likely desire (i.e., "Do people respond this way, and if so, how many?"), and that it is possible to have sets of associated experimental effects which describe zero persons. Additionally, the current research 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, our research shows that a majority of laypeople and social psychology researchers interpret authors of psychology articles as intending to make claims that represent a majority of their study's participants. Moreover, a majority of laypeople and researchers believe that this ought to be the case if authors are using a study's results to claim support for a general, person-level psychological theory.

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—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 inference (e.g., Moeller et al., *preprint;* Navarro, 2019; Liew, Howe, & Little, 2016) and proper theory development (e.g., van Rooij & Baggio, 2021). The current research suggests that even if theorizing indeed improves, inference can still go wrong if familiar statistical methods are privileged over ones that address specific psychological questions. 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 in *persons*.

Responses to Potential Objections

One objection is that, if the modal person-level pattern matches the group-level pattern, then there is no group-to-person generalizability problem. We suggest, however, that if the

modal person-level pattern matches the group-level pattern but is not *also* the pattern describing most participants, then the majority of responses would be unexplained. Moreover, most polled laypeople and social psychology researchers agreed that for a claim to support a theory, it ought to represent a majority of participants. Extrapolating from this, it stands to reason that *any* reported effect that does not describe a majority of participants (regardless of its use for theory) ought to be tagged as unrepresentative.

A related objection is that the proportion of participants showing the group-level pattern may be above chance given how many possible qualitative patterns exist, meaning that there is no group-to-person generalizability problem. For example, in a simple two-cell design, there are only three possible person-level effects (i.e., positive, negative, or zero). Based on this objection, if any of those effects described more than 33% of the sample and was consistent with the group-level pattern, then the study would not contain the group-to-person generalizability problem. Our reasoning against the first objection also applies here: if the modal pattern does not also describe most participants, then the group-level pattern represents only a minority of responses. In addition, there is unfortunately no ground truth to a "better than chance" criterion. That is, designs could be modified to include additional conditions that are likely to engender uniform responses across participants, creating more possible qualitative patterns while artificially reducing the chance of any single pattern. We note that if psychologists are truly *only* interested in testing a person-level pattern's pervasiveness against "randomness" or "chance," there are well-established methods for doing so (i.e., randomization tests, Grice, 2021; but see our SOM for how to conduct proportion tests against chance to make sample-to-population inferences for Table 1's claims).

A final objection is that there are other sources of 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 accepted that person-level patterns cannot be inferred from group-level analyses.

Recommendations

Given the group-to-person generalizability problem, what should experimental psychologists do? Our recommendations are consistent with those in a recent critique (Yarkoni, 2020). We note that our generalizability critique refers to generalizing across levels of analysis, whereas the recent critique refers to generalizing across stimuli, tasks, etc. Specifically, we propose predicting specific orderings of observations based on theory (e.g., A1 higher than A2 in B1, but A2 higher than A1 in B2), while specifying a proportion of participants whose responses should match predictions for the theory-derived hypothesis to survive. To conduct a "severe test" (Mayo, 2018) or corroborate a "risky prediction," (Meehl, 1990a, 1990b), the empirical proportion should be close to the theory-predicted proportion, and other theories should not predict this proportion.

Unfortunately, there may not be many psychological theories (especially in [social] cognition which typically lacks formal models) that can make such predictions (for examples and

discussion of this issue see, Crockett, 2016; Hamlin et al., 2013). However, theoretical progress can still be made. To test face validity, researchers can make minimum proportion predictions. To do this, researchers can tabulate the proportion of participants whose responses match a predicted pattern. They can then make a frequentist inference by testing this proportion against a majority null (i.e., 50%) using a binomial test. We *strongly* recommend supplementing the binomial testing approach with an estimation approach (e.g., Cumming, 2014) in which the estimated intervals are interpreted to rule out certain proportions. To be clear, we do not advocate 50% as the benchmark against which psychologists should test theory; we are simply suggesting a method that enables researchers to identify evidence that fails to support general psychological regularities (Hamaker, 2012). Others have suggested even higher proportions as convincing evidence (e.g., 80%; Speelman & McGann, 2020), though responses from the current work suggest that psychologists disagree about the appropriate cutoff. We also do not recommend ignoring theory-inconsistent patterns, or patterns represented by a small minority of participants. Understanding if and why other patterns exist allows refinement of theory by postulating and testing whether there are substantive moderating variables (e.g., individual differences), or simple violations of auxiliary assumptions (e.g., divergent interpretations of measures; see Quintana, 2021, for a discussion).

Limitations and Future Directions

One constraint of this person-level approach is that it ignores magnitude information (e.g., participants who use two extreme ends of a measure are treated identically to participants who use two close points of a measure). However, magnitude information can be incorporated into this approach. Researchers can choose an "imprecision value" (Grice et al., 2020), allowing only certain magnitudes to support a qualitative 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).

Relatedly, there are other (potentially better) methods for evaluating person-level effects in high-repetition studies that yield magnitude information, such as person-level standardized effect sizes and confidence intervals (e.g., Kurz, Johnson, Kellum, & Willer, 2019; Schuurman, Houtveen, & Hamaker, 2015). Additionally, recent research suggests first testing effects within each person (using typical statistical approaches at the person-level) and then using those tests to calculate a Bayesian "prevalence" estimate (Ince, Kay, & Schyns, 2022; Ince et al., 2021). These methods are not unlike our suggested approach of using a binomial test and its estimate, as they both allow for inferences from samples to populations, a feature unsupported in some other approaches (e.g., randomization tests). However, the intensive sampling that all these methods require are often infeasible in experimental psychology research, as person-level statistical tests would be subject to the same issues that have pervaded the replicability movement (e.g., number of observations and therefore statistical power). Therefore, unless all experimental psychologists start designing high-repetition studies, our proposal is the only method available, to our knowledge, that will consistently allow person-level pervasiveness claims to be generalized from samples to populations. Importantly then, our method is also the most generalizable, as it can also be used for high-repetition studies, single-response-per-condition studies, and anything in between[5]. In repetition designs, an analyst can simply average over a person's multiple responses within each condition and then apply the person-level analysis to those averages. This suggestion is not inconsistent with our argument to avoid making inferences from averages, as the suggested averaging would occur within rather than across persons. Other strengths of our approach are

clear: it requires no advanced statistical knowledge, is easy to implement and interpret, and therefore, is easy to communicate.

Another limitation 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 more rigorous disciplines to ensure generalizability. Relatedly, we note that our proposed method does not solve the problem of measurement error. Therefore, as suggested elsewhere (see Speelman & McGann, 2020), future research would benefit from collecting many trials per person in order to get observed person-level scores that better approximate the true person-level scores.

Finally, we did not assess the ubiquity of the group-to-person generalizability problem. We simply documented (and replicated) existence proofs. Documenting its ubiquity is a necessary next step to examine its generalizability (see Simons, Shoda, & Lindsay, 2017). We expect the complexity of the experimental designs employed and the phenomenon under investigation will be important in determining the ubiquity of the group-to-person generalizability. 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 work may suffer less from the group-to-

person generalizability problem. Ultimately, we suggest that the problem is an issue for any area of psychological research that does not routinely investigate or model person-level data.

In terms of specific subdiscipline predictions, we hypothesize that the problem will be least frequent in low-level research, such as visual and auditory cognition/perception, due to assumptions that most people share basic physiological similarities, and some lack of conscious control, which underlie these disciplines' studied effects. As the content of study becomes more concerned with more complex higher-level cognition, the problem should be more prevalent, as responses will rely more on individual differences (e.g., values and knowledge). It may even be the rule rather than the exception in some areas, such as social cognition. 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). Therefore, social psychology, and any other disciplines which habitually rely on between-subjects designs to make inferences about psychology may be especially prone to committing the error of assuming that between-subjects patterns generalize to the person-level. However, these are empirical questions for which future research is needed. We hope this paper gives researchers both the motivation and tools to examine group-to-person generalizability in their own areas of interest.

Conclusion

Psychological scientists 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 the pervasiveness of some person-level 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 sampled persons. Put simply, psychology is a feature of persons, not averages or distributions. Therefore, person-level design and analytic approaches should be customary in psychological science.

Footnotes

1. 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 a predicted attenuation interaction effect. In an attenuation interaction prediction, 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).

2. 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.

3. In the researcher sample, 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. (Despite our methods experiments, we ironically did not think to allow participants to opt out). To conduct the most stringent test of our hypothesis, we recoded all of these 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.

4. 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.

5. In an ideal situation, we advocate for any of the possible intensive sampling approaches (when time and resources are virtually unconstrained). Having many trials per condition per person would yield observed person-level effects that are closer to the true person-level effects. Moreover, such intensive data would allow conduction of typical and high-powered statistical tests within each person. See our SOM for an example of a frequentist approach to "prevalence" using one of the datasets from Table 1.

References

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.

Crockett, M.J. (2016). How formal models can illuminate mechanisms of moral judgment and decision-making. *Current Directions in Psychological Science, 25*(2), 85-90.

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.

Devezer, B., Navarro, D.J., Vandekerckhove, J., & Buzbas, E.O. (2021). The case for formal methodology in scientific reform. *Royal Society Open Science, 8*, 200805.

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.

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

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.

Hamlin, K.J., Ullman, T., Tenenbaum, J., Goodman, N., & Baker, C. (2013). The mentalistic basis of core social cognition: Experiments in preverbal infants and a computational model. *Developmental Science, 16*(2), 209-226.

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.

Mayo, D.G. (2018). *Statistical inference as severe testing.*

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.

Meehl, P.E. (1990a). Appraising and amending theories: The strategy of Lakatosian defense and two principles that warrant it. *Psychological Inquiry, 1*(2), 108-141.

Meehl, P.E. (1990b). Why summaries of research on psychological theories are often uninterpretable. *Psychological Reports, 66*(1), 195-244.

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.

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.

Simons, D.J., Shoda, D.Y., & Lindsay, D.S. (2017). Constraints on generality (COG): A proposed addition to all empirical papers. *Perspectives on Psychological Science, 12*(6), 1123-1128.

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.