



Psychology is a Feature of Persons, Not Averages or Distributions: The Group-to-Person Generalizability Problem in Experimental Psychology

Journal:	<i>Advances in Methods and Practices in Psychological Science</i>
Manuscript ID	AMPPS-22-0036.R1
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.R1**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Q: Does your paper rely on new or previously unpublished empirical data from your lab?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

1
2
3 If you answered Yes to either of these questions, provide a URL (either public or view-only) where the
4 data can be accessed by the editors and reviewers.
5

6 All URLs are contained on our OSF page, which is available here: <https://osf.io/xyse4/>
7

8 If your paper relies on existing data that are available via third parties, please indicate who controls
9 access to those data and how other researchers can access them in the same way you have.
10

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

12 If needed, add any additional explanation about the data used in your paper.
13

14 CUST_TRANSPARENCY_TEXT_DATA_OTHER :No data available.
15
16
17

18 Available Materials

19
20 Q: Have you made available any and all materials necessary to reproduce your experiments, analyses, or
21 other paper contents?
22

23 A: CUST_TRANSPARENCY_MATERIALS :No data available.
24

25 If No, please explain which materials are unavailable and explain why they are not available.
26

27 CUST_TRANSPARENCY_MATERIALS_TEXT :No data available.
28

29 Q: Does your paper rely on any materials, code, or other resources that are new to this project (i.e., they
30 were developed or created as part of the research reported in this paper)?
31

32 A: Yes

33 If you answered Yes, provide a URL (either public or view-only) where reviewers and editors can view
34 those materials and resources. Enter "Not Applicable" if your paper does not rely on any such materials.
35

36 All materials are contained on our OSF page, which is available here: <https://osf.io/xyse4/>
37

38 Q: Does your paper rely on any materials, code, or other resources that are in the public domain or
39 previously made available by you or by other researchers?
40

41 A: No

42 If Yes, please indicate who controls access to those materials and how other researchers can access
43 them in the same way you have. Enter "Not Applicable" if you are not relying on data from other
44 researchers or third parties.
45

46 CUST_TRANSPARENCY_EXISTING_MATERIALS_TEXT :No data available.
47
48
49

50 Preregistration

51
52 Q: Were any of the studies or analyses reported in the manuscript preregistered?
53

54 A: Yes " all studies were preregistered
55
56
57
58
59
60

1
2
3 If only a subset of the reported studies were preregistered, indicate which ones were and which ones
4 were not. For any studies that were not preregistered, please indicate why not. If your manuscript does
5 not report new studies, please enter, "Not applicable."
6

7 CUST_TRANSPARENCY_PREREG_STUDIES :No data available.
8

9 Please provide a URL for the main project page where reviewers and editors can access the
10 preregistration documentation (leave blank if there are no studies or none was preregistered). This may
11 be an anonymous view-only link for the review process.
12

13 CUST_TRANSPARENCY_PREREG_URL :No data available.
14

15 Which aspects of your project were preregistered? (check all that apply. Note that if your preregistration
16 includes a complete analysis script that handles coding of measures, missing data, exclusions, analyses,
17 etc., you could check multiple boxes on this list based on preregistering that script.)
18

19 Theoretical hypotheses . Tasks/measures used for confirmatory hypothesis tests. Data exclusion criteria and
20 procedures. Data analysis plan. Planned interpretation for different patterns of results (could be part of the
21 "Theoretical hypotheses" if each hypothesis states how different patterns of results would support or
22 disconfirm it).. Target sample size
23

24 When did you complete the preregistration:

25 Prior to any data collection for the study (not including pilot testing of procedures)
26

27 Did you make any changes to the preregistered procedures when completing your study?
28

29 A: No
30

31 If you made changes to your preregistered plans at any stage of the process of completing your
32 research, list all of those changes below.
33

34 CUST_TRANSPARENCY_PREREG_CHANGES_TEXT :No data available.
35

36 Such changes must also be documented in the manuscript itself. Are all of the changes specified above
37 fully reported in the manuscript text?
38

39 CUST_TRANSPARENCY_PREREG_CHANGES_DOCUMENTED :No data available.
40

41 If you have other comments or explanations about your preregistration that not covered by the
42 questions above, enter them here (leave blank if no you have no comments/notes or if your paper has
43 no preregistered studies):
44

45 CUST_TRANSPARENCY_PREREG_OTHER :No data available.
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2
3 Q: Authors were asked to select items from the list below to indicate what was preregistered:
4

- 5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
- 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

21 The author checked the following boxes:
22

23 A: Theoretical hypotheses . Tasks/measures used for confirmatory hypothesis tests. Data exclusion criteria
24 and procedures. Data analysis plan. Planned interpretation for different patterns of results (could be part of
25 the “Theoretical hypotheses” if each hypothesis states how different patterns of results would support
26 or disconfirm it).. Target sample size
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2
3
4
5
6
7
8 **Psychology is a Feature of Persons, Not Averages or Distributions:**
9
10 **The Group-to-Person Generalizability Problem in Experimental Psychology**
11

12
13 Ryan M. McManus^{1*}

14
15 Liane Young¹

16
17 Joseph Sweetman²
18
19
20
21
22

23 ¹Department of Psychology and Neuroscience, Boston College, Boston, MA, USA

24
25 ²Department of Psychology, University of Exeter, Exeter, Devon, UK
26
27

28
29 * Corresponding author email:

30
31 mcmnurd@bc.edu
32
33

34
35 *Acknowledgements:* We would like to thank Stefano Anzellotti, Hiram Brownell, Richard
36
37 Morey, Ehri Ryu, and Jordan Theriault for helpful feedback at the beginning of this project. We
38
39 would like to thank Adam Bear, Tony Chen, Isaac Handley-Miner, Robin Ince, Minjae Kim,
40
41 Aditi Kodipady, Gordon Kraft-Todd, Matthew Leitao, Shangzan (Sunny) Liu, Michael (Mookie)
42
43 Manalili, Julia Marshall, Joshua Rottman, and Abraham Rutchick for helpful conversations at
44
45 various stages of this project, as well as providing feedback on an early draft of the manuscript.
46
47 We thank Nathan Liang and Sunny Liu (again) for investigating and diagnosing coding/output
48
49 issues in R during the revision process. Finally, we thank James W. Grice, an anonymous
50
51 reviewer, and Katie Corker, who provided invaluable feedback during the review process.
52
53
54
55
56
57
58
59
60

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.

1
2
3 Fourth, we conduct four pre-registered experiments to rule out several methodology-based
4 explanations of group-to-person generalizability failures. Fifth, we survey laypeople and
5 psychology researchers to understand what is inferred about person-level phenomena from
6 group-level analyses. Finally, we argue that experimental psychologists, if interested in person-
7 level psychology, ought to deploy different design and analytic strategies than those typically
8 used.
9

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

19 Psychology is often defined as “the study of the mind and behavior.” Therefore, its
20 essential goals are describing cognitive functions and uncovering their antecedents and
21 consequences. We contend that researchers intend to apply these goals to the study of persons, as
22 it is minds that possess psychological processes, and minds reside in persons. To strengthen this
23 argument, we ask readers to engage in a thought exercise. Recall your most recent meeting with
24 collaborators in which you discussed hypotheses and experimental designs to test them. At any
25 point in that meeting, did you reason about possible patterns in a way that reflected how *persons*
26 may respond to different stimuli, or did you exclusively reason in a way that reflected how
27 different stimuli would affect *averages or locations of distributions*? Furthermore, given the
28 seeming frequency with which studied phenomena are described as applying to people generally,
29 we also contend that most experimental psychologists intend to uncover phenomena that describe
30 a *majority* of persons (i.e., “general psychological laws”; Hamaker, 2012). Therefore, what
31 follows are the most important takeaways from this paper:
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49

- 50 1. Psychologists sometimes fail to design experiments that permit investigation of
51 person-level hypotheses.
52
53
54
55
56
57
58
59
60

1
2
3 2. Even when appropriate experimental designs are used, psychologists often report
4 *only* their group-level analyses and interpret them *as if* they support or falsify person-
5
6 level hypotheses.
7
8
9

10
11
12 Because it is possible for the above statements to be misinterpreted or overgeneralized,
13
14 we first communicate what we mean by “person-level” hypothesis or result, and we then clarify
15
16 our position on designing studies to test such hypotheses or yield such results.
17
18

19 *What is a “Person-Level” Result?*

20
21 Specifically, a “person-level” result is one which provides information about a pattern of
22
23 responding or activity for a single person (e.g., an experimental effect direction and magnitude
24
25 for person X). If there are many experimental trials or “intensive” sampling in a longitudinal
26
27 design, then one can compute single-subject statistics (e.g., effect sizes, confidence intervals,
28
29 factor structures, etc.) in order to make person-level inferences (e.g., Kurz, Johnson, Kellum, &
30
31 Willer, 2019). Alternatively, once all person-level patterns are investigated and counted, a
32
33 “pervasiveness” estimate can be obtained (Speelman & McGann, 2020). By “pervasiveness,” we
34
35 mean the choosing of one possible person-level pattern and investigating, “How many persons
36
37 match this pattern?” Finally, if the chosen person-level pattern is the one that was predicted or
38
39 claimed, then that person-level pattern’s pervasiveness can be tested against a value considered
40
41 to be theoretically meaningful (e.g., 50% if one is intending to test or make a *general*
42
43 psychological claim). This pervasiveness approach has the advantage of not requiring a high
44
45 level of statistical expertise or a large number of observations.
46
47
48
49
50

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

1
2
3 Between-subjects experiments do not permit tests of person-level hypotheses (Speelman
4 & McGann, 2020; Whittsett & Shoda, 2014). These common designs make it impossible to ask
5 the simple question, “How many people respond this way?” (see Speelman & McGann, 2020),
6 and they prohibit examination of unfolding person-level processes (e.g., Brandt & Morgan, 2022;
7 Fisher, Medaglia, & Jeronimus, 2018; Moeller, 2022). For example, consider the following
8 research question: “Do people prefer Coca-Cola to Mountain Dew?” To assess this, the leading
9 soda cognition lab designs an experiment which randomly assigns half of participants to rate
10 their likelihood of buying Coca-Cola in the next month, and the remaining participants to rate
11 Mountain Dew in the same way. They analyze the resulting data with an independent-samples t-
12 test, finding that the average likelihood judgment is higher for the Coca-Cola condition.
13
14 However, a rival soda cognition lab also attempts to answer this question, but they use a within-
15 subjects design and find an average difference in the opposite direction. Assuming the within-
16 subjects effect generalizes to the person-level, which of these designs better answers the
17 question, “Do people prefer Coca-Cola to Mountain Dew?” If *prefer* implies a psychological
18 comparison, we suggest that the within-subjects design reigns supreme. Moreover, there are
19 many plausible non-substantive mechanisms for the between-subjects results (e.g., for
20 participants who rated Coca-Cola, the counterfactual soda of Pepsi was referenced).
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41

42 To further illustrate this, Birnbaum (1999) showed that “People judge 9 as larger than
43 221” can be derived from a between-subjects design, as 9 invokes a context of 2-digit numbers
44 whereas 221 invokes a context of 3-digit numbers. We argue that no serious experimentalist
45 would interpret these results as people truly believing 9 is larger than 221 (and we again note that
46 “judge... as larger than...” implies a psychological comparison). Of course, Birnbaum’s methods
47 could have been better (e.g., measuring “largeness” on a true ratio scale), but we argue that
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2
3 criticizing his results on methodological grounds is only possible because we all believe that
4
5 there is a truth of the matter, with better and worse ways of verifying it. In most psychological
6
7 experiments, however, the truth of the matter is *inferred* from the data, meaning that it is
8
9 unknown how often between-subjects results are taken to reflect within-subject phenomena when
10
11 the between-subjects results are truly akin to Birnbaum's findings (see also Speelman &
12
13 McGann, 2020). If some non-trivial proportion of between-subjects experiments in psychology
14
15 are designed with the intention to reveal a psychological process or its outcome, this problem
16
17 may be pervasive.
18
19
20
21

22 To clarify, we are not suggesting that between-subjects are never useful. These designs
23
24 may be preferable when within-subjects designs are impossible or practically infeasible. For
25
26 example, many intervention(-like) research questions may be best answered with between-
27
28 subjects designs. Additionally, hypotheses about population(-like) differences require at least
29
30 some component of the study to be a between-subjects factor, such as testing whether
31
32 psychopaths show different experimental effects than non-psychopaths. Similarly, if the primary
33
34 object of study is not the psychological process/outcome itself (e.g., the variability of responses
35
36 as a function of some manipulated ability), a between-subjects design may be most feasible (see
37
38 our SOM experiments). Finally, between-subjects designs are unproblematic when the research
39
40 goal is to provide generalization evidence (e.g., finding similar effects across stimulus sets; see
41
42 Yarkoni, 2020).
43
44
45
46

47 We note, however, that between-subjects designs cannot conclusively provide person-
48
49 level evidence of an experimental effect, just as group-level correlations among variables cannot
50
51 provide evidence of person-level correlations among those variables (see Fisher et al., 2018). We
52
53 argue that within-subjects designs at least offer researcher the opportunity to engage with person-
54
55
56
57
58
59
60

1
2
3 level inferences. For example, in our own recent moral cognition research, we assessed moral
4 character judgments to test their sensitivity to social relationship information in the context of
5 helping behavior (McManus, Mason, & Young, 2021). Among other variations, participants in
6 our experiments were given two scenarios: one in which someone helps a total stranger, and
7 another in which someone helps a distant family member. Group-level analyses suggested that
8 participants—*on average*—judged agents who helped strangers as more morally good than agents
9 who helped family members, presumably because people believe that there is no obligation to
10 help strangers. Importantly, this was tested using a within-subjects design. Therefore, although it
11 was not reported, our design permitted investigation of the question, “How many people respond
12 this way?” A between-subjects design would have disallowed such investigation.
13
14
15
16
17
18
19
20
21
22
23
24
25

26 However, having within-subjects designs does not automatically prevent inference errors
27 from occurring. Researchers can still commit ecological or ergodic fallacies (Kuppens & Pollet,
28 2014; Speelman & McGann, 2020), due to special instances of Simpson’s paradox—when
29 group-level patterns poorly represent lower-level units constituting the group (Simpson, 1951;
30 Kievit, Frankenhuys, Waldorp, & Borsboom, 2013; also see Hamaker, 2012, for an illustrative
31 example on the relation between typing speed and mistake frequency). To reiterate, even when
32 psychologists deploy appropriate experimental designs, they often, if not always, only report
33 their group-level analyses.
34
35
36
37
38
39
40
41
42
43

44 Overall, we are suggesting that, if a theory or research question is a person-level one, and
45 the goal of a study is to make a general claim (Hamaker, 2012), then researchers ought to choose
46 appropriate designs and analytic procedures. However, such careful matching does not always
47 occur in practice. The rest of this paper focuses on instances in which within-subjects group-
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2
3 level effects fail to describe the majority of sampled persons. From here on, we refer to this as
4
5 the “group-to-person generalizability problem.”
6

7 8 **Group-to-Person Generalizability Problems in the Wild** 9

10 We examined open data from the past five years in psychological research (2016–2021),
11
12 looking for the group-to-person generalizability problem. Due to the larger reform movements in
13
14 psychology, publications from this era should be relatively more rigorous than prior eras (e.g.,
15
16 larger samples, better statistical inferences). Our investigation was not systematic in the sense
17
18 that we can say, “X% of publications contain the problem.” Rather, using a person-level analytic
19
20 approach, we re-analyzed open data with the goal of finding five instances of the problem from
21
22 moral cognition—as we ourselves are moral psychologists—and five instances from social
23
24 cognition generally (e.g., race, gender, humor, etc.; see Table 1). Anecdotally, many seemingly
25
26 person-level claims (not reported here) were made from between-subjects designs that
27
28 fundamentally disallowed reanalysis. We also note that even though we investigated examples
29
30 from social cognition in particular, this problem is not limited to social cognition, as others have
31
32 identified pitfalls of averaging across persons in somewhat lower-level cognitive research on
33
34 visual perception of faces (Grice et al., 2020) and context effects in decision-making (Liew et al.,
35
36 2016).
37
38
39
40
41

42 To accomplish person-level analysis, we adopted “pervasiveness” or “persons-as-effect-
43
44 sizes” approaches (see Grice et al., 2020; Spelman & McGann, 2020). Put simply, we created
45
46 variables in each dataset that distinguished participants based on whether their response patterns
47
48 supported the reported group-level patterns. If a participant’s responses had at least *some*
49
50 distance between experimental conditions (e.g., 1-point on a Likert/sliding scale) and were
51
52 directionally consistent with a group-level pattern, then that participant was categorized as
53
54
55
56
57
58
59
60

1
2
3 supporting group-to-person generalizability. Because some measures were averages of multiple
4 items, this meant that, in practice, we counted participants as supporting generalizability even if
5 their experimental effects were near- but not exactly zero. Therefore, we used an extremely
6 liberal threshold.
7
8
9
10

11
12 An important nuance is that all of the investigated claims are based on *sets* of group-level
13 tests (e.g., multiple paired t-tests). We therefore extended the person-level approaches to
14 accommodate such claims. Specifically, we categorized participants as supporting
15 generalizability if their full set of responses matched the full set of group-level patterns. For
16 example, if a 2x2 interaction pattern underlied the claim, we counted person-level responses as
17 supporting generalizability if a participant's simple effects' directions and differential
18 magnitudes reflected the group-level pattern. But the ordering of all four condition averages was
19 not accounted for. Readers can imagine (and if they wish, investigate) what these analyses look
20 like under stricter constraints (see our OSF page: <https://osf.io/xyse4/>).
21
22
23
24
25
26
27
28
29
30
31
32

33 We statistically defined claims as being unsupported at the person-level if we could not
34 rule out the possibility that fewer than a simple majority of people would show the group-level
35 pattern, conducting a binomial test against 0.50. If the 95% CIs contained a proportion of 0.50 or
36 lower, then the claim was unsupported at the person-level. We chose this 0.50 proportion as the
37 null value because most claims in psychology articles do not use language that suggests an
38 experimental effect is one that describes only a (minority) subset of participants. This means that,
39 at least by implication, the effect is being communicated as applying to *most* participants
40 (something we test empirically in *An Important Objection*). Therefore, testing against this null
41 value, too, is a liberal threshold for categorizing claims as favoring generalizability (i.e., a simple
42 majority). As Table 1 shows, proportions of participants favoring generalizability varied across
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2
3 these publications but was low overall (3%-50%, with most proportions ranging between 20%-
4
5 40%). Importantly, this occurred across a variety of dependent variables (e.g., sliding scales,
6
7 Likert scales, reaction times, error rates) and pattern types (crossover interactions, attenuation
8
9 interactions, ordinal patterns, conjunctive differences). We note that authors of these studies may
10
11 not believe that their claims are describing most of their participants (although see *An Important*
12
13 *Objection*). However, descriptions of results are, at the least, ambiguous enough to warrant
14
15 uncertainty. We next break down a moral cognition example showing how the group-to-person
16
17 generalizability problem can occur.
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

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.”	<u>Experiments 1a-b</u> -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.”	<u>Experiments 1 & 4</u> -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.”	<u>Experiment 2</u> -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 <i>relatively less</i> partiality in their judgments of what is morally right, compared to judgments of how they would act.”	<u>Experiment 2</u> -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.”	<u>Experiments 1-3</u> -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.”	<u>Experiment 4</u> -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.”	<u>Experiments 2 & 4</u> -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.”	<u>Experiment 1</u> -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.”	<u>Experiment 4</u> -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.”	<u>Experiment 3</u> -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.

1
2
3 To investigate the above claim at the person-level, each simple effect and the interaction
4 can be described by a set of directional patterns. The No Choice simple effect can be computed
5 by subtracting the “helped a family member” ratings from the “helped a stranger” ratings,
6
7 whereas the Choice simple effect can be computed by subtracting the “helped a family member
8 instead of a stranger” ratings from the “helped a stranger instead of a family member ratings.”
9
10 An interaction effect can then be computed by subtracting the Choice effect from the No Choice
11 effect (see Table 2 for an example of 13 hypothetical participants who reflect all possible
12 qualitative patterns, and Table 3 for example R code to create generalizable 2x2 person-level
13 patterns, investigate their frequencies, and conduct a binomial test). The person-level
14 combination which matches the published claim is pattern number 6 (i.e., the “Positive,
15 Negative, Positive” pattern: No Choice simple effect, Choice simple effect, Interaction effect).
16
17 Conversely, a person-level combination which does not match the published claim but can still
18 be categorized as showing a “Positive” interaction value is pattern number 10 (i.e., the “Positive,
19 Zero, Positive” pattern).
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

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

<p>Step 1</p>	<p>Use wide-formatted data (i.e. 1 row per participant) to create simple effects of interest.</p>	<pre>data_wide <- data_wide %>% mutate(SimpleEff1 = A1 - A2) %>% mutate(SimpleEff2 = B1 - B2)</pre>
<p>Step 2</p>	<p>Create variables which constitute person-level pattern possibilities.</p>	<pre>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"))</pre>
<p>Step 3</p>	<p>Investigate frequencies of all person-level patterns.</p>	<pre>data_wide %>% group_by(`2x2_Pattern`) %>% summarize(freq = n())</pre>
<p>Step 4</p>	<p>Test the predicted effect’s frequency against a meaningful proportion value, using a binomial test.</p>	<pre>binom.test(x = Predicted Effect Freq, n = Total N, p = Prop Value, alternative = "two.sided")</pre>

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.

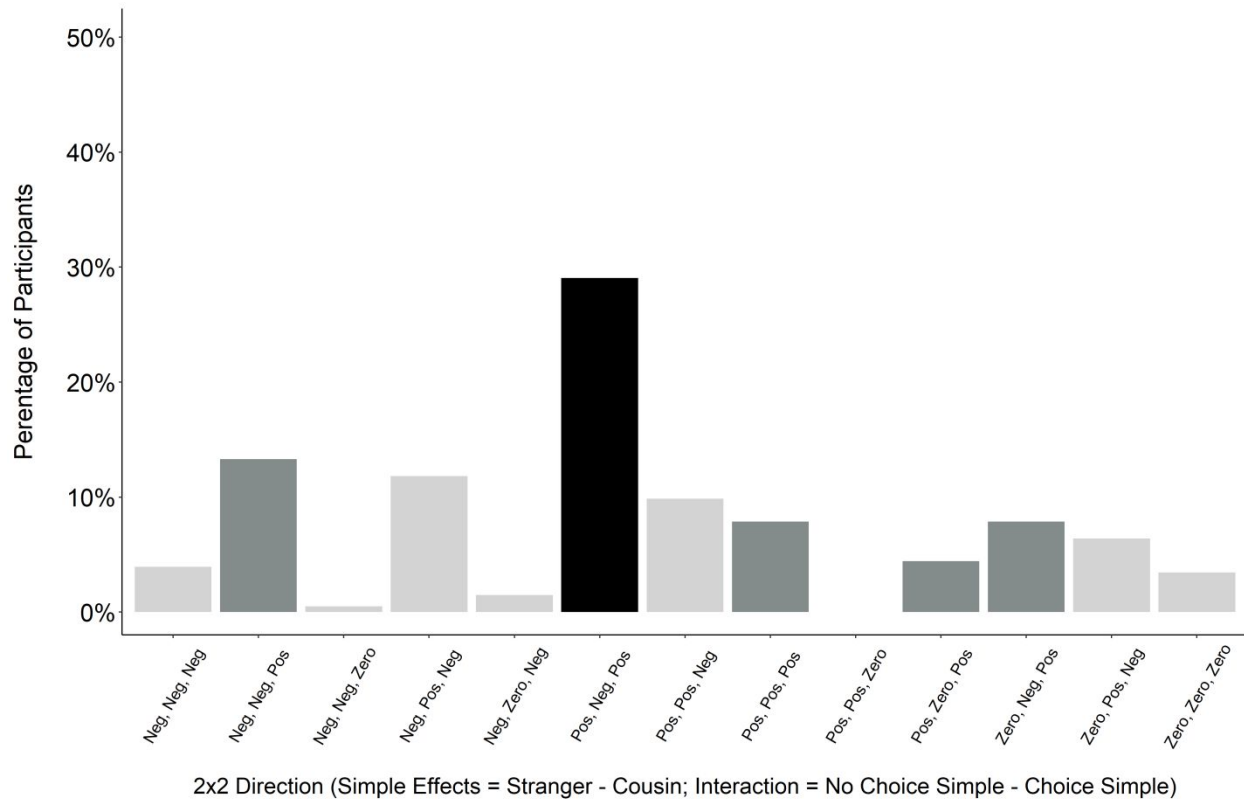


Figure 1. Person-Level Patterns from McManus, Mason, & Young (2021). Pattern descriptions (e.g., Pos, Neg, Pos) communicate the No Choice difference, Choice difference, and Interaction difference, respectively. The black bar represents the claimed group-level patterns. Dark grey bars represent patterns 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

1
2
3 method involves an extra step, creating a third difference score variable—the interaction score—
4
5 by subtracting the second simple effect’s difference score from the first simple effect’s
6
7 difference score. The one-sample t-test method is completed by conducting a t-test (against zero)
8
9 on the interaction scores. If either t-test returns a below-alpha p-value, then an interaction effect
10
11 exists. Importantly, in this context, the p-value from both t-test methods would be identical to
12
13 one another and to the p-value of the ANOVA’s interaction F-test, as all methods are testing for
14
15 a difference in differences (see SOM for a demonstration).
16
17
18

19 Why does this matter? As shown in Table 2, there are five patterns which yield a positive
20
21 interaction value, only one of which is the claimed pattern¹. This is problematic considering that
22
23 the interaction test is simply assessing whether the interaction scores’ average differs from zero,
24
25 nothing more. Therefore, it is possible that more participants had a positive interaction value
26
27 constituted by the “incorrect” set of simple effects than had a positive interaction value
28
29 constituted by the “correct” set of simple effects. Indeed, more than 60% of our sample had a
30
31 positive interaction value that contributed to the group-level interaction test (see Figure 1).
32
33
34

35 Now consider the opposite-signed simple effects. It is an obvious but crucial point that a
36
37 person-level claim about the full interaction pattern requires that participants show *both* simple
38
39 effects. However, what seems non-obvious is that *sets* of typical inferential tests cannot provide
40
41 this evidence. Because the units of analysis for a single paired-samples t-test are the person-level
42
43 difference scores, two separate paired-samples t-tests cannot connect units across analyses. The
44
45 only way to ensure that a particular proportion of participants show both group-level patterns is
46
47 to first count how many show each individual pattern. Tabulations of within-person differences
48
49 showed that the first simple effect described 51% of participants, whereas the second simple
50
51 effect described 55% of participants. Consequently, the *maximum* proportion of participants who
52
53
54
55
56
57
58
59
60

1
2
3 could have shown both patterns was 51%. As has been established, however, fewer than 30% of
4
5 participants showed both patterns.
6

7
8 Given this re-analysis and explanation, we suggest that the goal of a psychological
9
10 experiment should not be to explain a large proportion of variance (e.g., as is often reported in an
11
12 ANOVA/regression context), but to instead explain a large proportion of persons, as
13
14 psychological experiences occur within (not across) persons. Once this is realized, psychologists
15
16 can instead focus on developing and testing causal models which attempt to explain the
17
18 underlying data generation process (e.g., Grice, 2015; Grice et al., 2017).
19
20

21 **The Problem Worsens (and is Difficult to Fix)!**

22
23
24 We believe that we have provided compelling reasoning that person-level claims need to
25
26 be tested using persons-as-effect-sizes or pervasiveness approaches— tabulating the proportion
27
28 of participants whose responses match predictions (Grice et al., 2020; Spelman & McGann,
29
30 2020). To provide further supporting evidence, we generated hypothetical datasets in which sets
31
32 of group-level analyses are extremely poor representations of person-level cognition. In these
33
34 datasets, we created 2x2 crossover interactions, 2x2 attenuation interactions, and three-level
35
36 ordinal effects, all of which yield group-level effects (and survive non-parametric tests) but
37
38 describe *zero* participants (see SOM). Although we are unaware of real-world instances, the
39
40 theoretical possibility of group-level patterns being perfectly unrepresentative of persons should
41
42 warrant caution².
43
44
45

46
47 Despite these existence proofs, it could be argued that most discrepancies between group-
48
49 level and person-level analyses are due to methodological features of experiments which can be
50
51 remedied. That is, most experiments may not be designed to minimize noise and therefore
52
53 maximize the probability of persons exhibiting the group-level pattern. If such barriers could be
54
55
56
57
58
59
60

1
2
3 addressed, then group-level patterns may better represent person-level patterns. To address this,
4 using our moral cognition paradigm described earlier, we conducted four pre-registered
5
6 experiments which systematically varied methodological features hypothesized as partial, noise-
7
8 inducing causes of the group-to-person generalizability problem.
9

10
11
12 As an example, consider the problem of sequential stimulus presentation in typical
13 judgment paradigms. When participants are presented with many stimuli, they are typically
14 presented with one stimulus at a time, after which a judgment is measured. This sequential
15 procedure continues until participants see and respond to all stimuli. This procedure can induce
16 noise in the following way. Some participants might not have judged an early stimulus with the
17 extreme response option if they knew that they would perceive a later stimulus as more extreme;
18 consequently, false ties between stimuli might emerge when participants truly wish to judge
19 them differently. Additionally, this same procedure can lead to some participants forgetting how
20 they made judgments of earlier stimuli, leading to, e.g., differences between stimuli that they
21 truly wish to judge similarly. Therefore, if this kind of noise is operant in typical judgment
22 paradigms (and it is systematically reducing the number of participants who respond in a manner
23 consistent with the predicted group-level effects), participants who have the ability to see all
24 stimuli before making their judgments may be more likely to match the predicted effect. Within
25 four experiments of this spirit, we replicated our original group-level effects, as well as the low
26 proportions of participants represented by them (17%-27%; see also Devezzer et al. [2021] for a
27 discussion on how replicability need not imply “true”). However, none of our experiments were
28 successful in explaining the problem and therefore shifting person-level patterns to be better
29 aligned with the group-level pattern (see Table 4 for a summary of the experiments’ logic and
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2
3 results, and SOM for full details). All four experiments were pre-registered at the following
4
5 links: <https://osf.io/wfz3b>, <https://osf.io/7utrg>, <https://osf.io/8x69c>, and <https://osf.io/fcbxe>.
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

For Review Only

Table 4. Underlying Logic and Results for Methodology-Based Experiments (see SOM for full details)

Manipulation	Underlying Logic	Results
<p>Absence/Presence of Calibration Trials</p>	<p>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.</p> <p>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.</p> <p>Solution: Before the main experimental task, give participants calibration trials and normative feedback about how most other people use the scale.</p> <p>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).</p>	<p>N per Condition <i>N</i>Control: 658 <i>N</i>Experimental: 589</p> <p>Predicted Interaction Control: 24% Experimental: 27%</p> <p>Eq of Proportions Test $\chi^2 = 1.17, p = .280$</p> <p>Hypothesis Decision Unsupported</p>
<p>Inability/Ability to Respond to Stimuli Simultaneously</p>	<p>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.</p> <p>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.</p> <p>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.</p> <p>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</p>	<p>N per Condition <i>N</i>Control: 628 <i>N</i>Experimental: 609</p> <p>Predicted Interaction Control: 24% Experimental: 19%</p> <p>Eq of Proportions Test $\chi^2 = 4.65, p = .031$ (Wrong direction)</p>

	<p>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).</p>	<p>Hypothesis Decision Unsupported</p>
<p>Absence/Presence of Matched Stimuli</p>	<p>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.</p> <p>Solution: Give participants matched-in-content stimuli across experimental conditions, varying only the experimental features of interest.</p> <p>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).</p>	<p>N per Condition NControl: 638 NExperimental: 641</p> <p>Predicted Interaction Control: 24% Experimental: 17%</p> <p>Eq of Proportions Test $\chi^2 = 10.94, p < .001$</p> <p>Hypothesis Decision Unsupported (Wrong Direction)</p>
<p>Inability/Ability to “Opt Out” of using Measures/Scales</p>	<p>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.</p> <p>Solution: Give participants the ability to opt out of using a measurement scale.</p> <p>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).</p>	<p>N per Condition NControl: 746 NExperimental: 691</p> <p>Predicted Interaction Control: 22% Experimental: 23%</p> <p>Eq of Proportions Test $\chi^2 = 0.09, p = .779$</p> <p>Hypothesis Decision Unsupported</p>

(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

1
2
3 **Participants.** All laypeople were U.S. residents recruited and compensated via CloudResearch’s
4 “approved participants” list. Participants from McManus et al. (2021) were unable to access the
5 current study. Additionally, participants from our methods experiments could not participate.
6
7 Researchers were affiliated with the Society for Personality and Social Psychology (SPSP),
8 recruited via SPSP’s Open Forum listserv and compensated with Amazon gift cards. Participants
9 who did not complete the entire study were not included in our final analyses. As pre-registered
10 (<https://osf.io/6qay8> and <https://osf.io/nucbf>), we aimed to collect at least 642 analyzable
11 laypeople and 280 analyzable researchers. In total, we were able to collect 705 and 256 unique
12 responses, respectively. After applying the pre-registered exclusion criterion (failing a
13 comprehension check), this resulted in $N_{Laypeople}=588$ (gender: 309 female, 273 male, 6 non-
14 binary; ethnicity: 457 White, 68 Black, 5 American Indian, 41 Asian; 1 Pacific Islander; 16
15 other; $M_{Age} = 38.69$, $SD_{Age} = 11.29$) and $N_{Researchers}=244$ (165 female, 68 male, 8 non-binary, 3
16 other; ethnicity: 158 White, 3 Black, 1 American Indian, 55 Asian; 17 other, 9 Biracial; 1
17 Multiracial; $M_{Age} = 33.09$, $SD_{Age} = 11.34$). Although we did not pre-register a stopping rule, we
18 decided not to resample due to still having high statistical power for our focal hypothesis tests
19 (see *Statistical Power & Hypotheses*).

20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40 **Design.** Participants were randomly assigned to one of two conditions. Half of participants
41 learned about a simple effect comparison, whereas the other half of participants learned about a
42 more complex, two-way interaction effect. We note that we used both simple and complex effect
43 examples to test the generality of our hypotheses. That is, had we only conducted the study using
44 one effect type, we could have capitalized on our hypothesis only being true of a specific effect
45 type. This is why our pre-registration refers to our design as “observational,” even though we
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2
3 randomly assigned participants to one effect type; we never intended to (nor did we) explicitly
4 compare the simple effect data to the complex effect data.
5
6

7 **Materials and Procedure.** At the beginning of the study, all participants were informed that they
8 would be answering questions about a moral cognition experiment. For the simple effect
9 condition, participants learned about an effect from the supplemental materials of Law,
10 Campbell, & Gaesser (2021). For the complex effect condition, participants learned about the
11 interaction effect from McManus et al. (2021).
12
13
14
15
16
17
18

19 Participants first read text communicating results in typical journal article format (with
20 means, SDs, t-values, p-values, within-subject standardized effect sizes for comparisons of
21 interest [d_z], and a barplot; see OSF for full materials). After learning the results, they then read
22 text that simulated how data-based claims are made in a general discussion section (e.g., “People
23 judged fictional agents who helped a stranger as more morally good than fictional agents who
24 helped a cousin, but they judged fictional agents who helped a stranger instead of a cousin as less
25 morally good than fictional agents who helped a cousin instead of a stranger”).
26
27
28
29
30
31
32
33
34

35 After learning about the claim, participants were then asked to respond to a series of true-
36 false questions about what the reported results suggested. However, these questions were not of
37 primary interest (see OSF for Rmarkdown results). Participants were then again shown the claim
38 in general discussion format, and asked “By *people*, approximately what percentage of the
39 study’s participants do you think the researchers mean?” We call this measure the “empirical
40 proportion estimate.” Responses ranged from 0-100% on a sliding scale, with the starting
41 position (0, 50, 100) counterbalanced across participants. This measure allows categorization of
42 responses into two categories: less than a simple majority (50% or less), and equal to or greater
43 than a simple majority (51% or more). To move on to the next page, participants had to at least
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2
3 click on the slider, meaning that the slider's starting value would have been recorded as the
4 participant's response. As can be seen in Figure 2, however, these starting values were very
5 infrequent, suggesting that participants indeed engaged with the task.
6
7
8
9

10 Next, participants learned about a (fictional) general, person-level theory that the authors
11 had developed pre-study. Participants were then asked to respond to a series of true-false
12 questions about how the reported results informed the theory (see OSF). Participants were again
13 shown the claim in general discussion format and told that, later in the paper, the authors used
14 their study's results to claim support for their theory. Participants were then asked, "In order for
15 the study's results to support the researchers' theory/model, approximately what percentage of
16 the study's participants do you think need to respond in the way described by [the general
17 discussion's language]?" We call this measure the "theoretical proportion estimate." Responses
18 were measured identically to the empirical estimate. Finally, participants could write an open-
19 ended response to communicate anything that they were unable to communicate thus far. After
20 the main task, participants answered several demographic questions.
21
22
23
24
25
26
27
28
29
30
31
32
33
34

35 **Statistical Power.** As pre-registered, we aimed for at least 321 participants per condition for the
36 laypeople sample, and 140 participants per condition for the researcher sample. The pre-
37 registered laypeople sample size yielded 95% power to detect a 10-point proportion difference
38 from 50% (e.g., 60%) using a two-tailed binomial test and assuming an alpha level = 0.05, the
39 focal test to examine whether a majority of empirical/theoretical proportion estimates reflect
40 inferences being made about a majority of a study's participants. As explained in our pre-
41 registrations, we planned the researcher sample based on the results of the laypeople sample. For
42 the researcher sample, the pre-registered sample size yielded 95% power to detect a 15-point
43 proportion difference from 50% using identical test specifications as the laypeople sample.
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2
3 In the laypeople sample, applying the pre-registered exclusion criterion (i.e., missing a
4 comprehension check question) led to $N_{Simple}=303$ and $N_{Complex}=285$. In the researcher sample, we
5
6 were unable to successfully recruit our entire desired sample size. After one attempt to get more
7
8 responses (via reposting to SPSP's Open Forum listserv), we decided to close the survey once
9
10 incoming responses completely stalled, which occurred after two weeks. Applying the same
11
12 exclusion criterion led to $N_{Simple}=123$ and $N_{Complex}=121$. We did not resample for either
13
14
15 population because sensitivity analyses revealed that we still had more than 90% power to detect
16
17 our pre-registered minimal effect sizes.
18
19

20 21 22 **Hypotheses**

23
24
25 1) Empirical Proportion: The majority of people (i.e., 51% or more) within each sample
26
27 (laypeople and researchers) will believe authors' claims are intended to describe at least a
28
29 simple majority (i.e., 51% or more) of their study's participants.
30
31

32
33 2) Theoretical Proportion: The majority of people within each sample will believe at least a
34
35 simple majority of a study's participants ought to be described by the authors' claims in order
36
37 for the results to support a general theory of person-level psychology.
38
39

40 41 **Results**

42 ***Empirical Proportion Estimate.*** The majority of laypeople believed authors intended to describe
43
44 at least a simple majority of their study's participants, for both simple (81%) and complex (88%)
45
46 effects. The majority of researchers agreed for both simple (73%) and complex (80%) effects³
47
48 (see Table 5 for additional descriptive statistics and Table 6 for inferential statistics). Strikingly,
49
50 as shown in Figure 2, there is no discernible pattern as a function of being relatively
51
52 inexperienced (e.g., layperson or undergraduate) and relatively experienced with academic
53
54 research (e.g., professor).
55
56
57

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 experience⁴.

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

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

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

Effect Type	Population	Proportion	<i>p</i> -value
Simple	<i>Laypeople</i>	81% [76% - 85%]	< .001
	<i>Researchers</i>	73% [64% - 81%]	< .001
Complex	<i>Laypeople</i>	88% [84% - 92%]	< .001
	<i>Researchers</i>	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	<i>p</i> -value
Simple	<i>Laypeople</i>	93% [90% - 96%]	< .001
	<i>Researchers</i>	80% [72% - 87%]	< .001
Complex	<i>Laypeople</i>	92% [89% - 95%]	< .001
	<i>Researchers</i>	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.

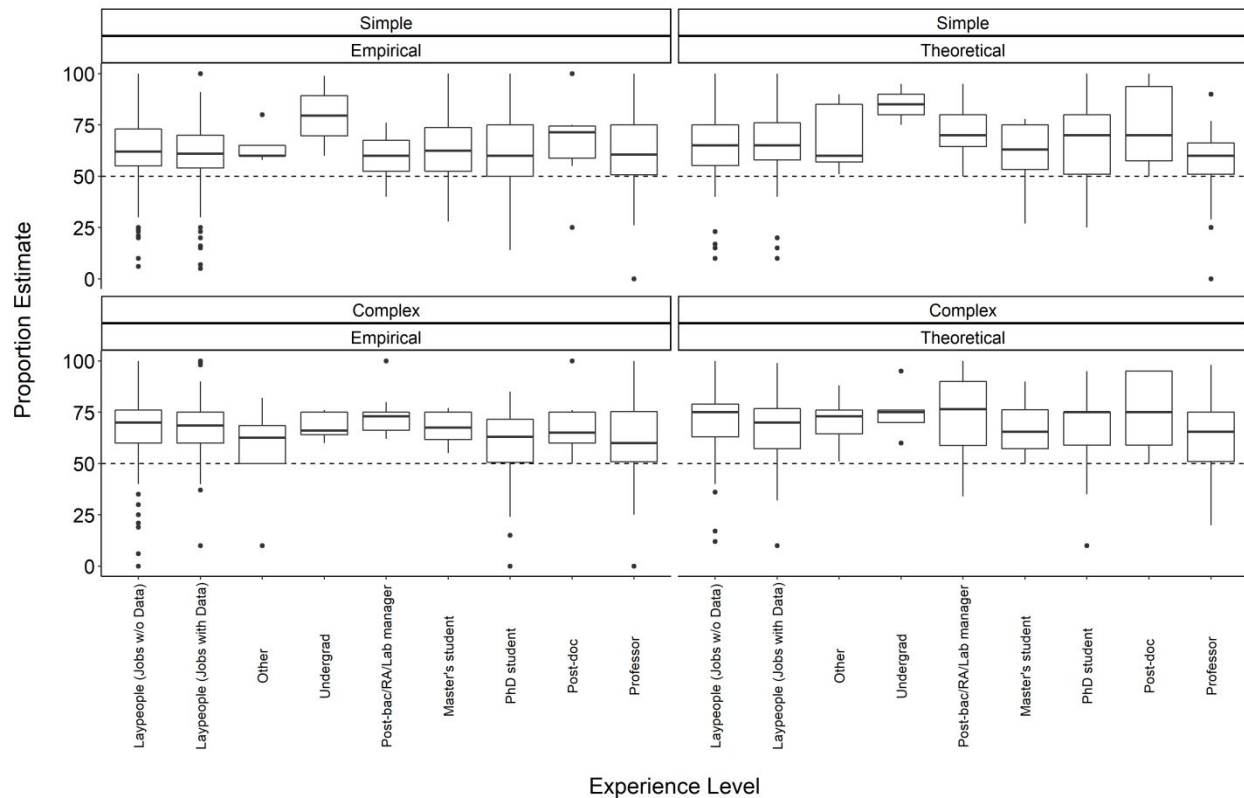


Figure 2. Boxplots of empirical/theoretical proportion estimates by effect type (simple versus complex), and by participants' level of experience. Note that “Other” refers to people involved in academic research in some way (via SPSP) but who indicated that they have never held an academic position.

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

1
2
3 noise explanations for this group-to-person generalizability problem in a moral judgment
4 paradigm, with obvious remedies proving unsuccessful. Finally, our research shows that a
5 majority of laypeople and social psychology researchers interpret authors of psychology articles
6 as intending to make claims that represent a majority of their study's participants. Moreover, a
7 majority of laypeople and researchers believe that this ought to be the case if authors are using a
8 study's results to claim support for a general, person-level psychological theory.
9
10
11
12
13
14
15
16

17 Our research is consistent with recent critiques put forth, in which some researchers (e.g.,
18 Richters, 2021; Speelman & McGann, 2020) have argued that there is a pervasive mismatch
19 between psychological theorizing and the analytic procedures used for testing—typical
20 theorizing occurs at the person-level but analytic procedures operate at the group-level. Over the
21 past decade, much effort has gone toward correcting, and promoting better, statistical inferences
22 (e.g., Lakens, 2021), but relatively fewer reform efforts have been aimed at appropriate
23 psychological inference (e.g., Moeller et al., *preprint*; Navarro, 2019; Liew, Howe, & Little,
24 2016) and proper theory development (e.g., van Rooij & Baggio, 2021). The current research
25 suggests that even if theorizing indeed improves, inference can still go wrong if familiar
26 statistical methods are privileged over ones that address specific psychological questions. Put
27 simply, psychologists seem to have put the statistical cart ahead of the psychological horse. This
28 problem, however, should not be judged as just another instance of “psychology in crisis.”
29 Instead, this is an opportunity to put past, current, and future research through more stringent
30 tests—to better ground our psychological claims in *persons*.
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48

49 *Responses to Potential Objections*

50
51 One objection is that, if the modal person-level pattern matches the group-level pattern,
52 then there is no group-to-person generalizability problem. We suggest, however, that if the
53
54
55
56
57
58
59
60

1
2
3 modal person-level pattern matches the group-level pattern but is not *also* the pattern describing
4 most participants, then the majority of responses would be unexplained. Moreover, most polled
5 laypeople and social psychology researchers agreed that for a claim to support a theory, it ought
6 to represent a majority of participants. Extrapolating from this, it stands to reason that *any*
7 reported effect that does not describe a majority of participants (regardless of its use for theory)
8 ought to be tagged as unrepresentative.
9

10
11
12 A related objection is that the proportion of participants showing the group-level pattern
13 may be above chance given how many possible qualitative patterns exist, meaning that there is
14 no group-to-person generalizability problem. For example, in a simple two-cell design, there are
15 only three possible person-level effects (i.e., positive, negative, or zero). Based on this objection,
16 if any of those effects described more than 33% of the sample and was consistent with the group-
17 level pattern, then the study would not contain the group-to-person generalizability problem. Our
18 reasoning against the first objection also applies here: if the modal pattern does not also describe
19 most participants, then the group-level pattern represents only a minority of responses. In
20 addition, there is unfortunately no ground truth to a “better than chance” criterion. That is,
21 designs could be modified to include additional conditions that are likely to engender uniform
22 responses across participants, creating more possible qualitative patterns while artificially
23 reducing the chance of any single pattern. We note that if psychologists are truly *only* interested
24 in testing a person-level pattern’s pervasiveness against “randomness” or “chance,” there are
25 well-established methods for doing so (i.e., randomization tests, Grice, 2021; but see our SOM
26 for how to conduct proportion tests against chance to make sample-to-population inferences for
27 Table 1’s claims).
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2
3 A final objection is that there are other sources of noise accounting for the group-to-
4 person generalizability problem, beyond those tested here (see SOM). For example, some
5 participants are distracted, leading to frequencies of person-level patterns which do not represent
6 the “true” frequencies. First, consistent with our experimental results, there is no reason to
7 believe, if such noise was reduced, that most person-level patterns would conveniently shift to
8 the group-level pattern. Second, as our tutorial and hypothetical datasets show, there are simple
9 non-method explanations for how group-level patterns can be (even perfectly) unrepresentative
10 of persons. Therefore, rather than assuming that there are solvable methodological issues
11 underlying the problem, it should be accepted that person-level patterns cannot be inferred from
12 group-level analyses.
13
14
15
16
17
18
19
20
21
22
23
24

25 26 *Recommendations*

27
28 Given the group-to-person generalizability problem, what should experimental
29 psychologists do? Our recommendations are consistent with those in a recent critique (Yarkoni,
30 2020). We note that our generalizability critique refers to generalizing across levels of analysis,
31 whereas the recent critique refers to generalizing across stimuli, tasks, etc. Specifically, we
32 propose predicting specific orderings of observations based on theory (e.g., A1 higher than A2 in
33 B1, but A2 higher than A1 in B2), while specifying a proportion of participants whose responses
34 should match predictions for the theory-derived hypothesis to survive. To conduct a “severe test”
35 (Mayo, 2018) or corroborate a “risky prediction,” (Meehl, 1990a, 1990b), the empirical
36 proportion should be close to the theory-predicted proportion, and other theories should not
37 predict this proportion.
38
39
40
41
42
43
44
45
46
47
48
49
50

51 Unfortunately, there may not be many psychological theories (especially in [social]
52 cognition which typically lacks formal models) that can make such predictions (for examples and
53
54
55
56
57
58
59
60

1
2
3 discussion of this issue see, Crockett, 2016; Hamlin et al., 2013). However, theoretical progress
4 can still be made. To test face validity, researchers can make minimum proportion predictions.
5
6 To do this, researchers can tabulate the proportion of participants whose responses match a
7
8 predicted pattern. They can then make a frequentist inference by testing this proportion against a
9
10 majority null (i.e., 50%) using a binomial test. We *strongly* recommend supplementing the
11
12 binomial testing approach with an estimation approach (e.g., Cumming, 2014) in which the
13
14 estimated intervals are interpreted to rule out certain proportions. To be clear, we do not
15
16 advocate 50% as the benchmark against which psychologists should test theory; we are simply
17
18 suggesting a method that enables researchers to identify evidence that fails to support general
19
20 psychological regularities (Hamaker, 2012). Others have suggested even higher proportions as
21
22 convincing evidence (e.g., 80%; Speelman & McGann, 2020), though responses from the current
23
24 work suggest that psychologists disagree about the appropriate cutoff. We also do not
25
26 recommend ignoring theory-inconsistent patterns, or patterns represented by a small minority of
27
28 participants. Understanding if and why other patterns exist allows refinement of theory by
29
30 postulating and testing whether there are substantive moderating variables (e.g., individual
31
32 differences), or simple violations of auxiliary assumptions (e.g., divergent interpretations of
33
34 measures; see Quintana, 2021, for a discussion).

35 36 37 38 39 40 41 42 *Limitations and Future Directions*

43
44 One constraint of this person-level approach is that it ignores magnitude information
45
46 (e.g., participants who use two extreme ends of a measure are treated identically to participants
47
48 who use two close points of a measure). However, magnitude information can be incorporated
49
50 into this approach. Researchers can choose an “imprecision value” (Grice et al., 2020), allowing
51
52 only certain magnitudes to support a qualitative pattern. Additionally, researchers can plot
53
54
55
56
57
58
59
60

1
2
3 frequencies of qualitative patterns by different imprecision values, allowing discernment
4
5 between participants who show small versus large effects (see Speelman & McGann, 2020,
6
7 Figure 4).
8
9

10 Relatedly, there are other (potentially better) methods for evaluating person-level effects
11
12 in high-repetition studies that yield magnitude information, such as person-level standardized
13
14 effect sizes and confidence intervals (e.g., Kurz, Johnson, Kellum, & Willer, 2019; Schuurman,
15
16 Houtveen, & Hamaker, 2015). Additionally, recent research suggests first testing effects within
17
18 each person (using typical statistical approaches at the person-level) and then using those tests to
19
20 calculate a Bayesian “prevalence” estimate (Ince, Kay, & Schyns, 2022; Ince et al., 2021). These
21
22 methods are not unlike our suggested approach of using a binomial test and its estimate, as they
23
24 both allow for inferences from samples to populations, a feature unsupported in some other
25
26 approaches (e.g., randomization tests). However, the intensive sampling that all these methods
27
28 require are often infeasible in experimental psychology research, as person-level statistical tests
29
30 would be subject to the same issues that have pervaded the replicability movement (e.g., number
31
32 of observations and therefore statistical power). Therefore, unless all experimental psychologists
33
34 start designing high-repetition studies, our proposal is the only method available, to our
35
36 knowledge, that will consistently allow person-level pervasiveness claims to be generalized from
37
38 samples to populations. Importantly then, our method is also the most generalizable, as it can
39
40 also be used for high-repetition studies, single-response-per-condition studies, and anything in
41
42 between⁵. In repetition designs, an analyst can simply average over a person’s multiple responses
43
44 within each condition and then apply the person-level analysis to those averages. This suggestion
45
46 is not inconsistent with our argument to avoid making inferences from averages, as the suggested
47
48 averaging would occur within rather than across persons. Other strengths of our approach are
49
50
51
52
53
54
55
56
57
58
59
60

1
2
3 clear: it requires no advanced statistical knowledge, is easy to implement and interpret, and
4
5 therefore, is easy to communicate.
6

7
8 Another limitation is that we used only one moral judgment paradigm to test method-
9
10 based noise explanations for the group-to-person generalizability problem. Additionally, much
11
12 research in moral cognition—including our current experiments (see SOM)—utilizes on-the-fly
13
14 measurement practices (see Flake & Fried, 2020). Future research is needed to determine
15
16 whether method manipulations fail to remedy the problem in other paradigms and areas of
17
18 psychology with better measurement practices. However, as shown earlier, there are obvious
19
20 non-method (and non-measurement) explanations for the problem. Therefore, a person-level
21
22 approach should still be used in more rigorous disciplines to ensure generalizability. Relatedly,
23
24 we note that our proposed method does not solve the problem of measurement error. Therefore,
25
26 as suggested elsewhere (see Speelman & McGann, 2020), future research would benefit from
27
28 collecting many trials per person in order to get observed person-level scores that better
29
30 approximate the true person-level scores.
31
32
33
34

35 Finally, we did not assess the ubiquity of the group-to-person generalizability problem.
36
37 We simply documented (and replicated) existence proofs. Documenting its ubiquity is a
38
39 necessary next step to examine its generalizability (see Simons, Shoda, & Lindsay, 2017). We
40
41 expect the complexity of the experimental designs employed and the phenomenon under
42
43 investigation will be important in determining the ubiquity of the group-to-person
44
45 generalizability. For example, when experiments have factors with more than two levels, or
46
47 multiple factors, the problem should be more likely to occur because the number of possible
48
49 person-level patterns explodes as design complexity increases. In contrast, simple binary choice
50
51 designs common to developmental and comparative work may suffer less from the group-to-
52
53
54
55
56
57
58
59
60

1
2
3 person generalizability problem. Ultimately, we suggest that the problem is an issue for any area
4
5 of psychological research that does not routinely investigate or model person-level data.
6
7

8 In terms of specific subdiscipline predictions, we hypothesize that the problem will be
9
10 least frequent in low-level research, such as visual and auditory cognition/perception, due to
11
12 assumptions that most people share basic physiological similarities, and some lack of conscious
13
14 control, which underlie these disciplines' studied effects. As the content of study becomes more
15
16 concerned with more complex higher-level cognition, the problem should be more prevalent, as
17
18 responses will rely more on individual differences (e.g., values and knowledge). It may even be
19
20 the rule rather than the exception in some areas, such as social cognition. Additionally, social
21
22 psychologists in particular are often interested in phenomena that participants do not have
23
24 introspective access to or are motivated to conceal, leading to the overuse of between-subjects
25
26 designs (rather than the creative use of within-subjects designs). Therefore, social psychology,
27
28 and any other disciplines which habitually rely on between-subjects designs to make inferences
29
30 about psychology may be especially prone to committing the error of assuming that between-
31
32 subjects patterns generalize to the person-level. However, these are empirical questions for
33
34 which future research is needed. We hope this paper gives researchers both the motivation and
35
36 tools to examine group-to-person generalizability in their own areas of interest.
37
38
39
40
41

42 Conclusion

43
44 Psychological scientists often make claims about, and interpret others' claims as being
45
46 about, person-level processes. Sometimes, however, these claims are made from experiments
47
48 that disallow investigation of person-level phenomena. Even when such investigation is possible,
49
50 these claims are typically derived from group-level patterns, interpreted *as if* they reveal the
51
52 pervasiveness of some person-level phenomenon. The current work confirms and builds upon
53
54
55
56
57
58
59
60

1
2
3 previous warnings that this practice can lead to serious errors in inference, as (sets of) group-
4
5 level patterns need not reflect even a simple majority of sampled persons. Put simply,
6
7 psychology is a feature of persons, not averages or distributions. Therefore, person-level design
8
9 and analytic approaches should be customary in psychological science.
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

For Review Only

Footnotes

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

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.

- 1
2
3 Fowler, Z., Law, K.F., & Gaesser, B. (2021). Against empathy bias: The moral value of
4 equitable empathy. *Psychological Science*, *32*(5), 766-779.
5
6
7 Galton, F. (1907). Vox populi. *Nature*, *75*, 450-451.
8
9
10 Grice, J.W. (2021). Drawing inferences from randomization tests. *Personality and Individual*
11 *Differences*, *179*, 110963.
12
13
14 Grice, J.W. (2015). From means and variances to patterns and persons. *Frontiers in Psychology*,
15 *6*, 1007.
16
17
18 Grice, J.W., Barrett, P., Cota, L., Felix, C., Taylor, Z., Garner, S., Medellin, E., & Vest, A.
19 (2017). Four bad habits of modern psychologists. *Behavioral Sciences*, *7*(3), 1-21.
20
21
22 Grice, J.W., Medellin, E., Jones, I., Horvath, S., McDaniel, H., O'lansen, C., & Baker, M.
23 (2020). Persons as effect sizes. *Advances in Methods and Practices in Psychological*
24 *Science*, *3*(4), 443-455.
25
26
27
28
29
30 Hamaker, E. (2012). Why researchers should think “within-person”: A paradigmatic rationale. In
31 M.R. Mehl & T.S. Conner (Eds.). *Handbook of Research Methods for Studying Daily*
32 *Life*, 43-61, NY, NY: Guilford.
33
34
35
36
37 Hamlin, K.J., Ullman, T., Tenenbaum, J., Goodman, N., & Baker, C. (2013). The mentalistic
38 basis of core social cognition: Experiments in preverbal infants and a computational
39 model. *Developmental Science*, *16*(2), 209-226.
40
41
42
43
44 Ince, R, A.A., Kay, J.W., & Schyns, P.G. (2021). Within-participant statistics for cognitive
45 science. *Trends in Cognitive Sciences*, *26*(8), 626-630.
46
47
48
49 Ince, R, A.A., Paton, A.T., Kay, J.W., & Schyns, P.G. (2021). Bayesian inference of population
50 prevalence. *eLife*, *10*, e62461.
51
52
53
54
55
56
57
58
59
60

- 1
2
3 Lakens, D. (2021). The practical alternative to the p-value is the correctly used p-value.
4
5 *Perspectives on Psychological Science*, 16(3), 639-648.
6
7
8 Law, K.F., Campbell, D., & Gaesser, B. (2021). Biased benevolence: The perceived morality of
9
10 effective altruism across social distance. *Personality and Social Psychological Bulletin*,
11
12 48(3), 426-444.
13
14
15 Liew, S.H., Howe, P.D.L., & Little, D.R. (2016). The appropriacy of averaging in the study of
16
17 context effects. *Psychonomic Bulletin and Review*, 23(5), 1639-1646.
18
19
20 Kievit, R.A., Frankenhuys, W.E., Waldorp, L.J., & Borsboom, D. (2013). Simpson's paradox in
21
22 psychological science: A practical guide. *Frontiers in Psychology*, 4, 513.
23
24
25 Kuppens, T. Pollet, T.V. (2014). Mind the level: Problems with two recent national-level
26
27 analyses in psychology. *Frontiers in Psychology*, 5, 1110.
28
29
30 Kurz, A.S., Johnson, Y.L., Kellum, K.K., & Wilson, K.G. (2019). How can process-based
31
32 researchers bridge the gap between individuals and groups? Discover the dynamic p-
33
34 technique. *Journal of Contextual Behavioral Science*, 13, 60-65.
35
36
37 Mayo, D.G. (2018). *Statistical inference as severe testing*.
38
39
40 McManus, R.M., Mason, J.E., Young, L. (2021). Re-examining the role of family relationships
41
42 in structuring perceived helping obligations, and their impact on moral evaluation.
43
44 *Journal of Experimental Social Psychology*, 96, 104182.
45
46
47 Meehl, P.E. (1990a). Appraising and amending theories: The strategy of Lakatosian defense and
48
49 two principles that warrant it. *Psychological Inquiry*, 1(2), 108-141.
50
51
52 Meehl, P.E. (1990b). Why summaries of research on psychological theories are often
53
54 uninterpretable. *Psychological Reports*, 66(1), 195-244.
55
56
57
58
59
60

- 1
2
3 Moeller, J. (2022). Averting the next credibility crisis in psychological science. Within-person
4 methods for personalized diagnostic and intervention. *Journal for Person-Oriented*
5
6 *Research*, 7(2), 53-77.
7
8
9
- 10 Moeller, J. et al. (preprint). Generalizability crisis meets heterogeneity revolution: Determining
11 under which boundary conditions findings replicate and generalize.
12
13
- 14 Navarro, D.J. (2019). Between the Devil and the Deep Blue Sea: Tensions between scientific
15 judgment and statistical model selection. *Computational Brain and Behavior*, 2(1), 28-34.
16
17
18
- 19 Quintana, D.S. (2021). Towards better hypothesis tests in oxytocin research: Evaluating the
20 validity of auxiliary assumptions. *Psychoneuroendocrinology*, 105642.
21
22
23
- 24 Richters, J.E. (2021). Incredible utility: The lost causes and causal debris of psychological
25 science. *Basic and Applied Social Psychology*, 43(6), 366-405.
26
27
28
- 29 Rottman, J., & Young, L. (2019). Specks of dirt and tons of pain: Dosage distinguishes impurity
30 from harm. *Psychological Science*, 30(8), 1151-1160.
31
32
- 33 Schuurman, N.K., Houtveen, J.H., & Hamaker, E.L. (2015). Incorporating measurement error in
34 $n = 1$ psychological autoregressive modeling. *Frontiers in Psychology*, 6, 1038.
35
36
37
- 38 Simons, D.J., Shoda, D.Y., & Lindsay, D.S. (2017). Constraints on generality (COG): A
39 proposed addition to all empirical papers. *Perspectives on Psychological Science*, 12(6),
40 1123-1128.
41
42
43
- 44 Simpson, E.H. (1951). The interpretation of interaction in contingency tables. *Journal of the*
45 *Royal Statistical Society. Series B (Methodological)*, 13(2), 238-241.
46
47
48
- 49 Soter, L.K., Berg, M.K., Gelman, S.A., & Kross, E. (2021). What we would (but shouldn't) do
50 for those we love: Universalism versus partiality in responding to others' moral
51 transgressions. *Cognition*, 217, 104886.
52
53
54
55
56
57
58
59
60

1
2
3 Speelman, C.P., & McGann, M. (2020). Statements about the pervasiveness of behavior require
4
5 data about the pervasiveness of behavior. *Frontiers in Psychology, 11*, 1-16.
6

7
8 Stroessner, S.J., Benitez, J., Perez, M.A., Wyman, A.B., Carpinella, C., Johnson, K.L. (2020).
9
10 What's in a shape? Evidence of gender category associations with basic forms. *Journal of*
11
12 *Experimental Social Psychology, 87*, 103915.
13

14
15 Surowiecki, J. (2005). *The wisdom of crowds*.
16

17
18 Thai, M., Borgella, A.M., & Sanchez, M.S. (2019). It's only funny if we say it: Disparagement
19
20 humor is better if it originates from a member of the group being disparaged. *Journal of*
21
22 *Experimental Social Psychology, 85*, 103838.
23

24
25 Van Rooij, I., & Baggio, G. (2021). Theory before the test. How to build high-verisimilitude
26
27 explanatory theories in psychological science. *Perspectives on Psychological Science,*
28
29 *16(4)*, 682-697.
30

31
32 Wallis, K.F. (2014). Revisiting Francis Galton's forecasting competition. *Statistical Science,*
33
34 *29(3)*, 420-424.
35

36
37 Whitsett, D.D., & Shoda, Y. (2014). An approach to test for individual differences in the effects
38
39 of situations without using moderator variables. *Journal of Experimental Social*
40
41 *Psychology, 50(1)*, 94-104.
42

43
44 Yarkoni, T. (2020). The generalizability crisis. *Behavioral and Brain Sciences, 45*, E1.
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60