From Moral Psychology to Methods Morale: How Studying Moral Obligation Turned into a Duty to Study Methods

Ryan M. McManus

A dissertation

submitted to the Faculty of

the department of Psychology and Neuroscience

in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

Boston College
Morrissey College of Arts and Sciences
Graduate School

March 2023

From Moral Psychology to Methods Morale:
How Studying Moral Obligation Turned into a Duty to Study Methods

Ryan M. McManus

Advisor: Dr. Liane Young

When (moral) 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. This dissertation first investigates a moral cognition hypothesis about the relation between perceptions of relationship obligations and moral evaluations of helping behavior. It is found that people, on average, judge agents who help strangers as more morally good than agents who help family members, but people also judge agents who help strangers instead of family members as less morally good than agents who help family members instead of strangers. Second, methodological issues with these studies are assessed, fixed, and thus the original psychological effect is retested with better experimental designs, measures, and analyses. Third, it is discovered that the moral cognition hypothesis consistently describes the psychology of only a minority of participants. Moreover, it is discovered that most psychologists misinterpret typical group-level analyses as revealing how prevalent a psychological phenomenon is. Finally, a set of simple and flexible methodological and statistical options are offered to better align typical psychological hypotheses with appropriate analyses, enabling researchers to confront this "group-to-person generalizability" problem in their own work.

NOTE:
Consistent with the Psychology and Neuroscience Department's "Three-Paper" option, portions of this proposal are taken directly from three manuscripts which were published or under review during my time at Boston College (McManus, Kleiman-Weiner, & Young, 2020; McManus, Mason, & Young, 2021; McManus, Young, & Sweetman, "invited revision").

Table of Contents

Professional Acknowledgements

Unfortunately, I will not have enough space to acknowledge and thank everyone who helped me professionally during this journey. However, I will do my best to thank those who have left the biggest impact on me, both as a scholar and as a fellow human.

Liane Young:
I first thank Liane for taking a chance on me as PhD student back in 2018, when I had a million ideas but almost no sense of how to sort and get started on any of them. I also thank Liane for being the best academic mentor I have had. There was no point at which she was not interested in research ideas or concerns I shared. Moreover, she always pushed back against things that deserved to be pushed back against, but she also conceded on things that I felt strongly about and argued for; she always made me feel heard. As an example, during the latter half of my PhD, she allowed me to study an area that was novel to both of us (but that I shared with her was very important to me); this ultimately resulted in my favorite research project. I firmly believe that Liane is an exemplar of how academic mentors should interact with their trainees. Finally, Liane has always been more than just another academic mentor. She was like a third parent to me for the past five years, always asking about my personal life, and not just for nicety's sake, but because she cared. When I would tell her about something bothersome in my personal life, she would always check in with me at our next meeting, and she would often email me to check in between meetings. So, on top of being a stellar academic mentor, she is also just a wonderful and caring human being. She taught me what a mentor could and should be, so I have tried to emulate her mentorship when mentoring undergraduate RAs. I honestly cannot think of a bad thing to say about Liane (and I would be highly suspicious of anyone who claimed otherwise).

Joe Sweetman:
I first thank Joe for taking the time to read and respond to the book-length set of emails I sent his way when I first embarked on my methods/statistics research. If he had ignored these emails, I may have never finished my favorite research project. I also thank Joe for acting as my primary mentor while Liane was on medical leave; he did not have to meet with me on a regular basis, nor did he need to be as responsive as he was to my never-ending book-length emails, but he was. Like Liane, Joe always made me feel heard; he pushed back against things that were necessary to push back against, but he, too, conceded on things that I felt strongly about and argued for. I do not believe that I could have finished my favorite project (or the latter half of my dissertation) without Joe's mentorship. If I could have finished it, it would have been much worse. Joe also cared about me as a person, starting every meeting with a 10-15 minute conversation about how life was going in general, offering insight and advice any time he could. Similarly, I can only think of good things to say about Joe.

Katie McAuliffe and Hiram Brownell:
Katie and Hiram were some of my first teachers at Boston College (BC), and now, almost

five years later, they are members of my dissertation committee. I am so thankful that Katie and Hiram were some of my first teachers at BC. In Katie's Origins of Virtue class, I learned, through her demonstration, how to interrogate methods and results of developmental studies in a way that I hadn't previously been exposed to for adult studies. It is fair to say, given the content of my dissertation, that her teaching of how to be critical left an impression on me. In Hiram's Research Methods and Statistics class, I learned introductory statistics for a second time. In this class, I learned, through Hiram's teaching, that our science depends critically on our making appropriate statistical inferences, and how those inferences can be wrong. Like Katie, it is fair to say, given the content of my dissertation, that his teaching of the importance of statistics left an impression on me. Both Katie and Hiram have been extremely responsive to random emails I've sent them over the years, always willing to help. My only regret regarding our relationships is that I didn't directly with them during my time at BC, but I am extremely thankful that they were still willing to be a part of my dissertation committee.

Abraham Rutchick and Debbie Ma:
Abe and Debbie were my first academic mentors. After I had been rejected from graduate school programs in 2015, I reached out and they were willing to meet with me to talk about moving forward. I first thank each of them for welcoming me into their labs, allowing me to gain research experience while I clearly had zero relevant knowledge. Between 2016 and 2018, they served as my MA advisors, and they both allowed me to take lead on projects in their labs, which ultimately resulted in my first few publications. I thank Abe specifically for encouraging me to study whatever I was interested in, and for the countless times he took me out to lunch to talk about and help plan my future. I thank Debbie specifically for teaching me (and being a model of) how to develop a program of research, and, like Abe, for the countless times she met with me to help plan my future and provide me with insight on the hidden curricula of academia. I could not have pursued a PhD without their continued support and mentorship. While Abe and Debbie are still mentors, I now consider them good friends.

Collaborators and Mentees:
There are too many collaborators who I've worked with over the past eight years to name, but I am thankful for every one of them for contributing to work I was involved with. In particular, I am thankful to Max Kleiman-Weiner and Heather Maranges. Max and Heather were two collaborators who played integral roles in projects that I really cared about. Max played a big role in, and co-authored, my very first publication in Liane's lab. He met with us and helped us to design the final experiment, which I think was responsible for ultimately getting the paper published. Heather led a project that I co-authored with her, which was my very first publication without any of my academic mentors. Being a part of this project with her (along with her continued communication and encouragement) helped me to realize that I could be an independent scientist. Heather has also become one of my closest friends in academia. It is fair to say that she is the collaborator who has left the biggest impression on my academic thinking.

Similarly, I've worked with so many mentees over the past eight years, and I am thankful for all of them. In particular, I am thankful to Jordyn Mason, Kayla Carew, and Helen Padilla Fong. Jordyn was the first mentee who I felt comfortable enough with to have her do every step of the project, while I just supervised her and checked in. She showed me how great undergraduate RAs could be, and she eventually co-authored a paper with me and Liane. Because of my experience with Jordyn, I felt comfortable having my next few RAs also do all steps of projects. Kayla not only did almost all the work for an in-prep manuscript, but she did so while also working on her senior thesis. Working with her raised my expectations of RAs even more than Jordyn did. Now, Helen has also done every step of a project for an in-prep manuscript, and she recently presented on this project at a national conference. She exceeded my expectations in every way, and I look forward to seeing what she does next. Working with these young women, seeing (and being a small part of) their progress, might be my proudest moments and accomplishments as an academic. I could not have done any my work without them, or the many other RAs who I've worked with over the years.

Prior and Current Lab Members:

There are too many prior and current lab members to name here, but again, I want to thank those who have either helped me the most or who have left the biggest impression on me. I first want to thank Matt Leitao, my former MA lab member. Matt was the first person to listen and encourage me to do the methods and statistics research that the latter half of my dissertation is about. During one of our biweekly Zoom-beer chats, I mentioned the idea that I ultimately sent to Liane and Joe, and Matt immediately encouraged me to work on it. Moreover, for many of our later chats, he would ask me to update him on the progress that was being made, and he would provide suggestions and argue with me. I don't know if I would have fully developed these ideas without him.

In my current lab, I first want to thank Minjae Kim for being the best senior lab member that I could ask for. No matter what I went to her for over the years, be it questions about the program or department, confusions about fMRI analyses, or simply to commiserate about the lows of grad school, she happily helped. I next want to thank my longest-term officemate, Isaac Handley-Miner. Like Matt, Isaac and I share many interests about metascience in psychology. Isaac was and still is always eager to talk methods and statistics; we've probably wasted weeks' worth of time exchanging and debating ideas in this space while sitting in McGuinn 326, or at Moogy's over beers, but it has probably been my favorite bit of wasted time in grad school. Like Matt, I don't think I would have fully developed the ideas I did without him. I also want to thank Mookie Manalili for being the friend and chess-nemesis that I needed when I no longer wanted to think about research, and instead wanted to have some down time. Mookie was always willing to listen when I was figuring out whether I wanted to continue on in academia, and he was always willing to play a few games of chess over a few beers so that we could both take our minds off of the real world. Finally, I want to thank all the lab managers (Josh, Aditi, and Lizy) who have helped support my research and time in Liane's lab over the past five years. I have probably asked more annoying questions myself than any lab manager

would want to answer from an entire lab, but they always (at least to my knowledge) happily helped. They have done so much behind-the-scenes work to enable research in the lab, and for that, I will always be grateful. I will always have a soft spot for Lizy in particular, as she is the lab manager who really made an effort to be good friends with my Boomer self (as she calls me).

Personal Acknowledgments

Similarly, there are many people who helped me in my personal life while on this journey toward my PhD. Though I do not have enough space to acknowledge and thank all of them, I will do my best to thank those who have left the biggest impact on me.

Amelia Kolenc:
Amelia, my lovely wife, has by far been my best and my most avid supporter since beginning my graduate school journey. I first thank Amelia for encouraging me to move across the country to pursue my dreams; if it weren't for her continued encouragement, I may have never gone back to school. I next thank Amelia for being my primary support system, emotionally, physically, and materially, throughout graduate school. Though these times were not plentiful, when I was at my lowest and considering quitting, she was there to tell me that she believed in me, reassuring me that I was smart enough and worthy of this pursuit. She also always made sure to ask what I was working on, a subtle way to communicate to me that she cared about what I was doing (even though I am almost certain that she did not). Materially, there were times when Amelia paid my rent, flew me home to visit family, bought my groceries, and outfitted me in new clothes. At this point, I must owe her at least a year's salary in time and money. Finally, I am thankful to her for being a role model of persistence and success. No matter how hard circumstances got for her over the years, either personally or professionally, she somehow always managed to thrive. She taught me, through demonstration, that it is possible to persevere, and that hard work eventually pays off. I am so ready to be back in Los Angeles with her and our (dog) family.

Amelia's Family:
Amelia's family was also a huge support system for me during my graduate school journey. Hannah Kolenc, Amelia's closest-in-age sister, probably had the biggest impact on me. Any time I was struggling, like Amelia, she always reassured me that I was smart and worthy of this pursuit. Also, like Amelia, there were times when she bought me meals and drove me to places that I needed to go. Most of all, she was fine with my moving into their shared apartment when I first decided to pursue graduate school, and we became good friends as a result. Starla Kolenc, Amelia's youngest sister, was and is always very interested in my research. Every time I see her, she asks about my latest experiment. When I was sick of a particular set of experiments or paper, having her ask a bunch of follow-up questions often gave me renewed energy, as I was reminded that what I was doing was in fact pretty damn cool. Finally, Amelia's parents have always supported my pursuit. Any time I have seen them over the years, they want to be updated on what I've been doing, out of genuine interest. Overall, Amelia's family has been nothing but helpful and supportive of my endeavor, and for that I will always be grateful.

My Hometown Friends:
I miss all of my hometown friends, and I wish I could have moved them with me when I decided to embark on this journey. But four of my hometown friends, John, Dane, Tom,

and Jarrik deserve special acknowledgements. First, John and Dane have been nothing but supportive of my journey. When I first talked with them about my interests and uprooting my life to move to Los Angeles, they reiterated that there was nothing for me in our hometown. Since then, they have visited me in Los Angeles and Boston, and they both made the trip to the Bahamas for my destination wedding. We all occasionally text each other in group messages, to get updates on one another's lives, offer advice, but most often to talk shit to one another (as we always have). Any time we've gone out together over the past eight years, whether it's back in Pittsburgh, in Los Angeles, or in Boston, they take care of me, telling me "We know you're just a poor college lad." I must owe each of them a few kegs of beer at this point. Similarly, Tom has always been supportive of my journey. He also has visited me in Los Angeles, Boston, and the Bahamas. Like Amelia's sisters, Tom has always expressed interest in what I was doing, and he often texts me just to check in. He also regularly takes me out of academic mode (which I am sometimes very thankful for) to share his opinions and complain about politics and culture in the U.S. If anyone were to uncover the number of text messages between us over the past eight years, they might think I was cheating on Amelia with him. Finally, I want to thank Jarrik, with whom I have a bit of a complicated relationship. I am thankful for his friendship in multiple ways. First, I am thankful for his various reminders that just because I left and became an academic, I'm not above joking around with the guys. Second, I am thankful for the way that he unintentionally motivated me to stay on course while pursuing this degree. When I first decided to move from Pittsburgh, he told me he thought I was crazy for leaving my job to give this a shot. I have remembered this any time I felt like maybe I hadn't made the right decision, and it kept me going. What can I say? Proving someone wrong is a powerful motivator.

My Family:

Finally, I want to thank my family. My mom and dad have been nothing but supportive of my journey, even though I'm almost certain that they don't understand what it is I'm doing. Even though I'm 33 years old, they still send me money for my birthday and holidays, and they both call me somewhat regularly to check in, asking if there's anything I need. My two sisters, Ashley and Kaitlyn, have been like second and third moms to me over the past eight years. When travelling home, I've stayed at their homes, where they've fed me, driven me to places I needed to go, and even gave me money. And like my parents, they still send me money for birthdays and holidays, because "I'm a poor college kid." Where was this behavior when we were kids? Overall, even though my family was sad to see me leave, they've been nothing but supportive, which has made this journey much easier than it could have otherwise been.

General Introduction

> "I know I should have been with you at such a difficult hour… but you would not have liked me to leave them [the lepers] uncared for. I would have had my pleasure of your company only at the cost of hopes and aspirations of the poor, helpless, lepers… Do you really think it is right to have me by your side in these circumstances?"
>
> - Excerpt from *Strangers Drowning* (MacFarquhar, 2015)

Morally speaking, many stories in Larissa MacFarquhar's book *Strangers Drowning* are simultaneously inspiring and confusing. In the excerpt above, readers are introduced to a man who devotes his life to living among and caring for sick stranger—lepers—rendering him a moral exemplar. However, the excerpt is taken from a letter to his wife, who at the time had fallen very ill herself and who, along with their ill infant, had to travel away from her husband for treatment. These details, side by side, seem to detract from the man's moral status and raise key questions about folk moral psychology.

The current proposal builds on recent research that has begun to investigate questions that arise from situations in which the welfare of unknown strangers is pitted against close others (e.g., Everett et al., 2018; Hughes, 2017; Marshall, Wynn, & Bloom, 2020; Marshall et al., 2022). Importantly, this proposal goes beyond prior work—in theory, measures, and methods—by conducting experiments that investigate whether and in what ways: (1) Beliefs about prosocial obligations to help depend on the closeness or relatedness between a helper and the helped; (2) Perceptions of prosocial obligations

shape moral evaluations of helpers; and (3) These proposed relationships survive methodological robustness checks.

Literature Review

Recently, many empirical studies (e.g., Curry, Jones Chesters, & van Lissa, 2019; Curry, Mullins, & Whitehouse, 2019; Everett et al., 2018; Hughes, 2017; Hughes, Creech, & Strosser, 2016; Kurzban, DeScioli, & Fein, 2012; Lee & Holyoak, 2020; Lieberman & Lobel, 2012; Marshall et al., 2020; Simpson, Laham, & Fiske, 2016; Sznycer, De Smet, Billingsley & Lieberman, 2016; Tepe & Aydinli-Karakulak, 2019; Uhlmann, Zhu, Pizarro, & Bloom, 2012; Waytz, Dungan, & Young, 2013; Weidman, Sowden, Berg, & Kross, 2019; Yudkin et al., 2021) and conceptual analyses (e.g., Berry, Lewis, & Sowden, 2021; Curry, 2016; Hester & Gray, 2020; Rai & Fiske, 2011; Schein, 2020; Tomasello, 2020) have focused on, and argued in favor of, the importance of understanding relationship ties in moral judgments and behavior. As the space of relationships is large and multidimensional, morality within this space will be incredibly complex. Here, this terrain starts to be mapped by zeroing in on a specific kind of relationship and its link to morality: kinship and the obligations it entails.

A Sense of Familial Obligation

First, evidence is reviewed suggesting that people report being more likely to help and protect their family membered, compared to non-family. It is argued that these first-person intentions point to a sense of obligation that is absent (or at least weaker) in the case of non-family.

Research abounds suggesting that people are more willing to help kin than non-kin. For example, people report being more willing to help a negligent sibling than an

acquaintance who was in danger through no fault of their own (Greitemeyer, Rudolph, & Weiner, 2003). Relatedly, people report being most likely to help a sibling, next most likely to help a cousin, and least likely to help an acquaintance if they had time to help only one person (Burnstein, Crandall, & Rudolph, 1994). Similarly, people report being more willing to help full siblings than half- or step-siblings (Sznycer et al., 2016), and family members compared to friends and strangers (Passarelli & Buchanan, 2020). Finally, in a convincing behavioral experiment, Madsen et al. (2007) monetarily incentivized engagement in uncomfortable physical exercise (i.e., wall squats), paying more to those who held the position for longer. People held the uncomfortable position for longer when the beneficiary was a 50% genetic relative (e.g., parent or sibling) than when the beneficiary was a 25% (e.g., grandparent) or a 12.5% genetic relative (e.g., cousin).

People are also sensitive to these distinctions when contemplating decisions that would protect their kin. For example, people were more certain that they would report a stranger than a sibling for committing identical crimes (Lee & Holyoak, 2020; Soter et al., 2021), an effect that is enhanced when people imagine a distant other versus a close other committing severe crimes (Weidman et al., 2019). In other work, people reported being less likely to "blow the whistle" (Waytz et al., 2013) and more willing to pay for a transgressor's crime (Linke, 2012) when hypothetical perpetrators were family members, compared to close friends, acquaintances, or strangers. People's predictions about their own behavior in sacrificial moral dilemmas also reveals their consideration of whether the to-be-sacrificed others are family (Bleske-Rechek, Nelson, Baker, Remiker, &

Brandt, 2010; Kurzban, DeScioli, & Fein, 2012; Petrinovich, O'Neill, & Jorgensen, 1993).

Overall, the available evidence suggests that people experience a sense of familial obligation, and perhaps an especially strong obligation to help or protect their closely related family members (e.g., siblings and parents), compared to more distantly related family members (e.g., cousins) or non-family (e.g., strangers).

Perceptions Of Familial Obligations And Their Role In Moral Evaluations

Although it seems clear that people themselves experience a sense of familial obligation, whether people believe *others* ought to adhere to these obligations, and whether these beliefs play a role in their moral evaluations of others, is an area of ongoing research.

For example, adolescents and young adults judged that ultimately not helping genetic relatives was more wrong than not helping non-genetic relatives (see Killen & Turiel, 1998). Relatedly, people judged not donating bone marrow to a critically in-need patient as more wrong when the in-need patient was the potential donor's cousin compared to a stranger (Baron & Miller, 2000). Additionally, when considering stories of agents whose personal desires conflicted with requests to spend time with or support a close other, more people judged an agent's fulfilling their personal desire as unacceptable when the requester was a family member versus a friend (Neff, Turiel, & Anshel, 2002). In the context of prosocial dilemmas, agents who chose to help (or simply endorsed helping) a larger number of strangers instead of a family member were judged as less moral than agents who did the opposite (Everett et al., 2018; Hughes, 2017). Similar effects emerge when people judge the moral acceptability of helping a larger number of

socially distant others (e.g., people in another country) instead of a smaller number of socially close others (e.g., friends; Law, Campbell, & Gaesser., 2021). Although a violation of an obligation was offered as one potential mechanism for the effects in these most recent studies, this hypothesis was not directly tested.

Importantly, these recent studies failed to experimentally control for features of the stimuli that could undermine the proposed mechanism. Specifically, in Everett et al. (2018) and Hughes (2017), family members were not equated with non-family members in terms of their relationship history. An inherent family obligation mechanism could not be disentangled from other features that may naturally covary with non-family versus family relationships, such as the frequency and nature of past and potential future interactions. Therefore, these features (but not inherent obligations to family) may be responsible for prior work's moral judgment effects. Overall, the available evidence suggests that perceptions of familial obligations *may* influence moral evaluations. However, the cited work leaves open the question of whether and exactly how perceived obligations influence moral evaluations. The current proposal, in part, aims to disentangle family obligation effects from other (e.g., reciprocity-based) effects. This is accomplished by using stimuli that describe family members as otherwise stranger-like (e.g., genetically and socially distant; see *Aim 1.1* methods), to attempt to rule out alternative explanations of prior work.

Current Proposal

The general logic behind this proposal's focal hypotheses is explained here. If people are more likely to help related versus unrelated (or closely related versus distantly related) others in hypothetical (e.g., Burnstein et al., 1994) and real-world situations (e.g.,

Madsen et al., 2007), this suggests that people believe that they have a *stronger obligation* to help closely related others than they do to help more distantly related (or unrelated) others. Therefore, if people believe that they have stronger obligations to closely related others than they do to more distantly related (or unrelated) others, then they may also use this information when evaluating others, resulting in differences in third-person moral evaluations. The logic of this assumption is consistent with research showing that first-person moral beliefs shape third-person moral judgments (e.g., Niemi & Young, 2016). More specifically, in contexts where helpers do not have to choose between multiple potential beneficiaries, it is predicted that people will judge a helper who fulfills a stronger obligation as less morally good than a helper who fulfills a weaker (or non-existent) obligation. This hypothesis is broadly consistent with attribution theory (Kelley, 1967), as an obligation is a situational feature that makes it relatively more difficult for a third-party judge to infer whether the helper has true prosocial motives. However, in contexts where helpers do (or must) choose between multiple potential beneficiaries, it is predicted that people will judge a helper who fulfills their stronger obligation as more morally good than a helper who fulfills their weaker (or non-existent) obligation. This hypothesis is broadly consistent with Relationship Regulation Theory (RRT; Rai & Fiske, 2011) and Morality-as-Cooperation (MAC; Curry, 2016). RRT suggests that communal sharing relationships (like those with family) carry with them inherent obligations that, if violated, will be judged negatively; similarly, MAC suggests that helping family is considered a universal moral good.

Results of the current proposal will shed light on whether and how third-person moral evaluations are influenced by prescriptive (i.e., obligation) beliefs. Although this

work focuses only on third-person perceptions, it may have practical implications for first-person prosociality and its promotion. For example, prosocial behavior is often zero-sum; the more one attends to or donates to distant strangers, the less one can attend to or give to family members or closer others. If the proposed hypotheses are supported (i.e., that third-person moral judgments are sensitive to whether specific others are being helped period versus being helped at the expense of specific others), this might result in counterintuitive downstream behavioral consequences. Specifically, attempts to convince people that it is rational and morally right to treat strangers and non-strangers (e.g., family) similarly—as in the "effective altruism" movement—may fail insofar as they make relationship obligations salient. That is, if people indeed believe those who help distant others at the expense of close others are morally worse than those who do the opposite, then, when reminded of their own relationship-based obligations, they may think of themselves as somewhat immoral if they were to use their time and/or resources to help distant others at the expense of their own close others. To the extent that being reminded of one's relationship-based obligations activates such self-facing moral judgments, this could result in people making more decisions to use one's resources *only* for close others. If the proposed hypotheses are indeed supported, then perhaps proponents of large-scale, impartial prosociality will be most effective with messaging that simply communicates the good one can do for strangers without making salient the simultaneous loss for one's family and closer others.

Study 1: Do (Relationship) Obligations Structure Moral Judgment?

AIM 1.1: Determine whether relationship information affects moral judgments of prosocial agents across different helping contexts.

Hypothesis 1.1.

Agents who help distant others (i.e., strangers) will be judged as more morally good than agents who help close others (i.e., family members), but agents who help distant others instead of close others will be judged as *less* morally good than agents who help close others instead of distant others. This result would provide indirect evidence that relationship obligations influence moral judgments. Specifically, going above and beyond one's obligation should result in more positive judgments, whereas violating one's obligation should result in less positive judgments.

Study 1.1 Methods

Participants. Participants were (Study 1.1a = 234 + Study 1.1b = 235 + Study 1.1c = 330) 769 U.S. residents compensated via Amazon's Mechanical Turk. Participants who failed to correctly answer memory checks and attention checks were excluded from all analyses, resulting in (Study 1.1a = 209 + Study 1.1b = 193 + Study 1.1c = 304) 706 analyzable responses.

Design and Procedure. Technically, Study 1.1 is based on three separate experiments that slightly varied the conditions under which participants made judgments. Studies 1.1a-b used a 2 (Relation: Stranger vs Family) x 2 (Choice Context: No Choice vs Choice) within-subjects design. Study 1.1c used a 2 (Relation: Stranger vs Family) x 3 (Choice Context: No Choice vs Choice vs Failed) within-subjects design. However, for the

present purposes, only the conditions that mirror Studies 1.1a-b are investigated in Study 1.1c. In Studies 1.1a-b, participants were presented with eight different scenarios, two from each condition. In Study 1.1c, participants were presented with only four scenarios, one from each condition. Participants were asked to make judgments of an agent who helped a: (1) stranger, (2) a family member, (3) a stranger *instead of* a family member, and (4) a family member *instead of* a stranger (see Table 1 for an example stimulus and its experimental variants). Importantly, family members were always described as otherwise "stranger-like," to isolate the effect of family membership as opposed to other potential factors (e.g., past social interactions / help, expectations of future social interactions / help, etc.). The presentation order of condition order was randomized across participants, and no scenario was repeated across conditions within the same participant. After each scenario, participants were asked to make moral character judgments (and trustworthiness judgments in Studies 1.1a-b) of the helper.

Table 1. Example stimulus and its experimental variants in Studies 1.1-1.2.

	Stranger	Family
No Choice	A new tenant is moving into an apartment down the hall from John. She is a stranger. John helps his new neighbor move their furniture in.	A new tenant is moving into an apartment down the hall from John. She is his cousin whom he has not seen or spoken to in years. John helps his cousin move their furniture in.
Choice	Two new tenants are moving into two separate apartments down the hall from John. One new tenant is a stranger. The other is his cousin whom he has not seen or spoken to in years. Rather than helping his cousin, John helps his other new neighbor move their furniture in.	Two new tenants are moving into two separate apartments down the hall from John. One new tenant is a stranger. The other is his cousin whom he has not seen or spoken to in years. Rather than helping the stranger, John helps his cousin move their furniture in.

Measures. After reading each scenario, participants made moral character judgments of the agent (Study 1.1a: *1 = extremely bad* to *4 = neither bad nor good* to *7 = extremely good*; Studies 1.1b-c: *1 = not at all good* to *5 = extremely good*), as well as trustworthiness judgments (Study 1.1a: *1 = extremely untrustworthy* to *4 = neither untrustworthy nor trustworthy* to *7 = extremely trustworthy*; Study 1.1b = *1 not at all trustworthy* to *5 = extremely trustworthy*). These measures were chosen because these characteristics appear most important in person perception (Cottrell, Neuberg, & Li, 2007; Goodwin, 2015; Goodwin, Piazza, & Rozin, 2014). Importantly, in Studies 1.1a-b, the two judgments of the same condition (e.g., two moral character judgments of agents who helped a stranger) were averaged together for each participant in order to conduct the main analyses. This indexing did not apply to Study 1.1c because there was only one judgment per condition per participant.

Statistical Power. Each of the three sample sizes yielded at least 80% power to detect within-subjects simple comparison effect sizes of $dz = 0.20$ (Faul et al., 2007).

Study 1.1 Results

Both moral character and trustworthiness judgments were highly correlated across studies. Therefore, only moral character judgment analyses are reported here (but analyses are qualitatively similar for trustworthiness judgments).

Across Studies 1.1a-c, repeated-measures ANOVAs revealed a consistent 2 (Relation) x 2 (Choice Context) interaction (Study 1.1a: $F(1, 208) = 34.72$, $p < .001$, $\eta_p^2 = .14$; Study 1.1b: $F(1, 192) = 42.63$, $p < .001$, $\eta_p^2 = .18$; Study 1.1c: $F(1, 303) = 20.27$, $p < .001$, $\eta_p^2 = .12$) on moral character judgments. To understand this interaction, and to

test *Hypothesis 1.1*, simple effects were investigated within each level of the Choice Context factor (see Figure 1).

In Study 1.1a, agents who helped a stranger were judged as more morally good (M = 6.28, SD = 0.81) than agents who helped family (M = 6.07, SD = 0.91), $t(208)$ = 3.55, $p < .001$, Cohen's d_z = 0.25, 95% CI [0.11, 0.38], Cohen's d_{av} = 0.24, 95% CI [0.11, 0.37], but agents who helped a stranger instead of kin (M = 4.91, SD = 1.18) were judged as less morally good than agents who helped family instead of a stranger (M = 5.34, SD = 1.01), $t(208)$ = -4.88, $p < .001$, Cohen's d_z = -0.34 [-0.20, -0.48], Cohen's d_{av} = -0.39 [-0.23, -0.56].

Similarly, in Study 1.1b, agents who helped a stranger were judged as more morally good (M = 4.23, SD = 0.70) than agents who helped family (M = 4.02, SD = 0.74), $t(192)$ = 4.97, $p < .001$, Cohen's d_z = 0.36, 95% CI [0.21, 0.50], Cohen's d_{av} = 0.30, 95% CI [0.18, 0.42], but agents who helped a stranger instead of family (M = 3.08, SD = 0.83) were judged as less morally good than agents who helped family instead of a stranger (M = 3.38, SD = 0.83), $t(192)$ = -4.42, p = < .001, Cohen's d_z = -0.32 [-0.17, -0.46], Cohen's d_{av} = -0.36 [-0.19, -0.52].

In Study 1.1c, again, agents who helped a stranger were judged as more morally good (M = 4.24, SD = 0.81) than agents who helped family (M = 4.14, SD = 0.79), $t(303)$ = 2.20, p = .029, Cohen's d_z = 0.13, 95% CI [0.01, 0.24], Cohen's d_{av} = 0.11, 95% CI [0.01, 0.22], but agents who helped a stranger instead of family (M = 3.02, SD = 1.04) were judged as less morally good than agents who helped family instead of a stranger (M = 3.45, SD = 0.93), $t(303)$ = -5.85, p = < .001, Cohen's d_z = -0.34 [-0.22, -0.45], Cohen's d_{av} = -0.43 [-0.28, -0.58].

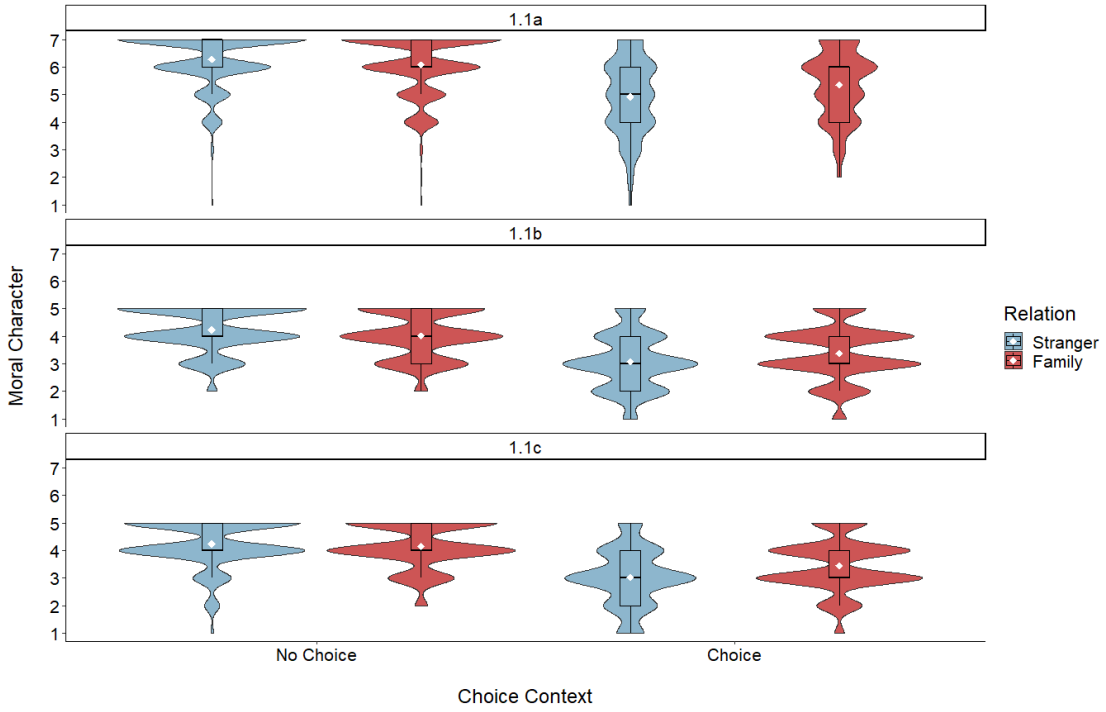

Figure 1. Violin plot of judgments of moral character (split by dataset).

Study 1.1 Discussion

Across all three studies, *Hypothesis 1.1* was supported. Specifically, people judged agents who helped a stranger as more morally good than agents who helped a family member, but they also 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. While obligation judgments were not measured, this study controlled for features of stimuli that could have been responsible for the reported moral judgment effects. Therefore, these results provide indirect, but not direct, evidence for the relationship obligation hypothesis. Therefore, the purpose of *Aim 1.2* is to directly investigate whether inferences about differential obligation underlie the moral character judgment effects found in *Aim 1.1.* This will be accomplished by directly measuring perceptions of obligations.

AIM 1.2: Assess whether social relationship information affects moral judgments due to perceived relationship-based obligations being violated or fulfilled across different helping contexts.

Hypothesis 1.2.

Agents who help distant others will be judged as fulfilling an obligation less than agents who help close others, and agents who help distant others instead of close others will be judged as fulfilling an obligation less (or violating an obligation more) than agents who help close others instead of distant others. These findings would suggest that the different-in-direction moral character judgment effects are due to identical-in-direction obligation judgment differences. Together with *Hypothesis 1.1,* this finding would suggest that people's moral character judgments in helping contexts hinge upon perceived obligations being violated or fulfilled.

Study 1.2 Methods

Participants. Participants were (Study 1.2a = 333 + Study1.2b = 217) 550 U.S. residents and compensated via Amazon's Mechanical Turk. Participants who failed to correctly answer memory checks and attention checks were excluded from all analyses, resulting in (Study 1.2a = 305 + Study 1.2b = 192) 497 analyzable responses.

Design and Procedure. Technically, Study 1.2 is based on two separate experiments that slightly varied the conditions under which participants made judgments (see *Measures*). Both studies used a 2 (Relation: Stranger vs Family) x 3 (Choice Context: No Choice vs Choice vs Failed) within-subjects design. However, for the present purposes, only the conditions that were investigated in *Aim 1.1* are investigated (i.e., the No Choice and Choice conditions). Across both studies, participants were presented with four different

scenarios, one from each of the four conditions investigated in *Aim 1.1*. The presentation order of condition order was randomized across participants, and no scenario was repeated across conditions within the same participant. After each scenario, participants were asked to make moral character judgments of the helper and to judge the extent to which the helper fulfilled or violated an obligation they had.

Measures. After reading the scenario, participants made moral character judgments of the agent (Study 1.2a: *1 = extremely bad* to *5 = neither bad nor good* to *9 = extremely good*; Study 1.2b: *1 = extremely bad* to *4 = neither bad nor good* to *7 = extremely good*), as well as obligation judgments (Study 1.2a: *1 = completely violated* to *5 = neither violated nor fulfilled* to *9 = completely fulfilled*; Study 1.2b: *1 = extremely bad* to *4 = neither bad nor good* to *7 = extremely good*).

Statistical Power. Each of the two sample sizes yielded at least 80% power to detect within-subjects simple comparison effect sizes of $dz = 0.20$ (Faul et al., 2007).

Study 1.2 Results

Moral Character. Across Studies 1.2a-b, repeated-measures ANOVAs revealed a consistent 2 (Relation) x 2 (Choice Context) interaction (Study 1.2a: $F(1, 304) = 42.79$, $p < .001$, $\eta_p^2 = .12$; Study 1.2b: $F(1, 191) = 35.07$, $p < .001$, $\eta_p^2 = .16$) on moral character judgments. To understand this interaction, and to replicate *Hypothesis 1.1*, simple effects were investigated within each level of the Choice Context factor (see Figure 2).

In Study 1.2a, agents who helped a stranger were judged as no more morally good ($M = 7.72$, $SD = 1.29$) than agents who helped family (($M = 7.62$, $SD = 1.27$), $t(304) = 1.57$, $p = .118$, Cohen's $d_z = 0.09$, 95% CI [-0.02, 0.20], Cohen's $d_{av} = 0.08$, 95% CI [-0.02, 0.18], but agents who helped a stranger instead of kin ($M = 5.81$, $SD = 1.69$) were

judged as less morally good than agents who helped family instead of a stranger (M = 6.35, SD = 1.60), $t(304)$ = -4.18, p = < .001, Cohen's d_z = -0.24 [-0.13, -0.35], Cohen's d_{av} = -0.33 [-0.17, -0.48].

In Study 1.2b, agents who helped a stranger were judged as more morally good (M = 6.21, SD = 1.03) than agents who helped family (M = 6.06, SD = 0.99), $t(191)$ = 2.27, p = .024, Cohen's d_z = 0.16, 95% CI [0.02, 0.31], Cohen's d_{av} = 0.15, 95% CI [0.02, 0.28], but agents who helped a stranger instead of kin (M = 4.56, SD = 1.42) were judged as less morally good than agents who helped family instead of a stranger (M = 5.39, SD = 1.22), $t(191)$ = -5.73, p < .001, Cohen's d_z = -0.41 [-0.27, -0.56], Cohen's d_{av} = -0.58 [-0.37, -0.78].

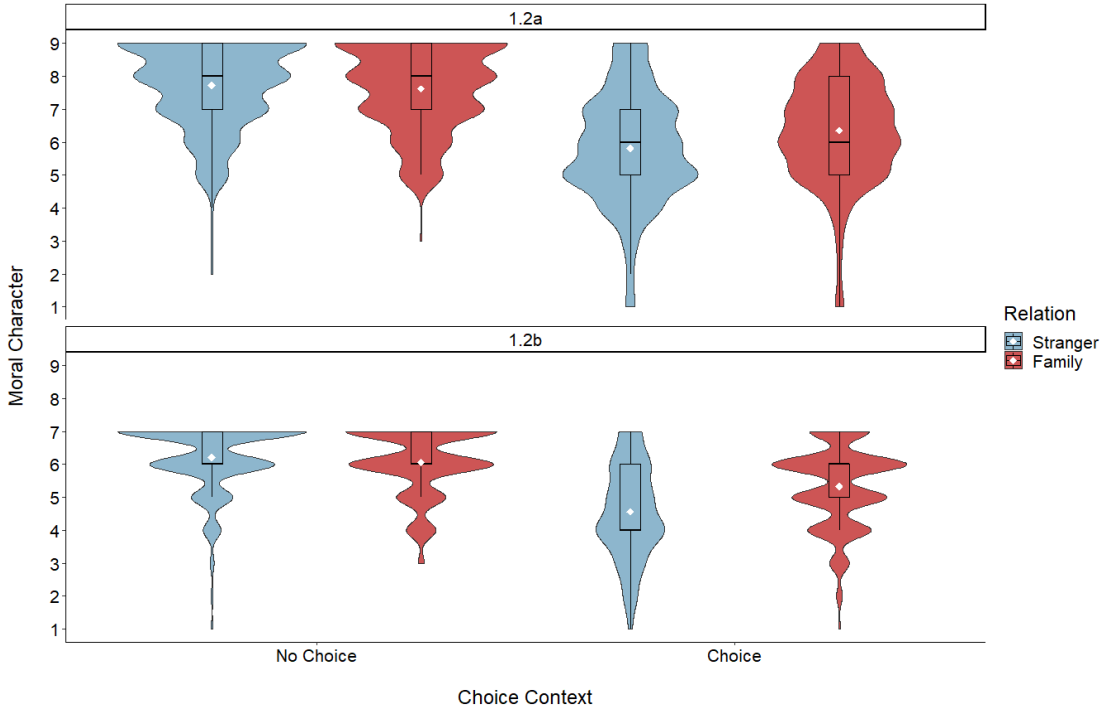

Figure 2. Violin plot of judgments of moral character (split by dataset).

Obligation. Interactions within Studies 1.2a-b were not investigated for obligation judgments because the goal of analyzing obligation judgments was to, in line with

Hypothesis 1.2, demonstrate that the same-direction effect emerged for obligation judgments within each level of the Choice Context factor (see Figure 3).

In Study 1.2a, agents who helped a stranger were judged as fulfilling an obligation less (M = 6.51, SD = 1.67) than agents who helped a family member (M = 6.89, SD = 1.73), $t(304) = -3.96$, $p < .001$, Cohen's $d_z = -0.23$ [-0.11, -0.34], Cohen's $d_{av} = -0.22$ [-0.11, -0.33]. Similarly, agents who helped a stranger instead of a family member were judged as fulfilling an obligation less (M = 5.20, SD = 1.82) than agents who helped a family member instead of a stranger (M = 6.27, SD = 1.71), $t(304) = -7.57$, $p < .001$, Cohen's $d_z = -0.43$ [-0.32, -0.55], Cohen's $d_{av} = -0.60$ [-0.44, -0.77].

In Study 1.2b, agents who helped a stranger were judged as fulfilling an obligation less (M = 5.23, SD = 1.35) than agents who helped a family member (M = 5.55, SD = 1.27), $t(191) = -3.30$, $p = .001$, Cohen's $d_z = -0.24$ [-0.09, -0.38], Cohen's $d_{av} = -0.25$ [-0.10, -0.39]. Similarly, agents who helped a stranger instead of a family member were judged as fulfilling an obligation less (M = 4.01, SD = 1.04) than agents who helped a family member instead of a stranger (M = 5.22, SD = 1.23), $t(191) = -8.52$, $p < .001$, Cohen's $d_z = -0.61$ [-0.46, -0.77], Cohen's $d_{av} = -0.92$ [-0.69, -1.15].

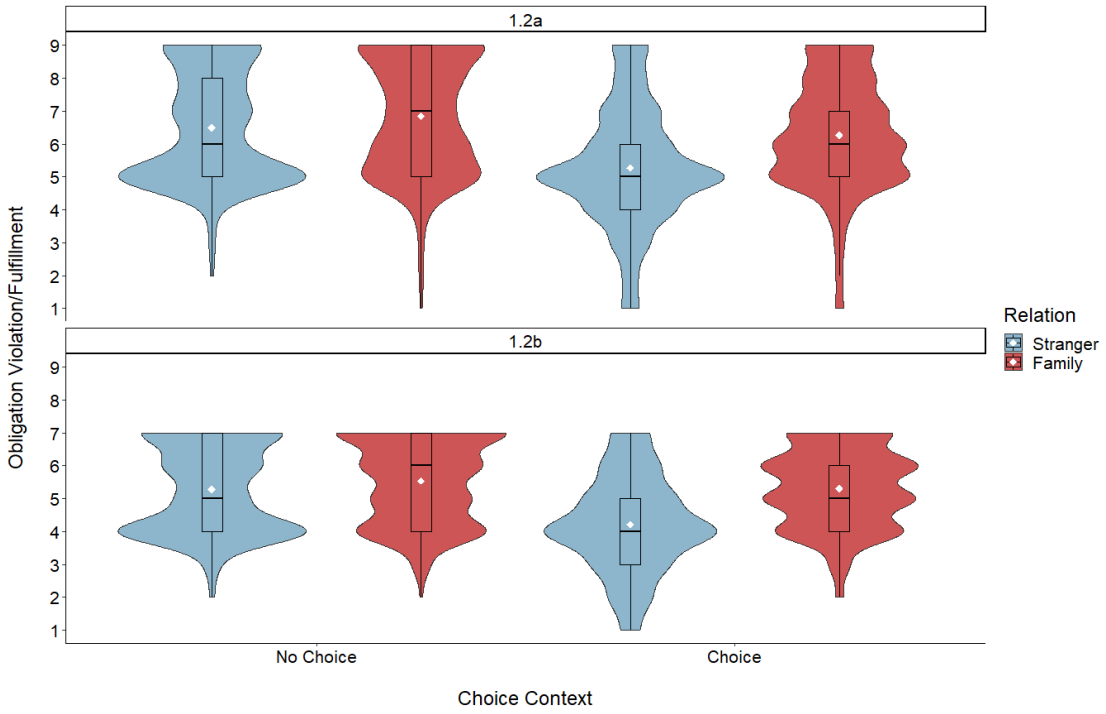

Figure 3. Violin plot of judgments of obligation fulfillment (split by dataset).

Study 1.2 Discussion

Across these two studies, *Hypothesis 1.1* was only partially replicated. Specifically, "people judged agents who helped a stranger as more morally good than agents who helped a family member, but they also 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" was only true in Study 1.2b. In Study 1.2a, people did not judge agents who helped strangers as more morally good than agents who helped family members. However, given the weight of the evidence for this effect (i.e., it emerged in 4 of 5 total studies), it is likely that Study 1.2a's effect was a sampling distribution outlier.

Importantly, the primary purpose of *Aim 1.2* was to directly investigate whether inferences about differential obligation underlie the moral character judgment effects found across all studies. Here, this hypothesis (*Hypothesis 1.2*) was tested by conducting

simple effects tests on obligation judgments at each level of the Choice Context factor. Across both studies, people consistently judged agents who helped strangers as fulfilling an obligation less than agents who helped family members, regardless of the choice context. That is, the same-direction distinction in obligation judgments emerged across contexts even though different-in-direction distinctions in moral character judgments emerged across contexts. These results provide more direct evidence that perceptions of obligations indeed underlie the moral character judgment effects seen across studies.

All studies reported thus far have relied on situations in which partiality (i.e., favoring specific others) is considered justified. However, the purpose of *Aim 1.3* is to investigate situations in which impartiality (i.e., not favoring specific others) is considered justified. Specifically, when occupying roles requiring impartiality, people assume additional obligations to non-family (e.g., professors have obligations to all students, doctors have obligations to all patients, etc.). Because helping family members at the expense of non-family in these contexts may be perceived as displaying inappropriate favoritism, obligation judgments and moral character judgments may take on a different form.

AIM 1.3: Investigate whether boundary conditions exist which change moral character judgments of prosocial agents who fulfill their relationship-based obligations.

Hypothesis 1.3a.

When agents are in workplace-authority positions which require impartiality (e.g., as a coach, professor, or supervisor), within that context, agents who help close others instead of distant others will be judged as less morally good than agents who help distant others instead of close others.

Hypothesis 1.3b.

When agents are in authority positions which require impartiality, within that context, agents who help close others instead of distant others will be judged as fulfilling an obligation less (or violating an obligation more) than agents who help distant others instead of close others. This finding would indicate that the moral character judgment effect is due to an identical-in-direction obligation judgment difference. Together with *Hypothesis 1.3a,* this finding would suggest that people's moral character judgments in workplace helping contexts hinge upon perceived obligations being violated or fulfilled. However, the perceived obligation violation would not be a relationship-based one; instead, the perceived obligation violation would be to the duties required of the agent in that position.

Study 1.3 Methods

Participants. Participants were 226 U.S. residents and compensated via Amazon's Mechanical Turk. Participants who failed to correctly answer memory checks and attention checks were excluded from all analyses, resulting in 196 analyzable responses.

Design and Procedure. Technically, Study 1.3 is a subset of a larger experiment that investigated the same 2 (Relation: Stranger vs Family) x 3 (Choice Context: No Choice vs Choice vs Failed) within-subjects design as *Aims 1.1-1.2*. However, for the present purposes, only the Choice conditions are considered (i.e., those in which an agent helps one person instead of another). Participants were presented with two different scenarios, one from each of the two Choice conditions. Importantly, the content of the scenarios used for Study 1.3 differed from those used in *Aims 1.1-1.2*. Specifically, all scenarios occurred in workplace(-like) contexts, where the target of judgment was in an authority position of some kind (see Table 2 for an example). The presentation order of condition order was randomized across participants, and no scenario was repeated across conditions within the same participants. After each scenario, participants were asked to make moral character judgments of the helper and to judge the extent to which the helper fulfilled or violated an obligation they had.

Table 2. Example stimulus and its experimental variants in Study 1.3.

	Stranger	**Family**
Choice	Debbie, a professor, received two e-mails from students who asked to meet on their only days off to talk about graduate school. Debbie did not recognize one of the student's names; she was a stranger. Debbie recognized the other student's name; she was her cousin's daughter whom she had not seen or spoken to in a while. Instead of e-mailing her cousin's daughter back, Debbie instead set up a meeting to drive to a coffee shop near the other student's hometown to chat more about graduate school.	Debbie, a professor, received two e-mails from students who asked to meet on their only days off to talk about graduate school. Debbie did not recognize one of the student's names; she was a stranger. Debbie recognized the other student's name; she was her cousin's daughter whom she had not seen or spoken to in a while. Instead of e-mailing the student she did not know, Debbie instead set up a meeting to drive to a coffee shop near her cousin's hometown to chat more about graduate school.

Measures. After reading the scenario, participants made moral character judgments of the agent (*1 = extremely bad* to *4 = neither bad nor good* to *7 = extremely good*), as well as

obligation judgments (*1 = extremely bad* to *4 = neither bad nor good* to *7 = extremely good*).

Statistical Power. This sample size yielded at least 80% power to detect within-subjects simple comparison effect sizes of $dz = 0.20$ (Faul et al., 2007).

Study 1.3 Results

Moral Character. Contrary to results from previous studies, in workplace-like contexts, agents in authority positions who helped a stranger instead of kin ($M = 4.64$, $SD = 1.38$) were judged as *more* morally good than agents in authority positions who helped family instead of a stranger ($M = 4.30$, $SD = 1.28$), $t(195) = 2.64$, $p = .009$, Cohen's $d_z = 0.19$ [0.05, 0.33], Cohen's $d_{av} = 0.26$ [0.07, 0.46].

Obligation. Also contrary to results from previous studies, in workplace-like contexts, agents in authority positions who helped a stranger instead of kin ($M = 4.52$, $SD = 1.52$) were judged as fulfilling an obligation *more* than agents in authority positions who helped family instead of a stranger ($M = 4.22$, $SD = 1.47$), $t(195) = 2.10$, $p = .037$, Cohen's $d_z = 0.15$ [0.01, 0.29], Cohen's $d_{av} = 0.20$ [0.01, 0.39].

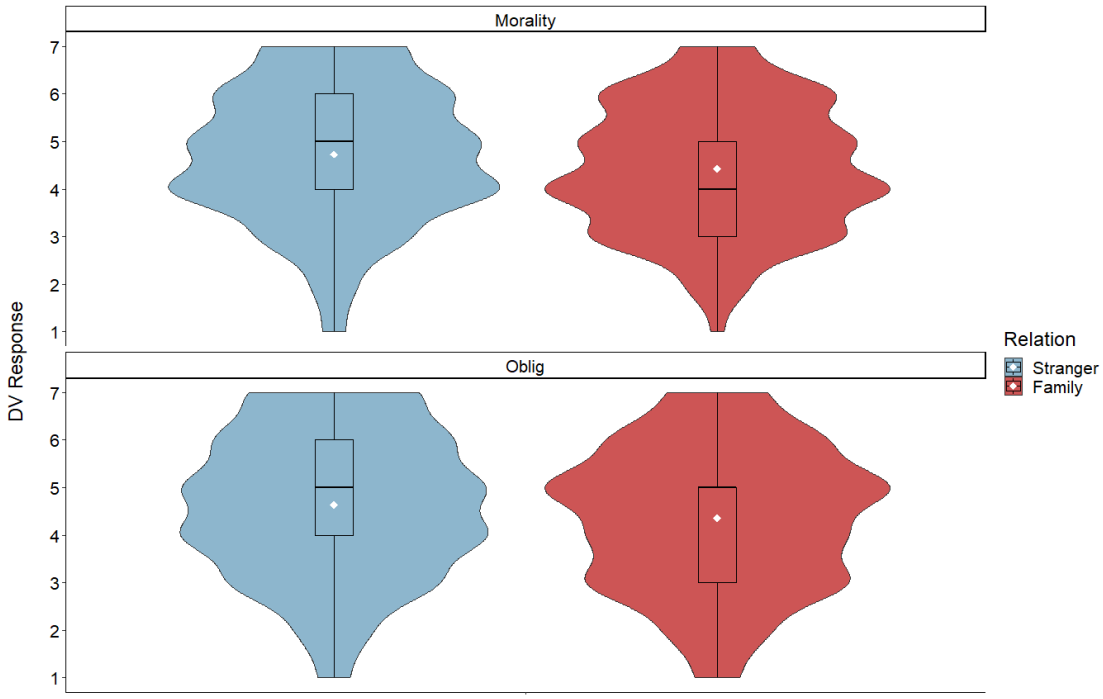

Figure 4. Violin plot of judgments of moral character and obligation fulfillment.

Study 1.3 Discussion

Study 1.3 suggests that there are indeed boundary conditions on how obligation judgments influence moral character judgments. Specifically, when an agent is in an authority position that requires impartiality (e.g., professor, doctor, etc.), people judged them as fulfilling an obligation more when they helped a stranger instead of a family member than when they helped a family member instead of a stranger. These results suggest that people's obligation judgments were not being directed at the relationship obligations of Studies 1.1-1.2, but rather to another obligation. Although Study 1.3 cannot lend direct evidence to the exact type of obligation people had in mind, given the context, people's judgments seem to be referencing an obligation inherent in one's workplace duties. These obligation perceptions also seemed to inform moral character judgments, as people judged agents who helped a stranger instead of a family member as more morally good than agents who helped a family member instead of a stranger.

Together with Studies 1.1-1.2, results of Study 1.3 suggest that people have complex beliefs about when (and which) obligations apply and how they ought to inform moral judgment. The current set of studies demonstrates how context (e.g., being in a workplace role) powerfully prioritizes certain obligations. Outside of workplace contexts, people's moral psychology seems to operate on the intuitive folk wisdom of "family first." Importantly, however, and in support of the idea that people have a nuanced moral psychology, in workplace contexts, people seem to go beyond this intuitive wisdom by reprioritizing the importance of family obligations.

Study 2: Refining the Paradigm

In Studies 1.1-1.2, people judged agents who helped strangers as more morally good than agents who helped family members, but they also judged agents who helped strangers *instead of* family members as *less* morally good than agents who helped family members instead of strangers. Supporting *Hypothesis 1.2* about the underlying mechanism, it was also found that people judge agents who helped strangers (regardless of the choice context) as fulfilling an obligation less than agents who helped family members (again, regardless of choice context). Therefore, these studies suggest that the same-direction distinction in obligation judgments leads to different-in-direction moral character judgments.

While Study 1 provides a comprehensive investigation of familial obligations and moral evaluations, there were two methodological issues which may have had serious consequences on proper inference. First, when participants made obligation judgments, they responded to an item that read "To what extend did X violate or fulfill an obligation they had?" with response options ranging from "completely violated" to "completely fulfilled." Because of the semantic anchors used and how this item was worded, it is unclear how participants interpreted it. For example, what would it mean for one agent to "somewhat" fulfill an obligation, and another agent to "completely" fulfill an obligation? A more interpretable measure of obligation would assess its presence or strength rather than its graded violation or fulfillment. Second, participants made obligation judgments and moral character judgments simultaneously only *after* the outcome of each scenario was known. However, prescriptive judgments (like obligation judgments) are, by their nature, future oriented (Malle, 2021). Because participants made a prescriptive judgment

after the outcome was known, Study 1 could not disentangle whether obligation judgments were inputs to moral character judgments, or moral character judgments were retroactively contaminating obligation judgments.

To address these problems, Study 2 adopts a pre-/post-outcome design (see Marshall et al., 2020; Marshall et al., 2022) that allows measurement of prescriptive (obligation) judgments *before* the outcome and measurement of evaluative (moral character) judgments *after* the outcome. Two important consequences follow from these methodological changes. First, it becomes possible to answer the question of whether, on average, differences in perceived obligation strength correspond to differences in perceived moral character without compromising the hypothesized temporal link between these two judgment types. This eliminates the possibility of moral character judgments retroactively contaminating obligation judgments. Second, it becomes possible for the primary research question to be answered at multiple levels of analysis. Beyond the conclusion of mean differences in each judgment, the research question can be further probed. Specifically, assuming that there is variability in by-relationship obligation judgments and by-relationship moral character judgments, there exists a question about their relation to one another. For example, in the "No Choice" context, is it the case that high discrimination in obligation judgments is associated with high but opposite-signed discrimination in moral character judgments? This difference score correlation analysis can reveal results that are consistent or inconsistent with mean difference analyses. Although these tests were possible in Study 1, the measurement and design issues would have disallowed strong inferences to be made from such analyses.

Study 2 Methods

Study 2's methods will be communicated here. However, results of specific hypotheses will be broken down by *Aims* in the results section.

Participants. All participants were U.S. residents recruited and compensated via Prolific (Palan & Schitter, 2018). It was decided a priori to collect data from 690 participants to obtain 600 analyzable responses. This sample size was chosen to yield 200 responses per three between-subjects conditions (see *Statistical Power*). Once data were collected and an exclusion criterion was applied (i.e., failing an attention check), the final N = 611.

Design and Procedure. Study 2 used a 2 (Relation: Distant vs Close) x 2 (Choice Context: No Choice vs Choice) x 3 (Relatedness Between Beneficiaries) mixed design in which "Relation" and "Choice Context" were manipulated within subjects, whereas "Relatedness Between Beneficiaries" was manipulated between subjects. Participants were randomly assigned to one of three conditions which varied how related the target agents were to *both* beneficiaries. One group of participants read stories involving agents helping strangers and siblings; a second group of participants read stories involving agents helping strangers and cousins; and a third group of participants read stories involving agents helping cousins and siblings. Importantly, the intention between this between-subjects design was not to conduct hypothesis tests by comparing between-subjects conditions. Rather, the between-subjects conditions serve as internal replications of within-subjects effects under slightly varied conditions. Additionally, this design serves the purpose of distinguishing between a general family versus non-family account of obligations' effects on moral judgment and a more granular genetic relatedness account of obligation judgments and effects.

After random assignment to between-subjects condition, all participants were told asked to make judgments of an agent who: (1) helped a genetically distant other (e.g., cousin), (2) a genetically closer other (e.g., sibling), (3) a genetically distant other *instead of* a genetically closer other, and (4) a genetically closer other *instead of* a genetically distant other. Consistent with Study 1, genetic relatives were always described as otherwise "stranger-like," to isolate the effect of relatedness. The presentation order of condition order was randomized across participants, and no scenario was repeated across conditions within the same participants (with all possible permutations of scenarios to conditions being evenly presented across participants). For each scenario, it was presented in a two-stage procedure, with a pre-outcome judgment task (i.e., obligation judgments) followed by a post-outcome judgment task (i.e. moral character judgments).

Measures. In the pre-outcome segment, participants made judgments about how much of an obligation the agent had to help (*0 = none at all* to *100 = a great deal*). In the post-outcome segment, when the agent's helpful behavior was revealed, participants made moral character judgments of the agent (*0 = extremely bad* to *50 = neither bad nor good* to *100 = extremely good*). 0 to 100 scales were adopted to allow participants to make finer-grained distinctions than was possible in Study 1.

Importantly, in contexts where agents did not have to consider whether to help one of two potential beneficiaries (i.e., "No Choice" conditions), participants made only one obligation judgment for each scenario. However, in conditions where agents considered whether to help one of two potential beneficiaries (i.e., "Choice" conditions), participants made two obligation judgments—one judgment about the target agent's obligation to help each potential beneficiary. When conducting analyses on obligation

judgments in the latter conditions, ratings were averaged across the potential beneficiaries of the same relation to the target agent. For example, when participants read a scenario in which an agent ultimately helped a stranger instead of a cousin, and a separate scenario in which an agents ultimately helped a cousin instead of a stranger, participants' two stranger obligation judgments were averaged into one stranger obligation judgment, whereas their two cousin obligation judgments were averaged into one cousin obligation judgment. This indexing did not apply to moral character judgments because participants judged the agent's character only *after* they knew who the agent had helped.

Statistical Power. The sample size goal of Study 2 was to collect at least 200 analyzable participants per between-subjects condition. This determination was based on an internal meta-analysis of the mean differences found in Study 1. Each of the final sample sizes ($N_{Stranger/Sibling}$ = 203; $N_{Stranger/Cousin}$ = 203; $N_{Cousin/Sibling}$ = 205) yielded at least 80% power to detect within-subject mean differences of dz = 0.20, and correlations of r = 0.20, assuming two-tailed tests at an alpha level = 0.05 (Faul et al., 2007).

AIM 2.1: Use a more appropriate experimental design and better measures to determine whether effects of relationship obligations on moral character judgments replicate.

Hypothesis 2.1a.

Agents who have the opportunity to engage in helping behavior will be judged as having a stronger obligation to help close others than distant others, regardless of the choice context. This finding would rule out the possibility that obligation judgment differences in Study 1 were due to retroactively changing perceptions of obligation as a consequence of already knowing the outcome of the potential helping situation.

Hypothesis 2.1b.

Agents who indeed help distant others will be judged as more morally good than agents who indeed help close others, but agents who help distant others instead of close others will be judged as less morally good than agents who help close others instead of distant others. This finding would corroborate the moral character findings from Study 1.

Aim 2.1 Results

Analyses will be repeated for each measure within each dataset. As a reminder, "dataset" refers to an independent sample of participants who completed the fully within-subjects design under slightly varied conditions (i.e., varying levels of relatedness between beneficiaries).

Obligation Strength. See Figure 5 for obligation judgments plotted by dataset and condition. In No Choice conditions, agents who could help a stranger were judged as less obligated to help (M = 28.02, SD = 30.19) than agents who could help a sibling (M = 45.34, SD = 33.34), $t(202) = -6.16$, $p < .001$, $d_z = -0.43$ [-0.57, -0.29], $d_{av} = -0.54$ [-0.73, -

0.36]; agents who could help a stranger were judged as less obligated to help (M = 22.62, SD = 29.18) than agents who could help a cousin (M = 40.87, SD = 34.51), $t(202)$ = -6.28, $p < .001$, d_z = -0.44 [-0.58, -0.30], d_{av} = -0.57 [-0.76, -0.38]; and agents who could help a cousin were judged as less obligated to help (M = 43.17, SD = 34.07) than agents who could help a sibling (M = 49.69, SD = 34.84), $t(204)$ = -2.49, p = .014, d_z = -0.17 [-0.31, -0.04], d_{av} = -0.19 [-0.34, -0.04].

In Choice conditions, agents were judged as less obligated to help a stranger (M = 25.27, SD = 22.96) than a sibling (M = 43.29, SD = 28.87), $t(202)$ = -11.48, $p < .001$, d_z = -0.81 [-0.96, -0.65], d_{av} = -0.67 [-0.80, -0.55]; agents were judged as less obligated to help a stranger (M = 25.19, SD = 24.22) than a cousin (M = 38.04, SD = 27.04), $t(202)$ = -10.25, $p < .001$, d_z = -0.72 [-0.87, -0.56], d_{av} = -0.49 [-0.59, -0.39]; and agents were judged as less obligated to help a cousin (M = 41.90, SD = 26.81) than a sibling (M = 48.14, SD = 28.86), $t(204)$ = -7.13, $p < .001$, d_z = -0.50 [-0.64, -0.35], d_{av} = -0.22 [-0.28, -0.16]. Therefore, *Hypothesis 2.1a* is supported.

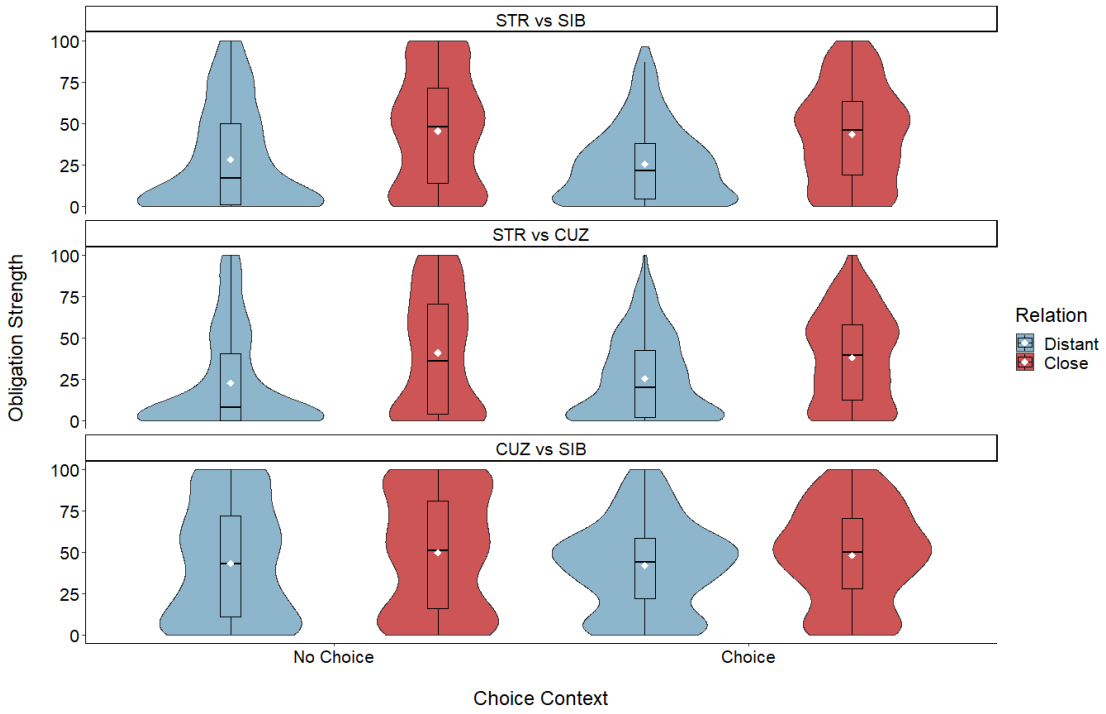

Figure 5. Violin plot of judgments of obligation strength (split by dataset).

Moral Character. See Figure 6 for moral character judgments plotted by dataset and condition. When agents helped strangers and siblings, a 2 x 2 within-subjects ANOVA revealed an interaction pattern on moral character judgments, $F(1, 202) = 44.26$, $p < .001$, $\eta_p^2 = 0.18$. This pattern replicated when agents helped strangers and cousins, $F(1, 202) = 13.14$, $p < .001$, $\eta_p^2 = 0.06$, as well as when agents helped cousins and siblings, $F(1, 204) = 21.49$, $p < .001$, $\eta_p^2 = 0.10$.

In No Choice conditions, agents who helped a stranger were judged as more morally good ($M = 83.85$, $SD = 14.53$) than agents who helped a sibling ($M = 81.07$, $SD = 15.46$), $t(202) = 2.69$, $p = .008$, $d_z = 0.19$ [0.05, 0.33], $d_{av} = 0.19$ [0.05, 0.32]; agents who helped a stranger were judged as more morally good ($M = 83.25$, $SD = 15.56$) than agents who helped a cousin ($M = 80.45$, $SD = 16.13$), $t(202) = 2.77$, $p = .006$, $d_z = 0.19$ [0.06, 0.33], $d_{av} = 0.18$ [0.05, 0.30]; however, agents who helped a cousin were judged no

differently (M = 81.11, SD = 15.48) from agents who helped a sibling (M = 80.76, SD = 16.34), $t(204) = 0.35$, $p = .730$, $d_z = 0.02$ [-0.11, 0.16], $d_{av} = 0.02$ [-0.10, 0.15]. Although this last test was unable to directly support the null hypothesis, it is noteworthy that the point estimates are closer to zero than they are to very small effects that some researchers may consider as theoretically meaningful (i.e., $|d_z/d_{av}| = 0.10$). In Choice conditions, agents who helped a stranger instead of a sibling (M = 58.92, SD = 19.48) were judged as less morally good than agents who did the opposite (M = 68.79, SD = 16.89), $t(202) = -5.90$, $p < .001$, $d_z = -0.41$ [-0.55, -0.28], $d_{av} = -0.54$ [-0.73, -0.35]; agents who helped a stranger instead of a cousin (M = 63.90, SD = 19.45) were judged as less morally good than agents who did the opposite (M = 68.08, SD = 17.71), $t(202) = -2.41$, $p = .017$, $d_z = -0.17$ [-0.31, -0.03], $d_{av} = -0.23$ [-0.41, -0.04]; and agents who helped a cousin instead of a sibling (M = 60.08, SD = 17.71) were judged as less morally good than agents who did the opposite (M = 67.53, SD = 17.08), $t(204) = -5.12$, $p < .001$, $d_z = -0.36$ [-0.50, -0.22], $d_{av} = -0.43$ [-0.60, -0.26]. Therefore, *Hypothesis 2.1b* is only partially supported due to the No Choice null effect in the cousin/sibling dataset.

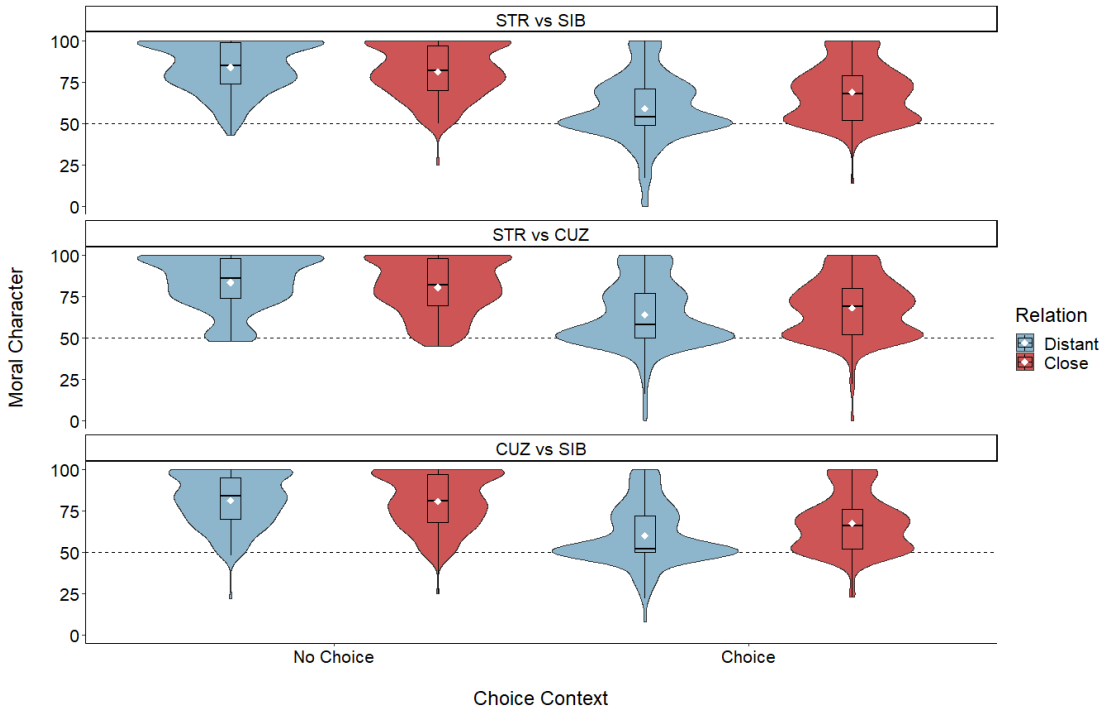

Figure 6. Violin plot of judgments of moral character (split by dataset).

AIM 2.2: Investigate whether differential obligation strength judgments drive differential moral character judgments.

Hypothesis 2.2a-b.

Participants whose obligation judgment differences ("Distant" – "Close") are relatively larger will show larger moral character judgment differences (also "Distant" – "Close"), regardless of choice context. However, the direction of this relationship should differ across choice contexts. Specifically, in the "No Choice" context, there should be a negative correlation between obligation differences and moral character differences. In the "Choice" context, there should be a positive correlation between obligation differences and moral character differences. These findings would provide direct mechanistic evidence for the relationship between perceived obligation strength and moral character judgments.

Aim 2.2 Results

See Figure 7 for difference score relationships plotted by dataset.

No Choice Conditions. In No Choice conditions, obligation difference scores were consistently *un*correlated with moral character difference scores. Specifically, in the Stranger/Sibling dataset, there was no relationship between obligation and moral character difference scores, $r = .07$ [-.07, .21], $t(201) = 1.01$, $p = .316$; in the Stranger/Cousin dataset, there was no relationship between obligation and moral character difference scores, $r = .03$ [-.11, .16], $t(201) = 0.38$, $p = .702$; and in the Cousin/Sibling dataset, there was no relationship between obligation and moral character difference scores, $r = -.03$ [-.17, .10], $t(203) = -0.48$, $p = .635$. This set of results does not support *Hypothesis 2.2a.*

Choice Conditions. In Choice conditions, obligation difference scores were consistently positively correlated with moral character difference scores. Specifically, in the Stranger/Sibling dataset, there was a positive relationship between obligation and moral character difference scores, $r = .18$ [.05, .31], $t(201) = 2.65$, $p = .009$; in the Stranger/Cousin dataset, there was a positive relationship between obligation and moral character difference scores, $r = .39$ [.27, .50], $t(201) = 6.00$, $p < .001$; and in the Cousin/Sibling dataset, there was a positive relationship between obligation and moral character difference scores, $r = .19$ [.06, .32], $t(203) = 2.78$, $p = .006$. This set of results supports *Hypothesis 2.2b.*

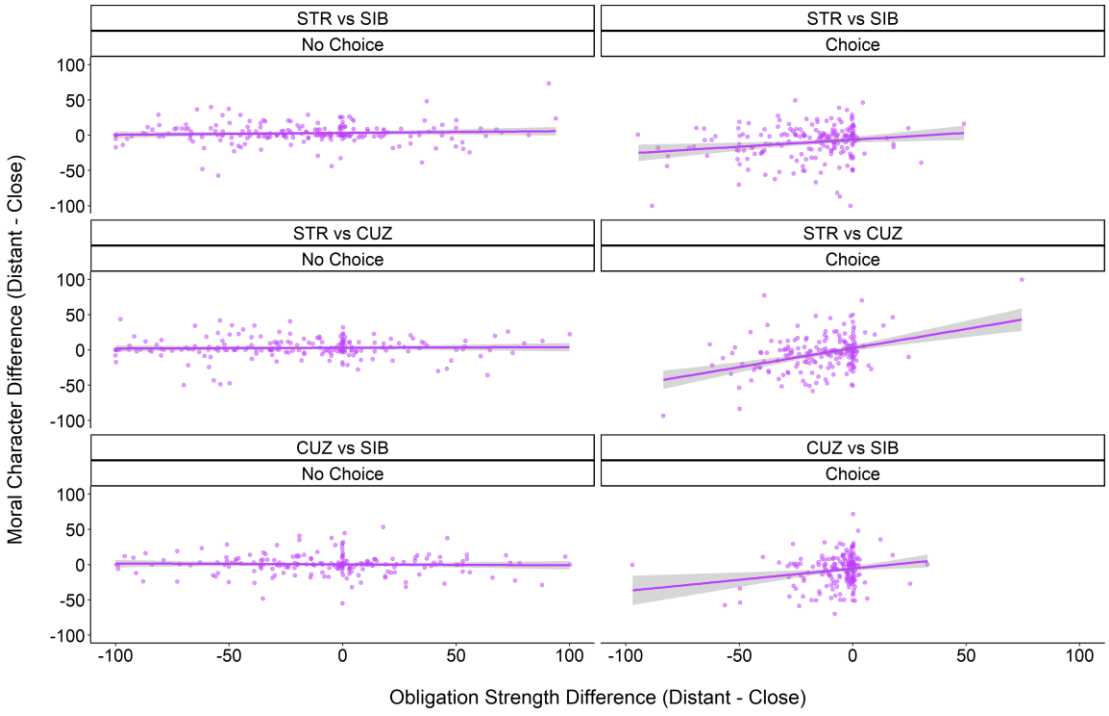

Figure 7. Scatterplots of moral character differences by obligation differences (split by dataset).

Study 2 Discussion

In Study 2, with a new paradigm and new measures, the focal effects from Study 1 were replicated. Specifically, people judged agents who had an opportunity to help a genetically closer other as having a stronger obligation than agents who had an opportunity to help a genetically more distant other. Furthermore, this differentiation seemed to affect moral evaluations. Specifically, people judged agents who helped strangers as more morally good than agents who helped family members (i.e., cousins or siblings), but they judged agents who helped strangers instead of family members as *less* morally good than agents who did the opposite. When examining the generalizability of this context-based reversal in judgments, however, it did not hold when people evaluated *only* agents who helped family members. That is, people did not judge agents who helped cousins as more morally good than agents who helped siblings (even though they judged agents as having stronger obligations to help siblings than cousins). People did, however, judge agents who helped cousins instead of siblings as less morally good than agents who helped siblings instead of cousins, replicating the simple effect found in the other samples.

To better understand the relationship between obligation judgments and moral character judgments, correlations among obligation differences and moral character differences were investigated. Results of these analyses suggest that in contexts where agents do not have to consider a choice about whom to help (i.e., No Choice conditions), differences in obligation strength were unassociated with differences in moral character. That is, there was not an association whereby the more people discriminated between distantly related (or unrelated) other and more closely related other in the obligation judgments, the more the discriminated (in the opposite direction) in the moral character

judgments. This result suggests that the link between differential obligations and differential moral evaluations may be more complicated than previously assumed. Even when there were average differences in both obligation judgments and moral character judgments (e.g., in the stranger/cousin dataset), these correlational relationships were still non-existent, suggesting that perhaps attribution theory (Kelley, 1967) is the wrong lens through which to frame these effects. However, it is also possible that these non-associations were an artefact of the experimental design, as judgments in this context were made across two entirely different scenarios rather than within a single scenario (see Marshall et al., 2020 for a different design). Regardless, these findings highlight the need for more work exploring the underlying mechanisms for differences in moral evaluations in this context.

On the other hand, in contexts where agents had to consider a choice about whom to help (i.e., Choice conditions), differences in obligation strength were indeed consistently positively associated with differences in moral character. Specifically, the more people discriminated between distantly related (or unrelated) others in their obligation judgments (with stronger obligations to help genetically closer others), the more they discriminated in their moral character judgments (with more positive moral character judgments for agents who helped closer others). These findings suggest that obligations may be especially salient in contexts where choices about whom to help can or must occur, and in turn, this is when perceived obligation strength will be especially likely to structure subsequent moral evaluations.

Study 3: Investigating Person-Level (not Group-Level) Responses

Upon further consideration of the analyses and results of Study 2, it was realized that the analysis technique proposed for *Hypotheses 2.2a-b* requires consistent inter-participant magnitude changes in obligation differences and moral character differences. Therefore, the broader aim (i.e., examining whether obligation differences drive moral character differences) may have been unsupported due to an inconsistent inter-participant association between obligation differences and moral character differences. That is, perhaps most participants made the hypothesized directional distinction between moral character judgments, and they varied in the magnitude of that distinction. Most of these participants could have also made the hypothesized directional distinction between obligation judgments, varying in the magnitude of that distinction as well. However, if some participants who made large moral character distinctions made small obligation distinctions, and some participants who made large moral character distinction made large obligation distinctions, then the difference score correlation analysis could yield a null result.

Repeated-measures correlations, which assess the *intra*-participant (as opposed to *inter*-participant) variation among variables, may have been a more appropriate analysis technique (Bakdash & Marusich, 2017). Therefore, in the "No Choice" context, there may indeed be a negative correlation between obligation judgments and moral character judgments. Within the repeated-measures correlation framework, this would mean that, within individuals, of the times that obligation judgments and moral character judgments are made, relatively lower obligation judgments tend to coincide with relatively higher moral character judgments (and vice versa). And mirroring the difference score

correlation results in the "Choice" context, there should be a positive repeated-measures correlation between obligation judgments and moral character judgments. This would mean that, within individuals, of the times that obligation judgments and moral character judgments are made, relatively higher obligation judgments tend to coincide with relatively higher moral character judgments. Such findings, independent of the difference score analyses, would provide person-level mechanistic evidence for the relationship between perceived obligation strength and moral character judgments. These analyses were indeed carried out. However, they yielded similar conclusions as reported in Study 2. Specifically, there was no association between obligation judgments and moral character judgments in the No Choice context, but there was a positive association between these judgments in the Choice context.

It is important to note, however, that repeated-measures correlations necessarily ignore experimental condition information. This analysis technique *only* assesses, within individuals, whether an increase in obligation judgments is related to an increase or decrease in moral character judgments. It is therefore possible that obligation judgments could have driven moral character judgments at the individual level, but they did so in different ways for different people (e.g., most participants' obligation judgments in the "No Choice" context could have been negatively related as predicted, but half of them could have shown higher obligation judgments and lower moral character judgments for strangers relative to family). Moreover, this analysis technique estimates a within-subject regression line that is a best fit across participants; it does not estimate a regression line and produce a single correlation value for each participant. Therefore, even if this analysis technique would have supported predictions from Study 2, there could have been

substantial heterogeneity in the magnitude (or even in the direction!) of person-level correlations.

Another important note of caution applies to all analyses used thus far. Specifically, all analyses were necessarily conducted within specific levels of one experimental factor. Repeated-measures correlations need to be conducted within, e.g., the No Choice level of the Choice Context factor, to investigate the predicted directional relationship. Difference score correlations also need to be conducted within specific levels of the Choice Context factor. Even simple effects tests (i.e., paired t-tests) need to be conducted within specific levels of the Choice Context factor. Therefore, each of these analysis techniques do not investigate whether *the same people* show *both* hypothesized patterns. Consequently, none of these analysis techniques can aid in answering a research question that asks whether most people's psychological experience corresponds to the complex set of predicted patterns described in Studies 1-2.

Study 3, in part, aims to address whether there are other, perhaps better, ways to assess the psychology of moral judgments from Studies 1-2. Recent meta-science research suggests that traditional group-level statistical tests (e.g., ANOVAs and t-tests) can yield conclusions which do a subpar job of describing the psychology of individual persons (Grice et al., 2020; Speelman & McGann, 2020). For example, Speelman and McGann (2020, Figure 2) reported on hypothetical two-cell comparisons, showing how the same group-level effect (i.e., a significant difference between conditions) can be constituted by many different sets of person-level response patterns. Specifically, they document how a simple two-cell effect can describe anywhere between 80-100% of individual participants. Additionally, they claim that an effect can emerge at the group-

level which is representative of even fewer individual participants (e.g., if a small subset of participants shows a large effect while most others show no effect or the opposite effect). Similarly, Grice et al. (2020, Figure 1) reported on a three-level ordinal pattern which was inferred from a one-way repeated-measures ANOVA with follow-up simple comparisons in which all simple comparisons were statistically significant. This set of simple effects was interpreted as strong evidence of a consistent linear decrease in responses as a function of the manipulated factor. However, when investigating how many individual participants showed this three-level, linear decrease in responses, only 24% of participants' responses reflected this pattern.

This issue is not just a problem that occurs in the everyday practice of data analysis; its occurrence is rooted in flawed theoretical assumptions. For results of typical group-level statistical tests to be properly applied to the level of individual persons in psychology research, psychologists must assume the principle of ergodicity (see Fisher et al., 2018; Speelman & McGann, 2020). The APA dictionary of psychology defines ergodicity in the following way (APA, 2022):

> "A principle stating that the average value of a variable over a set of individuals in a defined space of time, such as a sample, will be the same as the average across a long time series of points for a single individual. For example, if ergodicity held for a measure of satisfaction in an organization, the average satisfaction score of all employees in the organization would be the same as the average satisfaction score across a one-year period for one employee."

Put simply, the ergodicity principle dictates that variation around the average of a particular response value (e.g., manipulation-based moral character differences)—which

is *obviously* due to different individuals responding in different ways—is equal to (or a good proxy for) the variation within a single individual's multiple responses over time. In addition to the research conducted by Grice et al. (2020) and Speelman and McGann (2020), metascience research on correlational data suggests that this principle cannot simply be assumed but needs to be tested. Fisher et al. (2018) investigated the bivariate relationships between many clinically relevant psychological constructs, both at the group-level and at the person-level, finding that person-level correlations showed anywhere between two to four times as much variance as group-level correlations. Moreover, person-level correlations suggested that some individuals do not even show the same-direction relation between variables as suggested by the group-level correlations.

Applying this ergodicity problem to the current research, it is possible that the replicable effects reported in Studies 1-2 are unrepresentative of individual participants. The purpose of Study 3.1 is to investigate whether the claims derived from the group-level patterns of Studies 1-2 indeed represent most of the individuals who constitute the group-level patterns. To simplify investigation of the group-level patterns' representativeness in Studies 1-2, Study 3.1 focuses on only the moral judgment pattern (not the combination of obligation and moral judgment patterns). Therefore, this investigation is an extremely liberal test of the research question put forth in Studies 1-2, as it does not consider the theorized, underlying mechanism for the pattern.

AIM 3.1: Using a person-level analytic technique, examine the validity of documented effects of social relationship information on moral character judgments.

Hypothesis 3.1.

When analyzing data using a person-level approach rather than a group-level approach (see Grice et al., 2020; Speelman & McGann, 2020), the 2x2 interaction observed in Studies 1-2 will hold. Specifically, the *majority* of participants from Studies 1-2 will judge agents who help strangers as more morally good than agents who help family members, but will simultaneously judge agents who help strangers instead of family members as less morally good than agents who help family members instead of strangers. If this is not found, it would be justification for investigating the person-level prevalence of other claims in the moral psychology literature.

Study 3.1 Methods

Participants. Participants included in this study are the same participants used in Studies 1-2. Participants' data were only used from the conditions which constitute the predicted 2x2 interaction on moral character judgments.

Design and Procedure. The design and procedure that participants completed were identical to those reported in Studies 1-2.

Measures. The dependent variable of interest was moral character judgments, measured on various scales across studies (i.e., 1-5, 1-7, and 1-9 Likert-like scales, as well as 0-100 sliding scales). Specifically, participants were asked, "How morally bad or good is [target] as a person?" Values below the mid-point indicated "somewhat" to "extremely bad," whereas values above the mid-point indicated "somewhat" to "extremely good."

Analytic Procedure. In order to investigate whether the interaction effect in Studies 1-2 indeed represented most participants, approaches from recent meta-science research were adopted (Grice et al., 2020; Speelman & McGann, 2020). Variables were created in each dataset that divided participants into categories of response patterns. For each participant, each simple effect, as well as the interaction, were described by a directional pattern. The No Choice simple effect was computed by subtracting the "helped a family member" ratings from the "helped a stranger" ratings, whereas the Choice simple effect was computed by subtracting the "helped a family member instead of a stranger" ratings from the "helped a stranger instead of a family member ratings." The interaction score was computed by subtracting the Choice effect score from the No Choice effect score. Therefore, a 2x2 design can lead to 13 different qualitative response patterns (see Table 3; see row 6, specifically, for the predicted person-level pattern). Once it is determined how many participants in each sample match the sample's group-level patterns, it can be determined, descriptively, whether the majority of participants' responses match the predicted pattern.

Table 3. Example hypothetical participants, showing all possible patterns in Studies 1-2.

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

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

Study 3.1 Results

In Study 1, the group-level crossover interaction and opposite-signed simple effects emerged in four datasets. In Study 2, the group-level crossover interaction and opposite-signed simple effects emerged in two datasets. Of these six total datasets, the group-level crossover interaction and two opposite-signed simple effects *never* described a simple majority of participants. Strikingly, the study-to-study estimates were considerably lower than a simple majority, ranging between 6(!)% – 31% of participants. See Figure 8 for person-level pattern breakdowns by study.

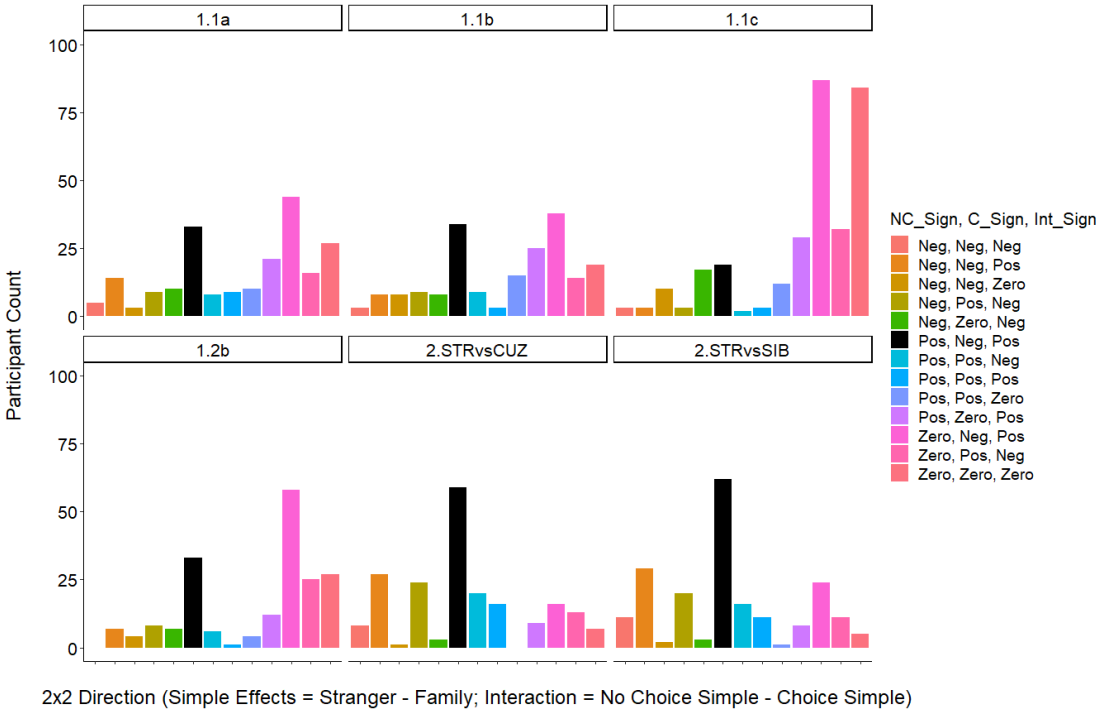

Figure 8. Person-level patterns from Studies 1-2. The black bar represents the group-level effect.

Study 3.1 Discussion

Even though the full set of group-level effects from Studies 1-2 were replicable, the focal claim of each paper—that people judge agents who help strangers as more

morally good than agents who help family, but they also judge agents who help strangers instead of family as less morally good than agents who help family instead of strangers—failed to ever describe even a simple majority of sampled individuals. How exactly can this happen? The answer to this question lies within how a statistical interaction (and its decomposition) is typically tested in psychological data.

A crossover interaction, like the one found in Studies 1-2, is typically tested for using a 2x2 repeated-measures ANOVA. Importantly, the interaction can be assessed using t-tests as well, which can help to explain the group-level versus person-level 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 (e.g., within No Choice and Choice conditions). The paired-samples t-test method is completed by conducting a t-test on the two difference scores. The one-sample t-test method involves an extra step, creating a third difference score variable—the interaction score—by subtracting the second simple effect's differences score from the first simple effect's difference score (e.g., the Choice difference score is subtracted from the No Choice difference score). The one-sample t-test method is completed by conducting a t-test (against zero) on the interaction scores. If either t-test returns a below-alpha p-value, then an interaction effect exists. Importantly, in this context, the p-value from both the t-test methods would be identical to one another and to the p-value of the ANOVA's interaction F-test, as all methods are testing for a difference in differences.

This is problematic for the claims made in Studies 1-2, as well as for any other claims derived from a 2x2 interaction test. As Table 3 shows, there are five distinct patterns which yield an interaction value that is directionally consistent with the

interaction value that emerges for the claimed pattern in Studies 1-2 (i.e., a positive interaction value). Therefore, it is possible that most interaction values which underlie a group-level interaction are not constituted by patterns that are consistent with the two, constituent group-level simple effects. In Study 2, for example, the number of participants who showed a positive interaction value underlied by the two predicted simple effects was consistently *lower* than the number of participants who showed a positive interaction value underlied by all other simple effects combinations. This led to more than half of the sample having a positive interaction value even though only ~30% of participants showed the predicted crossover pattern. The reason for this interaction's consistent emergence across Studies 1-2 is, in addition to a majority of the samples having the positive interaction value, that the nature of a crossover interaction dictates that participants who show it are likely to have higher interaction scores than all other participants. Therefore, this subset of participants will bias the average interaction value to be higher when compared to a situation in which the same subset of participants showed a different set of simple effects while still showing a (non-crossover) positive interaction value.

For the group-level simple effects, there exists another problem. *Sets* of group-level inferential tests cannot provide evidence about the co-occurrence of person-level simple effects. Because the units of analysis for a single paired-samples t-test are the person-level differences scores, two separate paired-samples t-test cannot connect units across analyses (and has already been established, the connection of units via the interaction test has its own problems). Therefore, even if both simple effects emerge

through typical statistical tests (i.e.., multiple t-tests), there is no guarantee that most (or *any*) participants show both simple effects.

Overall, person-level analyses of Studies 1-2 is consistent with demonstrations from recent meta-science research (Grice et al., 2020; Speelman & McGann, 2020). Even though sets of group-level effects consistently emerged, the predicted set never described even a simple majority of sampled participants. Although this problem occurred in a specific moral psychology paradigm, there is no principled reason that it should only apply to that paradigm. This problem may also occur in other moral psychology paradigms, rendering other published claims false, unrepresentative, or at best, imprecise. Therefore, the purpose of Aim 3.2 is to investigate whether this problem exists in other moral psychology paradigms.

AIM 3.2: Investigate whether a person-level analytic technique renders other findings in the moral psychology literature as problematic.

Hypothesis 3.2.

When analyzing published open data using a person-level approach rather than a group-level approach, some moral psychological claims will be undermined. Specifically, some moral psychological claims derived from typical group-level analyses will be unrepresentative of person-level responses. This finding would be justification for further research on ways to solve the problem.

Study 3.2 Methods

Participants. Participants included in this study are those from four additional (very recently) published papers in moral psychology. Very recent articles were chosen because of the assumption that they are more likely to be methodologically and statistically rigorous than less recent articles, due to the influence of the Open Science movement in social psychology in particular. Three of these papers are from the same subfield as Studies 1-2: how social relationships influence moral judgment or moral reasoning broadly (Fowler, Law, & Gaesser, 2021; Law, Campbell, & Gaesser, 2021; Soter, Berg, Gelman, & Kross, 2021). One of these papers is from a different subfield: how and why qualitatively distinct moral violations differentially influence moral wrongness judgments (Rottman & Young, 2019). These publications were chosen for convenience reasons. They were relevant to my own research projects and all data were posted on OSF and therefore downloadable and re-analyzable.

Designs and Procedures. This section is broken down by each publication. In Fowler et al. (2021), the general research question of interest was: "How do people judge the moral

appropriateness of feeling biased empathy versus equal empathy for those who are close versus distant to them?" Participants in this study were first told to think about someone who was extremely close to them in terms of social distance (e.g., family member) and to think about someone who was extremely distant to them in terms of social distance (e.g., a stranger that they may recognize). Participants reported names for each target and their actual relationship information. Participants then imagined themselves, across various scenarios, feeling empathy for each target who were in similar need of help. After each scenario, participants were instructed to judge how morally right or wrong it was to feel a certain amount of empathy for each target. For current purposes, the conditions of interest were when participants imagined feeling: more empathy for a distant other than a close other, more empathy for a close other than a distant other, and equal empathy for the distant other and close other. Group-level analyses suggested that, on average, participants judged feeling equal empathy as more morally right than feeling more empathy for a close other, and that feeling more empathy for a closer other as more morally right than feeling more empathy for a distant other.

In Law et al. (2021), the general research question of interest was: "How do people judge the moral appropriateness of actual helping behavior toward distant others at the expense of closer others?" Participants in this study read several hypothetical scenarios where an agent had to choose between donating money to a socially distant cause (e.g., overseas) or a socially closer cause (e.g., people in the same town). The scenarios varied in how close the socially closer cause was (same country versus same town versus friend versus family member). After each scenario, participants were instructed to judge how morally acceptable it was for the helping agent to act in the way

they did. Importantly, the helping agent always chose the socially distant cause. Group-level analyses suggested that, on average, participants judged the help toward socially distant causes as less and less morally acceptable as the socially close cause became closer (i.e., Family < Friend < Same Town < Same Country).

In Soter et al. (2021), the general research question of interest was: "In situations where people can protect others who have committed a crime, do people's beliefs about what they should versus would do differ depending on whether the others are socially close versus socially distant?" Participants in this study first thought of four target people, two who were close and two who were distant to them. Then participants provided names for each person and actual relationship information. In the main experimental task, participants read eight "punish-or-protect" dilemmas, with two dilemmas per target. For each dilemma, participants judged how likely they *would be* to protect the target and how much they *should* protect the target. Group-level analyses suggested that, on average, people report being more likely to protect close others than distant others and being more certain that they should protect close others than distant others. Importantly, their analysis yielded a two-way interaction indicating that the "would protect" effect was stronger than the "should" effect, suggesting that people show more partiality in their "would" judgments than their "should" judgments.

In Rottman & Young (2019), the general research question of interest was: "How does the type of moral violation and its frequency affect moral wrongness judgments?" Participants in this study were presented with stories about agents committing various moral violations that varied by moral domain (i.e., harm vs purity) and frequency (e.g., once versus frequently). After learning about each violation, participants judged how

morally wrong the agent's action was. Group-level analyses suggested that, on average, participants judged harmful violations as more wrong when they were done more frequently compared to less frequently, and they judged purity violations as more wrong when they were done more frequently compared to less frequently. Importantly, their analysis yielded a two-way interaction indicating that the effect of frequency on harm was stronger than the effect of frequency on purity violations, suggesting that people's judgments of harm violations are more sensitive to "dosage" than purity violations.

Measures. The dependent variable of interest varied across studies but was always a Likert-like scale or a sliding scale. In Fowler, Law, and Gaesser (2021), participants judged and rated how morally right or wrong it was to feel empathy for various social targets (1 = extremely morally wrong; 5 = neither right nor wrong; 9 = extremely morally right). In Law, Campbell, & Gaesser (2021), participants judged and rated the moral acceptability of donation behaviors to various social targets at the expense of other social targets (1 = completely unacceptable; 9 = completely acceptable). In Soter, Berg, Gelman, and Kross (2021), participants judged and rated both what they would and should do in situations where they could protect or report various social targets who committed crimes (1 = definitely would / should protect; 6 = definitely would / should). In Rottman and Young (2019), participants judged the moral wrongness of various types of moral violations (0 = not at all wrong; 100 = extremely wrong).

Analytic Procedure. To investigate whether claimed effects in these studies represented most participants, the same approach from Aim 3.1 was taken. Specific claims were found by reading the general discussion sections of each article. If a general discussion's claim could not be neatly mapped onto a specific set of group-level tests, or it was

unclear which group-level tests were being referenced, then a specific study's discussion and results sections were read to find a clear mapping between claims and tests. Results of this reading yielded the claims documented in Table 4 (in Results).

Study 3.2 Results

Similar to re-analysis of Studies 1-2, re-analysis of each published claim suggested that the claim does *not* represent a majority of sampled participants. Across the four publications, the proportion of participants who were represented by the claims ranged between 3%(!) - 46%, with most proportions ranging between 20% - 40%. See Table 4 for exact claims from these publications.

Table 4. Quotes, relevant tests, and person-level statistics for Study 3.2.

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

Note: Person-level proportions are the proportion of participants whose response patterns matched the full set of group-level patterns required given the published claim.

Study 3.2 Discussion

In addition to re-analysis of data from Studies 1-2, re-analysis of four other published moral psychological claims suggests that this "group-to-person generalizability" problem may be somewhat pervasive. Across all four publications, the published—and importantly, focal—claims were unrepresentative of most sampled participants. While this evidence may suggest, as others have argued, that psychologists ought to analyze their data differently (Grice et al., 2020; Speelman & McGann, 2020), it could also be argued that all the investigated claims may indeed still be representative of most people.

Specifically, it could be argued that most discrepancies between group-level and person-level analyses are due to methodological features of experiments which can be remedied. That is, most experiments may not be designed to minimize noise and therefore maximize the probability of participants exhibiting the group-level pattern. If such barriers could be addressed, then group-level patterns may better represent person-level patterns. To test this idea, a specific paradigm can first be chosen (ideally a paradigm that has generated consistent group-level replications). Next, this paradigm can be modified in ways that are hypothesized to reduce participant-level noise in responses. Then, these modified versions of the paradigm can be presented to half of all participants while the unmodified version can be presented to the remaining half of participants. Finally, assuming that the group-level effects are consistently replicated, the proportion of participants who are represented by them can be compared to determine whether the methodological modifications improve the correspondence between group-level and person-level analyses. The purpose of Aim 3.3 is to examine this possibility.

AIM 3.3: Assess whether simple (somewhat obvious) methodology adjustments explain and solve the discrepancy between person-level and group-level patterns.

Hypothesis 3.3a.

Simple methodological adjustments will be able to reduce the severity of the problem, leading to less of a discrepancy (or no discrepancy) between group-level and person-level effects. Specifically, methodological adjustments which reduce participant-level noise will lead to higher proportions of participants' responses being aligned with the group-level effects. This finding would yield easy-to-implement suggestions for moral psychology researchers to reduce participant-level noise which in turn would result in better alignment among levels of analysis.

Hypothesis 3.3b.

An alternative hypothesis is that simple methodological adjustments will not solve or reduce the severity of the problem, leading to the same discrepancy between group-level and person-level effects. Specifically, methodological adjustments designed to reduce participant-level noise will lead to similar proportions of participants' responses being aligned with the group-level effects. If it is assumed that participant-level noise was successfully reduced, this finding would be consistent with the notion that there are real and meaningful individual differences in responses for the tested paradigm. Furthermore, this finding would suggest that it is strictly necessary to investigate person-level patterns if the research goal is to make general psychological claims.

Study 3.3 Methods

Participants. For each methods experiment, all participants were U.S. residents recruited and compensated via CloudResearch's "approved participants" list. Each of the four experiments in Study 3.3 were conducted sequentially; therefore, participants who took

part in one methods experiment could not participate in another methods experiment. After excluding participants who failed a pre-task attention check, this led to the following analyzable *N*s: Experiment 1 = 1,247; Experiment 2 = 1,237; Experiment 3 = 1,279; Experiment 4 = 1,437. See *Statistical Power* for sample size justifications.

Designs and Procedures. For each experiment, participants were randomly assigned to one of two conditions, corresponding to the absence/presence of each methodological feature of interest. For all experiments, the experimental paradigm from McManus et al. (2020; 2021) was used to test the effect of each methodological feature. Therefore, in each experiment, there was a 2 (Relation: Stranger vs Cousin) x 2 (Choice Context: No Choice vs Choice) within-subject task in which participants responded to four different stimuli: agent helps a stranger, agent helps a cousin, agent helps a stranger instead of a cousin, and agent helps a cousin instead of a stranger. In all conditions, participants were asked to judge and rate the moral character of the helpful agent. In total, there were four methodological features that were tested as somewhat obvious candidates to explain the group-to-person generalizability issue discussed thus far: the absence/presence of calibration trials; the inability/ability to respond to all stimuli simultaneously; the absence/presence of perfectly matched stimuli; and the inability/ability to "opt out" of using the measures/scales provided. The procedure for, and the underlying logic for manipulating, these specific features is explained below (but see Table 5 [in Study 3.3 Results] for a succinct summary of the underlying logic for manipulating these specific features).

In the calibration trials experiment, participants were randomly assigned to one of two between-subjects conditions. In the "Calibration Trials" condition, participants were

given five calibration trials after initial instructions, whereas in the "No Calibration Trials" condition, participants started the experimental task immediately after instructions. Calibration trial participants made moral character judgments of agents who were described as: an extreme moral saint, someone who is morally good in ways that most people are not, someone whose behavior reflects nothing about their underlying moral character, someone who is morally bad in ways that most people are not, and an extreme moral monster. These stimuli were designed to incentivize use of the entire scale range and therefore give participants anchors for the main experimental task. Calibration trial participants were also given fake (normative) feedback after each calibration trial, explaining whether their response was consistent with most other participants' responses.

This calibration trials experiment serves to potentially solve two problems. First, 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. Second, 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. If participants engage in calibration trials (and receive normative feedback), then they should have similar understanding of similar scale points. Moreover, they should be less likely to use extreme ends of the scale during the main experimental task due to the experimental stimuli being everyday helping behaviors. Therefore, if the group- versus person-level discrepancy is due to noise induced by lack of calibration trials and normative feedback, then participants in the 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 the control condition (i.e., those who do not engage in pre-task calibration trials).

In the simultaneous responses experiment, participants were randomly assigned to one of two between-subjects conditions. In the "Simultaneous Judgments" condition, participants first read through all main task stimuli before making judgments, whereas in the "Sequential Judgments" condition, participants, in typical fashion, made judgments after each stimulus. In the pre-task instructions, simultaneous judgment participants were told that they would only make judgments about the stimuli's characters after reading through all stimuli, and that they would be reminded of what took place in each story when making those judgments. After reading through all stimuli, these participants encountered a page with five slider scales (one as an attention check), each corresponding to a different within-subjects condition, where they made moral character judgments of agents within each stimulus. Above each slider, participants were provided with a summary of each story. Importantly, participants were also instructed to make each judgment while simultaneously considering their judgments of the other agents. Assuming instruction obeyance, any "ties" between conditions (in addition to person-level ceiling or floor effects) should be interpreted as participants truly believing that two stimuli ought to be judged identically.

This simultaneous judgments experiment also serves to potentially solve two problems. First, if participants cannot consider all stimuli simultaneously, then some participants may fail to distinguish between stimuli that they truly which to distinguish

between. Second, 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. If participants are able to consider all stimuli before making judgments and are reminded of the important details of all stimuli before making their judgments, then they should be able to make desired distinctions without problem. Therefore, if the group- versus person-level discrepancy is due to noise induced by being unable to consider stimuli simultaneously, then participants in the experimental condition (i.e., those who can see all stimuli before making judgments simultaneously) should be more likely to show the person-level response pattern that matches the group-level pattern, compared to participants in the control condition (i.e., those who see stimulus after stimulus while making judgments sequentially).

In the matched stimuli experiment, participants were randomly assigned to one of two between-subjects conditions. In the "Matched Stimuli" condition, all stimuli were perfectly matched in content except for the critical within-subject manipulations of "Relation" and "Choice Context." In the "Different Stimuli" condition, all stimuli were different in content across conditions within a single participant. This matched stimuli experiment serves to potentially solve one 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 factors when responding. If participants are given matched-in-content stimuli across all experimental conditions (which vary only the experimental features of interest), then they should be able to focus only on the experimental features of interest.

Therefore, if the group- to person-level discrepancy is due to noise induced by a lack of carefully matched stimuli, then participants in the 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 the control condition (i.e., those who see different-in-content stimuli).

In the ability to "opt out" experiment, participants were randomly assigned to one of two between-subjects conditions. In the "Possible to Opt Out" condition, participants could communicate not wanting to use the provided measurement scale, whereas in the "Impossible to Opt Out" condition, participants were, in typical fashion, unable to communicate this. In the pre-task instructions, opt-out-possible participants were told that they could check a box above the slider scale that read, "I can't use this scale to make this kind of judgment / I don't think it's possible to answer this." They were also shown an example of the slider scale and checkable box with an example stimulus. Importantly, these participants were told that they would not be penalized (i.e., forfeit their payment) for checking the box, and so to choose it if it truly reflected their thinking. During the main task, if participants chose this option, they were directed to a page asking them to consider answering the moral character question with a simpler trinary scale (*Morally bad; Neither bad nor good; Morally good*). If participants opted out again, they were directed to a final page which asked them to explain their reasoning. For current purposes, the focus is only on whether participants opted out of the slider scale (and these participants are excluded from all analyses).

This ability to opt out experiment serves to potentially solve one 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. If participants are given the ability to opt out of using the measurement scale, then this should rid the sample of measure-transforming or randomly responding participants. Therefore, if the group- versus person-level discrepancy is due to noise induced by an inability to opt out of using the measurement scale, then participants in the 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 the control condition (i.e., those who cannot opt out).

Measures. As in Study 2, the dependent variable of interest was moral character judgments, measured on 0-100 sliding scales across all experiments. Participants were asked, "How morally bad or good is [target] as a person?" Values below the mid-point indicated "somewhat" to "extremely bad," whereas values above the mid-point indicated "somewhat" to "extremely good."

Analytic Procedure. To investigate whether claimed effects in these studies represented most participants, the same person-level grouping approaches from Aims 3.1-3.2 were taken. However, rather than testing individual subsamples against a proportion of 50%, the goal of Study 3.3 was to compare experimental conditions to control conditions within each methodological experiment. To test whether the proportion of participants who showed the group-level was higher in the experimental condition, two-sample equality of proportion test will be conducted within each experiment.

Statistical Power. The sample size goal for each experiment was 590 participants per between-subjects condition. This sample size yields high-powered two-sample equality of proportion tests, the focal test to determine whether the methodological manipulation increases the proportion of participants showing the crossover interaction (assuming the set of group-level effects from Studies 1-2 replicate). To conduct a power analysis, the baseline proportion of participants who showed the crossover interaction in McManus et al. (2021)—30%--was used. The smallest effect of interest (SESOI) was a 10% proportion increase, which an $N = 590$ per condition yields 95% power for, assuming a two-tailed test with an alpha level = 0.05. Importantly, this sample size yields more than 95% power to detect group-level simple effects effect sizes consistent with Studies 1-2 ($dz = 0.15 – 0.25$), assuming two-tailed tests with an alpha level = 0.05. Because of the nature of the predicted crossover interaction, this sample size yields even higher power (than the predicted simple effects) to detect it.

Study 3.3 Results

Group-Level Replications. Across all experiments, within each between-subjects condition (i.e., eight different conditions), the 2x2 within-subjects crossover interaction pattern replicated (No Calibration Trials: $F(1,657) = 79.21$, $p < .001$, $\eta_p^2 = 0.11$; Calibration Trials: $F(1,588) = 76.10$, $p < .001$, $\eta_p^2 = 0.11$); (Sequential Judgments: $F(1,627) = 83.92$, $p < .001$, $\eta_p^2 = 0.12$; Simultaneous Judgments: $F(1,608) = 116.40$, $p < .001$, $\eta_p^2 = 0.16$); (Different Stimuli: $F(1,637) = 97.86$, $p < .001$, $\eta_p^2 = 0.13$; Matched Stimuli: $F(1,640) = 50.04$, $p < .001$, $\eta_p^2 = 0.07$); (Impossible to Opt Out: $F(1,745) = 109.10$, $p < .001$, $\eta_p^2 = 0.13$; Possible To Opt Out: $F(1,690) = 105.30$, $p < .001$, $\eta_p^2 =$

0.13). Non-parametric robustness checks (i.e., Wilcoxon signed-rank tests) also favor this interaction across experiments.

Across all experiments, within each between-subjects condition, the "No Choice" simple effect—in which agents who helped strangers are judged as more morally good than agents who helped family members—replicated, ts > 2.39, ps < .020, d_zs > 0.09, d_{av}s > 0.09. Non-parametric robustness checks (i.e., Wilcoxon signed-rank tests) also favored this simple effect across experiments. Similarly, the "Choice" simple effect—in which agents who helped strangers *instead of* family members are judged as less morally good than agents who helped family members instead of strangers—replicated, ts > 4.27, ps < .001, d_zs > 0.17, d_{av}s > 0.13. Non-parametric analyses favored this simple effect across five of six conditions. In the "Matched Stimuli" experiment, a Wilcoxon signed-rank test suggested that this simple effect was not robust for participants in the experimental condition. It is unclear what caused this null effect; it may be stimulus-specific, or simply a sampling distribution outlier.

Methodology-Based Proportion Changes. Across all experiments, the proportion of participants showing the group-level effect never substantially increased. That is, not a single methodological manipulation significantly improved the discrepancy between the person-level and group-level interaction patterns. Two methods significantly *worsened* the discrepancy; specifically, making simultaneous judgments and responding to matched stimuli, relative to making sequential judgments and responding to different stimuli, led to lower proportions of participants showing the interaction pattern (see Table 5 for detailed statistics and Figures 9a-d for person-level frequencies of all possible patterns across experiments).

Table 5. Underlying Logic and Results for Study 3.3.

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

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

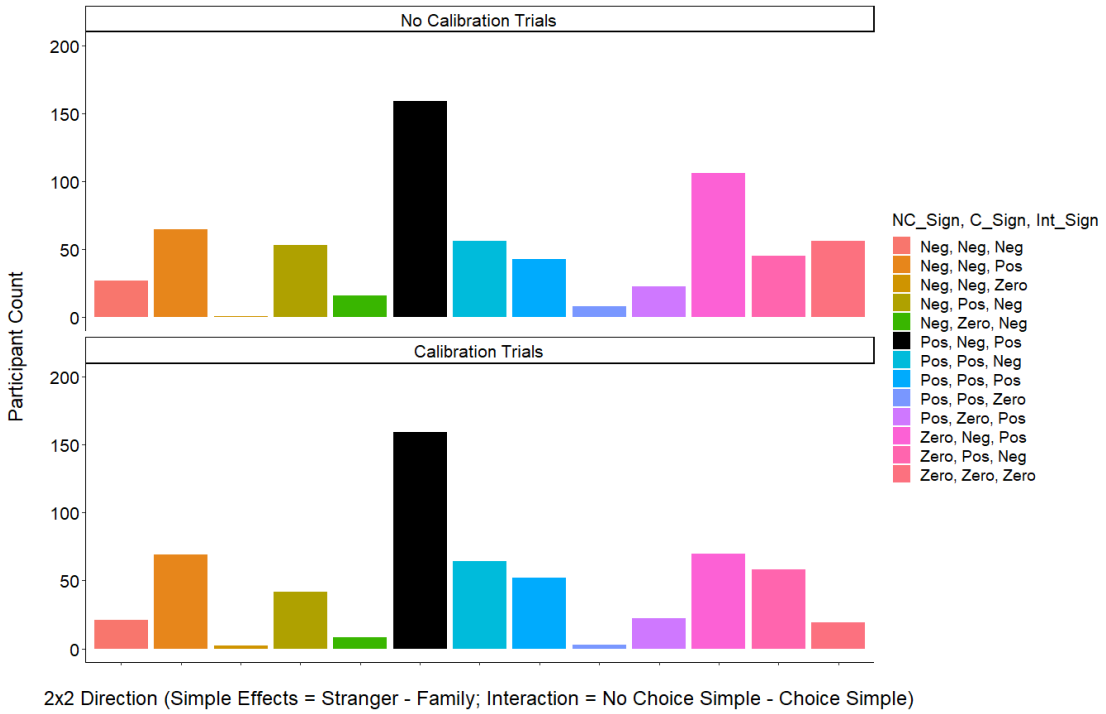

Figure 9a. Person-level patterns from Calibration Trials experiment. The black bar represents the group-level effect.

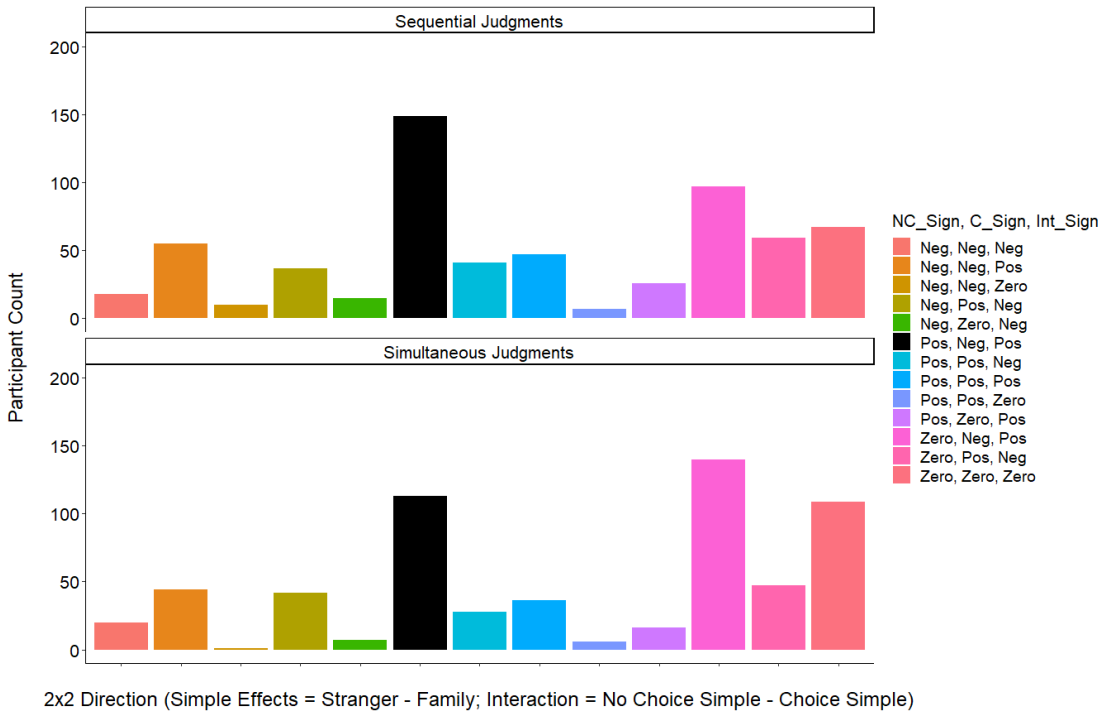

Figure 9b. Person-level patterns from Simultaneous Judgments experiment. The black bar represents the group-level effect.

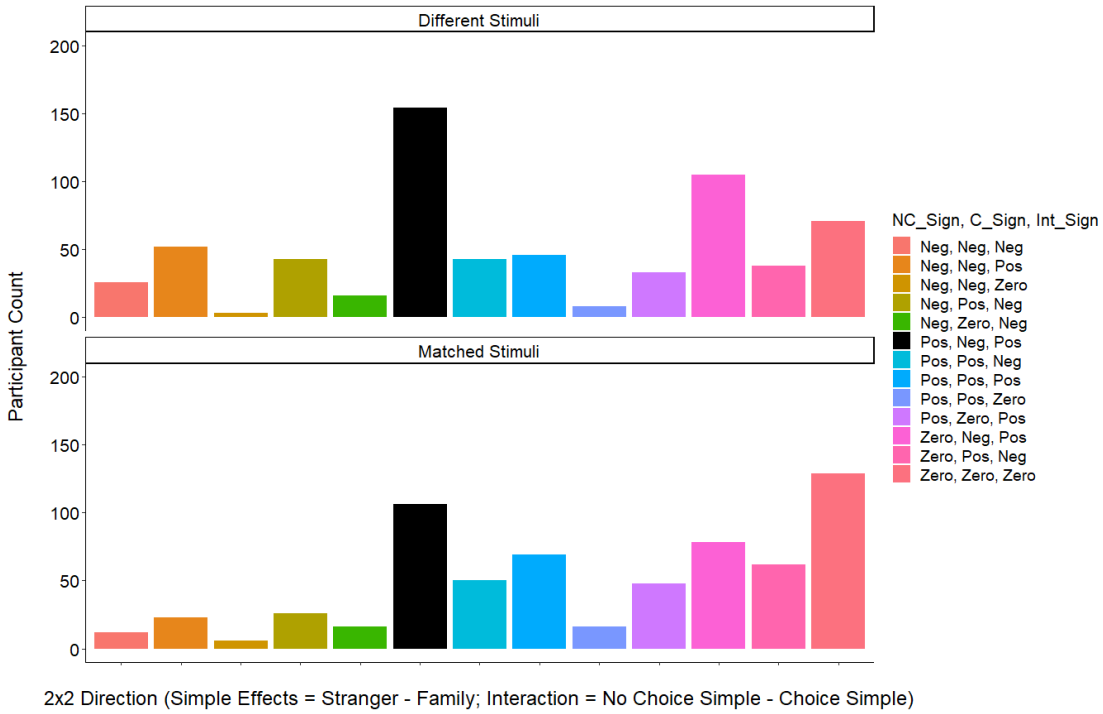

Figure 9c. Person-level patterns from Matched Stimuli experiment. The black bar represents the group-level effect.

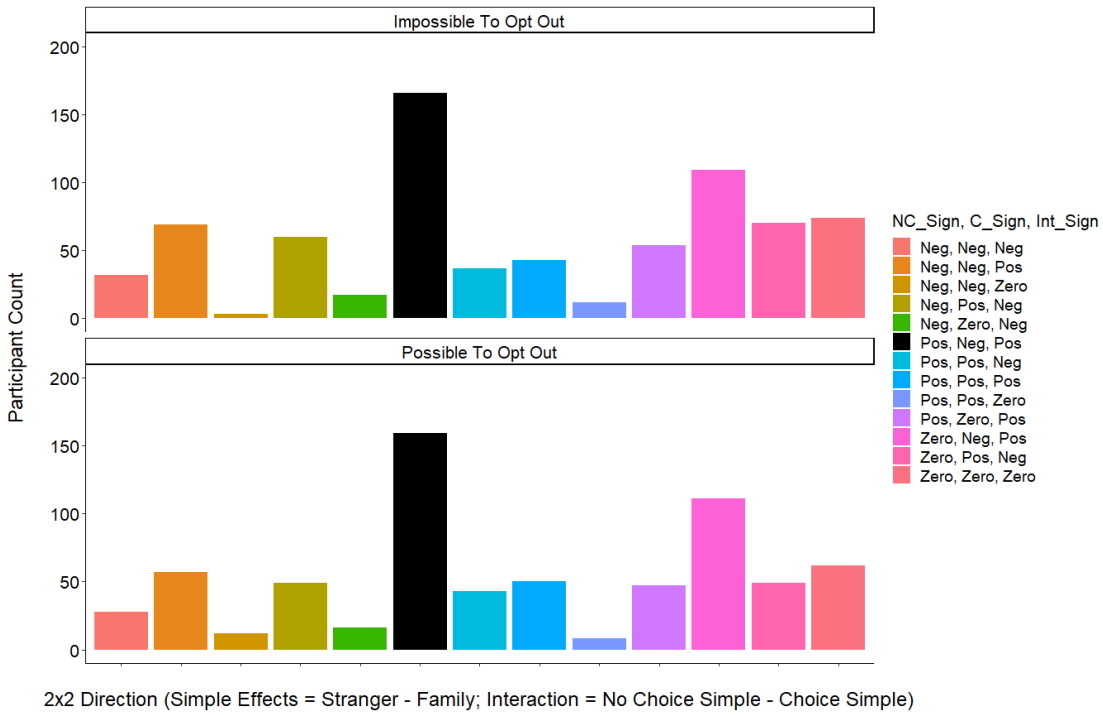

Figure 9d. Person-level patterns from Calibration Trials experiment. The black bar represents the group-level effect.

Study 3.3 Discussion

Study 3.3 Discussion

Although the focal group-level effects from Studies 1-2 were consistently replicated, the associated low person-level proportions were also replicated. Additionally, each methodological manipulation failed to increase the proportion of participants whose responses matched the group-level pattern. Therefore, the methodology-based proportion change hypotheses are considered falsified. On one hand, this is bad news; simple methodological manipulations, at least the ones employed here, may not be a remedy for the discrepancy between person-level and group-level patterns. On the other hand, these results suggest that, at least in the context studied here, there may be meaningful individual differences in moral reasoning. This means that psychologists interested in how social relationships affect moral reasoning across contexts have more research to do. Perhaps there are extant individual difference measures that explain the variability in person-level responses in this paradigm. To the extent that this issue generalizes to other subdomains of moral psychology (and perhaps to experimental psychology generally), the results of Study 3.3 suggest a need to incorporate person-level analyses.

Thus far, it has been argued that there is a group-to-person generalizability problem in some moral cognition research (and its resistance to obvious method-based remedies has been documented). However, there is clear subjectivity involved when deciding what should count as person-level evidence for a claim. For example, many claims which were categorized as instances of the problem (see Table 4 from Study 3.2) may seem unproblematic to most researchers. It could be argued that percentages of participants in the 20-40% range, who show the group-level patterns, are quite high. Perhaps readers of psychology research (e.g., social psychology researchers) do not

interpret authors as intending to make claims that represent at least a simple majority of participants. Moreover, psychology researchers may not believe that a claim ought to describe at least a simple majority of participants in order for the data to provide evidence for a psychological theory. The purpose of Aim 3.4 is to examine these possibilities.

AIM 3.4: Determine how psychology researchers interpret claims derived from typical group-level analyses.

Hypothesis 3.4a.

Most social psychology researchers will infer that a claim based on group-level patterns is representative of a majority of the study's participants. This finding would suggest that, even though inferences about person-level proportions cannot be made from group-level statistical tests alone, these inferences indeed are made.

Hypothesis 3.4b.

Most social psychology researchers believe that, for a claim to support a general, individual-level psychological theory, the group-level patterns must be representative of the majority of the study's participants. This finding would suggest that theoretical claims in psychology are held to some implicit standard which is rarely (if ever) tested. Moreover, this finding would suggest that psychologists ought not use traditional group-level approaches (e.g., ANOVAs and t-tests) to answer psychological questions, as these approaches do not assess the pervasiveness of a psychological experience.

Study 3.4 Methods

Participants. Psychology researcher participants were affiliated with the Society for Personality and Social Psychology (SPSP), recruited via SPSP's Open Forum listserv and compensated with Amazon gift cards. The sample size goal for Study 3.4 was $N = 280$. Due to the data collection format (asking for survey responses on an academic listserv), the sample size goal was not fully reached. After one attempt to get more responses (via reposting to SPSP's Open Forum listserv), the study was closed once incoming responses completely stalled, which occurred after two weeks. Applying the exclusion criterion

(failing a comprehension check) results in $N = 244$. Resampling was not attempted due to still having high statistical power for the focal hypothesis tests (see *Statistical Power*).

Design and Procedure. Participants were randomly assigned to one of two conditions. Half of participants learned about a simple effect comparison, whereas the other half of participants learned about a more complex, two-way interaction effect. Both simple and complex effects were used to test the generality of the hypotheses.

At the beginning of the study, all participants were informed that they would be answering questions about a moral cognition experiment. For the simple effect condition, participants learned about an effect from the supplemental materials of Law, Campbell, & Gaesser (2021). For the complex effect condition, participants learned about the interaction effect from Studies 1-2 (McManus et al., 2020; 2021). Participants first read text communicating results in typical journal article format (with means, SDs, t-values, p-values, within-subject standardized effect sizes for comparisons of interest [d_z], and a barplot). After learning the results, they then read text that simulated how data-based claims are made in general discussion sections (e.g., "People judged fictional agents who helped a stranger as more morally good than fictional agents who helped a cousin, but they judged fictional agents who helped a stranger instead of a cousin as less morally good than fictional agents who helped a cousin instead of a stranger").

After learning about the claim, participants were asked "By *people,* approximately what percentage of the study's participants do you think the researchers mean?" This measure is called the "empirical proportion estimate." Next, participants learned about a (fictional) general, individual-level theory that the authors had developed pre-experiment. Participants were again shown the claim in general discussion format and told that, later

in the paper, the authors used their study's results to claim support for their theory. Participants were then asked, "In order for the study's results to support the researchers' theory/model, approximately what percentage of the study's participants do you think need to respond in the way described by [the general discussion's language]?" This measure is called the "theoretical proportion estimate." Finally, participants could write an open-ended response to communicate anything that they were unable to communicate throughout the study. After the main task, participants answered several demographic questions, a few of which provided insight on research experience in academia.

Measures. As described above, participants provided two focal responses during the study. The first response was to a question which followed the discussion section formatted description of the moral cognition experiment's results, "By *people,* approximately what percentage of the study's participants do you think the researchers mean?" Responses ranged from 0-100% on a sliding scale, with the starting position (0, 50, 100) counterbalanced across participants. This measure allows categorization of responses into two categories: less than a simple majority (50% or less), and equal to or greater than a simple majority (51% or more). The second response was to a question which followed information indicating that the authors of the experiment used their results to claim support for a general, individual-level moral psychological theory, "In order for the study's results to support the researchers' theory/model, approximately what percentage of the study's participants do you think need to respond in the way described by [the general discussion's language]?" This measure is called the "theoretical proportion estimate." Responses were measured identically to the empirical proportion estimate.

Analytic Procedure. To investigate whether claimed effects in these studies represented most participants, the same person-level grouping approaches from *Aims 3.1-3.2* were utilized. To test whether the majority of participants (i.e., social psychology researchers) indeed showed the pattern which matched *Hypotheses 3.4a-b*, binomial tests will be conducted on the empirical and theoretical proportion estimates. Specifically, the estimates will first be categorized as hypothesis-consistent versus hypothesis-inconsistent. Second, the frequency of hypothesis-consistent estimates will be calculated for each estimate type. Finally, for each estimate type, a binomial test will be conducted to test the proportion of hypothesis-consistent responses against a proportion of 50%.

Statistical Power. The sample size goal for each Study 3.4 was 140 participants per between-subjects condition (simple versus complex effects). This sample size yields high-powered binomial tests, the focal test to determine whether the proportion of hypothesis-consistent responses exceeds 50%. To reason through a power analysis, pilot data from a laypeople sample was used. Specifically, this same study was conducted first on laypeople, finding that the smallest deflection from 50% was 31% (i.e., 81%). That is, more than 80% of laypeople's responses supported *Hypotheses 3.4a-b*.

It was reasoned that social psychology researchers would be less likely to assume group-level effects generalize to a majority of sampled participants. It was further reasoned that social psychology researchers may be less likely to believe that an effect ought to describe a simple majority of sampled participants for the effect to provide support for a general, individual-level psychological theory. The smallest effect size of interest (SESOI) was therefore chosen to be a 15% deflection from 50% (i.e., 65%), which an $N = 140$ per condition yields 95% power for, assuming a two-tailed test with an

alpha level = 0.05. As noted in *Participants*, due to sampling difficulties, the total final analyzable $N = 244$ ($N_{Simple} = 123$ and $N_{Complex} = 121$). Resampling was not carried out because sensitivity analyses revealed that each sample size still yielded more than 90% power to detect the SESOI.

Study 3.4 Results

Empirical Proportion Estimate. The majority of social psychology researchers believed that authors intended to describe at least a simple majority of their study's participants, for both simple effects (73% [64% - 81%]), $p < .001$, and complex effects (80% [72% - 87%]), $p < .001$. Strikingly, as shown in Figure 10, there is no discernible pattern as a function of being relatively inexperienced (e.g. being an undergraduate) and relatively experienced with academic research (e.g., being a professor).

Theoretical Proportion Estimate. The majority of social psychology researchers also 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 general, individual-level psychological theory, for both simple effects (80% [72% - 87%]), $p < .001$, and complex effects (90% [83% - 95%]), $p < .001$. As shown in Figure 10, there again is no discernible pattern as a function of research experience.

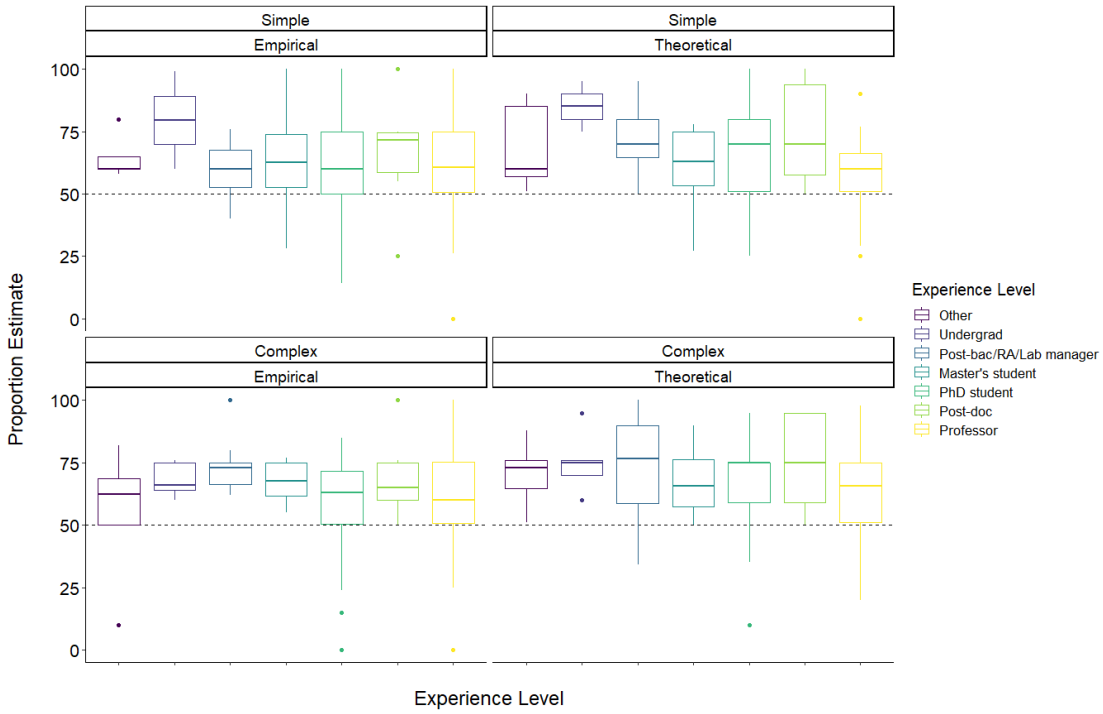

Figure 10. Boxplots of empirical/theoretical proportion estimates by effect type (simple versus complex), and by participants' level of experience. Note that "Other" (n = 13) refers to people involved in academic research in some way (via SPSP) but who indicated that they have never held an academic position.

Study 3.4 Discussion

Recent meta-scientific critiques of psychology have pointed out that there could be (and is) a pervasive mismatch between psychological theorizing and the analytic procedures used for testing (e.g., Richters, 2021; Speelman & McGann, 2020). Specifically, typical theorizing occurs at the person-level but psychology's analytic procedures typically operate at the group-level. This may have serious consequences on how psychologists make inferences generally, especially when attempting to corroborate or revise theory. Currently, psychologists may believe that their data corroborates a general theory and then claim it does. However, if their claim about a psychological experience is one that does not represent most of their participants, then their beliefs about their data would be wrong.

The results of Study 3.4 suggest that psychology researchers typically interpret estimates from group-level statistical tests to represent the majority of sampled participants. Moreover, researchers typically believe that for a psychological claim (e.g., People judged X as less morally acceptable than Y) to provide support for a general, individual-level theory, the claim must represent *at least* a simple majority of sampled participants. In fact, the majority of researchers believed a theory-supporting claim can be a theory-supporting claim only if it represents more than 60% of sampled participants. These findings are problematic when considering the way in which analyses are typically conducted and reported, as well as how verbal claims are derived from these analyses. First, if most psychology researchers interpret results of group-level tests as representing most sampled participants, it is unknown how often this interpretation is incorrect, as person-level statistics are rarely (if ever) reported in psychology journal articles. Second, if a criterion for a claim's theory-based validity is that the claim represents most sampled participants, then there are multitudes of psychological claims in the published literature which have not yet been properly tested, as aggregation or nomothetic approaches (e.g., averaging across different participants' responses) are ubiquitous in experimental psychology. Therefore, if psychologists' claims indeed ought to rely on the empirical pervasiveness of the phenomenon of interest (as the results of Study 3.4 suggest), then experimental psychologists need to adopt different analytic strategies from those typically used.

General Discussion

The current work accomplished three broad aims. First, it was established that, when evaluating others' prosocial behavior, people *on average* judged agents who helped strangers as more morally good than those who helped family members, but agents who helped a stranger instead of a family member were judged as less morally good than agents who helped a family member instead of a stranger. However, when agents were in roles that required impartiality (e.g., teacher/doctor), agents who helped a stranger instead of a family member were judged as more morally good than agents who helped a family member instead of a stranger. Moreover, all these patterns were underlied by differential perceptions of obligations, with agents helping strangers judged as fulfilling an obligation less so than agents helping family members (except for the impartiality scenarios, where this difference was, as expected, reversed). These results suggest that people flexibly incorporate important social information into their moral judgments, hinging their evaluations on inferred obligations that are absent or present across scenarios.

Second, though, the above-mentioned results were inferred via potentially problematic experimental methods. Specifically, it was claimed in the first aim that perceived obligations underlied the discovered moral judgment patterns, but the design of the first aim's experiments made it difficult to infer whether obligation judgments were being retroactively contaminated when considered in conjunction with moral judgments. In the second aim, rather than making moral judgments and obligation judgments simultaneously only *after* the outcome of each scenario was known, people instead made obligation judgments *before,* and moral judgments *after,* the outcome of each scenario was known. Additionally, this necessitated a change in how obligation judgments were

measured. Rather than asking people to judge how much an agent violated or fulfilled an obligation, the second set of experiments asked people to judge the presence and strength of an obligation. This second set of experiments replicated the results of the first set of experiments, which ruled out the possibility that moral judgments were retroactively contaminating obligation judgments in the earlier research. Specifically, people, *on average,* judged agents as having a weaker obligation to help strangers than to help family members, and having weaker obligations to help cousins than siblings. Again, people, *on average,* judged agents as more morally good for helping strangers compared to family members, but they judged agents as less morally good for helping strangers instead of family compared to helping family members instead of strangers. However, obligation judgment differences were only correlated with moral judgment differences in "Choice" conditions (i.e., when agents had to make a choice about whom to help), but not in "No Choice" conditions (i.e., when agents helped the only person in need of help). These results suggest that obligation judgments and moral judgments may be systematically aligned in "Choice" conditions, but only incidentally aligned in "No Choice" conditions.

Third, however, it was realized that there were shortcomings of many of the statistical approaches used throughout the first two aims. Specifically, all main analyses were unable to answer a question about whether *the same people showed each set of hypothesized effects*. For example, difference score correlations assess the relation between obligation differences and moral judgments differences, but only within one condition at a time. Additionally, simple effects tests (e.g., t-tests) assess judgment differences within one level of a factor at a time. Therefore, these tests could not allow an

inference about the pervasiveness or prevalence of the complex cognition that was hypothesized. When investigating the prevalence of the moral judgment effects from the previous two aims, it was never the case that a majority of people's responses matched the nuanced psychology that was originally inferred. This problem was also found in others' recent, similar, moral psychology research. Moreover, several methods experiments, which were designed to reduce noise and therefore increase the number of people showing the predicted group-level patterns, failed to do so. Finally, the potential gravity of the problem was further investigated by documenting how social psychology researchers themselves tend to interpret group-level statistical tests. Here, most researchers interpret (sets of) group-level statistical tests as representing the majority of sampled participants, and most researchers believe this ought to be true if the claim derived from such analyses is being used to support a general, person-level, psychological theory.

In the following sections, implications of this research are discussed. First, implications will be discussed with respect to the moral psychology literature specifically. It is noted that this first section is relatively shorter than the second section, as important discussion points were built into each Aim's discussion section earlier on. Moreover, much of this first section assumes that the group-level effects from Aims 1-2 are adequate representations of person-level psychology, though it is clearly noted when this assumption is not being made, and when other (untested) method-based features of these studies could have been altered to result in the hypothesized effects being representative of most persons. Second, implications will be discussed with respect to psychological science broadly. It is noted that this second section is much longer. In this

second section, it is assumed that the group-level effects from Aims 1-2 (as well as those from Aim 3) are *in*adequate representations of person-level psychology.

Implications for the Moral Psychology Literature

Ten years ago, Bloom (2011) identified a gap in our understanding and investigation of moral psychology:

> The problem is that most research in this field, including my own, focuses almost entirely on how people make sense of, judge, and respond to interactions of unrelated strangers. We have little to say about how people think of interactions that occur between parent and child, brother and sister, and closely related individuals. We also often ignore moral judgments and moral feelings that concern spouses, close friends, colleagues, allies, and compatriots.

Bloom goes on to argue that these are precisely the interactions that matter most, and, in turn, that the field's failure to explore them leads to the development of theories that do not capture our everyday moral psychology.

Similarly, much recent conceptual and empirical work has pointed out that, either implicitly or explicitly, the field of moral psychology has often operated as if social relationships are not important for third-person moral judgment (e.g., Earp et al., 2021*;* Everett et al., 2018; Gray & Hester, 2020; Law et al., 2021*;* Marshall et al., 2020; Marshall et al., 2022*;* Schein, 2020). Indeed, many prominent theories do not make predictions about how social relationships will impact moral judgment (e.g., Gray, Young, & Waytz; 2012; Schein & Gray, 2016; 2018; Young et al., 2007; 2010). Some prominent theories, however, stress the importance of social relationships to different degrees (e.g., Graham et al., 2011; Rai & Fiske, 2011). Specifically, Relationship Regulation Theory (Rai & Fiske, 2011), which partly draws on relationship psychology (e.g., Clark & Mills, 1979), suggests that communal sharing relationships, like those with family, are characterized by providing preferential aid to those within the group.

Therefore, RRT predicts that failure to uphold this relationship obligation will be judged negatively. Similarly, Morality-as-Cooperation (Curry, 2016) predicts that helping family is considered a universal moral good. The current findings support these predictions, but add greater nuance.

Assuming Group-Level Effects are Representative of Person-Level Moral Psychology

Contributing to the recent resurgence in research on everyday moral psychology, findings from Aims 1-2 suggest that people view relatedness between potential helpers and beneficiaries to be a key determinant in perceived strength of obligations to help. Agents were judged as having stronger obligations to help close genetic relatives (i.e., siblings) than more distant genetic relatives (i.e., cousins) and non-relatives (i.e., strangers), and agents were also judged as having stronger obligations to help distant genetic relatives than non-relatives. Replicating prior research (Marshall et al., 2020; McManus et al., 2020), perceived obligations appeared to inform moral judgments. 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. These data further suggest, however, that in contexts where agents do not have to consider a choice about whom to help, differences in obligation strength are unassociated with differences in moral character. That is, there was not an association whereby the more people discriminated between distantly related (or unrelated) others and more closely related others in their obligation judgments, the more they discriminated (in the opposite direction) in their moral character judgments. This non-association also held when using repeated-measures

correlations which analyze within-person variability as opposed to between-person variability (Bakdash & Marusich, 2017).

As noted earlier, these results suggest that the link between differential obligation judgments and differential moral evaluations may be more complicated than previously assumed (e.g., Marshall et al., 2020; Marshall et al., 2022; McManus et al., 2020). Even when there were average differences in both obligation judgments and moral evaluations, these relationships were still non-existent, suggesting that perhaps attribution theory (Kelley, 1967) is the wrong lens through which to frame these effects. Therefore, these data can be interpreted in the following way: Most people who made differential obligation judgments did not also make differential moral evaluations, or, most people who made differential obligation judgments did not tend to agree on how to differentiate in their moral evaluations. However, it is also possible that this non-association was an artefact of the experimental design, as judgments in this context were made across two entirely different scenarios rather than within a single scenario. Regardless, these findings highlight the need for more work exploring the underlying mechanisms for differences in moral evaluations in this context. Further, the implications of these results echo recent calls to avoid using sets of aggregate estimates to draw conclusions about person-level psychology (e.g., Fisher, Medaglia, & Jeronimus, 2018; Grice et al., 2020).

On the other hand, in contexts where agents had to consider a choice about whom to help, differences in obligation strength were consistently positively associated with differences in moral character. Specifically, the more people discriminated between distantly related (or unrelated) others and closely related others in their obligation judgments (with stronger obligations to help closer others), the more they discriminated

in their moral character judgments (with more positive moral character judgments for agents who helped closer others). This association also held when using repeated-measures correlations. Together, results from Aims 1-2 suggest that obligations may be especially salient in contexts where choices about whom to help can or must occur, and in turn, this is when perceived obligation strength will be especially likely to structure subsequent moral evaluations.

Overall, if the discipline of moral psychology has the goal of understanding day-to-day moral psychology, results from Aims 1-2 suggest the need to continue to incorporate real-world social information into its stimuli. Without doing, so moral psychologists may continue down the path that Bloom (2011) admonished, developing theories that only explain abstract, contextless, moral judgments (Hester & Gray, 2020).

Assuming Group-Level Effects are Unrepresentative of Person-Level Moral Psychology

As was shown in Study 3.1., it was never the case that the hypothesized set of moral judgments represented most participants. Most of the time, this was driven by a minority of participants not making the predicted distinction between strangers and family in the No Choice context. Therefore, the following recommendations are made for scholars who wish to build upon the research from Aims 1-2, investigating the effects of relationship obligations on moral judgment.

In general, it is recommended that researchers shift their focus from single- or few-trial studies to high-trial studies. First, researchers should create many pairs of helping stimuli that differ only in relationship information (i.e., controlling for other features of stimuli that could covary with relationship information). Second, in addition to measuring moral judgments and obligation judgments, researchers should use

additional measures that tap other potential mechanisms (e.g., expectation of help, typicality of help, etc.; see the SOM of McManus et al., 2020 for candidate mechanisms). Third, researchers should use the aforementioned pairs of stimuli and measures, together, to create a procedure which allows participants to make relative judgments of agents helping strangers versus agents helping family across a high number of trials (see "Implications for the Practice of Psychological Science Broadly," for more details). Fourth, for moral judgments and obligation judgments, researchers should use a method called "high-trials Bayesian prevalence estimation," a method which runs typical group-level frequentist tests *within* each person's high-trial data (again see "Implications for the Practice of Psychological Science Broadly," for more details). Finally, to rule out non-obligation mechanisms as the primary explainer of moral judgments, researchers should conduct person-level mediation analyses that control for the other candidate mechanisms. The suggested method of "high-trials Bayesian prevalence estimation" can also be applied to this final step, allowing an inference about how many people in the population are likely to show a particular set of moral judgments as a function of obligation judgments (again see "Implications for the Practice of Psychological Science Broadly," for more details).

Following the above steps would provide strong evidence that people's relationship obligation judgments structure their moral judgments in opposite directions (i.e., as predicted in No Choice versus Choice contexts). While it may not seem necessary to apply these methods across No Choice and Choice contexts (as the Choice context effects were much more consistent across Aims 1-2, at both the group- and person-level), it is noted that even the Choice context effects were inferred from only one to two trials

per condition. As is known in the multilevel modelling literature (e.g., Yarkoni, 2020), this is not enough data to test the generality of an effect across stimuli. Overall, although these suggestions will require a much higher bar for positive evidence than moral psychologists typically attempt to meet, their implementation (and success) would provide person-level evidence of not just any set of (complex) moral judgment effects, but also of the theorized underlying mechanisms.

Implications for the Practice of Psychological Science Broadly

Psychology is often defined as "the study of the mind and behavior." Therefore, its essential goals are describing cognitive functions and uncovering their antecedents and consequences. This work contends (and provides positive evidence supporting the idea) that researchers intend to apply these goals to the study of individual persons, as it is minds that possess psychological processes, and each mind resides in a single person. Moreover, this work contends (and provides positive evidence supporting the idea) that many psychologists intend to (and interpret others as intending to) uncover phenomena that describe a majority of persons (i.e., general psychological laws; Hamaker, 2012). However, as demonstrated in the current work, even though it can lead to serious errors in inference, it is typical to conduct and report *only* group-level analyses and interpret them *as if* they support or falsify person-level hypotheses. In the following sections, it is argued that only certain kinds of experimental designs permit tests of person-level hypotheses, additional evidence is offered to dissuade researchers from continuing to use group-level analyses in the typical way to answer psychological questions, recommendations are provided for the future, and finally, objections and limitations are discussed.

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

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

To illustrate this possibility in a different domain, Birnbaum (1999) had participants judge the largeness of numbers on a 10-point scale ranging from *very very*

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

Clarifying the Problem

This does not mean that between-subjects designs are never useful. These designs may be preferable when within-subjects designs are practically infeasible or impossible.

For example, many intervention(-like) research questions may be best answered with between-subjects designs (e.g., see Study 3.3). Additionally, hypotheses about population(-like) differences require at least one between-subjects factor, such as testing whether psychopaths show different experimental effects than non-psychopaths. Finally, between-subjects designs are unproblematic when the research goal is to provide generalization evidence (e.g., finding similar effects across instructions/measures; see Yarkoni, 2020).

However, between-subjects designs cannot conclusively provide person-level evidence of an experimental effect, just as group-level correlations among variables cannot provide evidence of person-level correlations among those variables (see Fisher et al., 2018). For example, in the current moral cognition research, moral character judgments were assessed to test their sensitivity to social relationship information in the context of helping behavior. Among other variations, participants were given two scenarios: one in which someone helps a total stranger, and another in which someone helps a distant family member. Group-level analyses suggested that participants–*on average*– judged agents who helped strangers as more morally good than agents who helped family members, presumably because people believe that there is no obligation to help strangers. Importantly, this was tested using a within-subjects design. Therefore, this design permitted investigation of the question, "How many people's responses match the pattern indicated by the difference between conditions?" A between-subjects design would have disallowed such investigation.

Importantly, as was documented across Aim 3's studies, using within-subjects designs does not automatically prevent inference errors from occurring. Researchers can

still commit ecological or ergodic fallacies (Kuppens & Pollet, 2014; Speelman & McGann, 2020), due to special instances of Simpson's paradox—when group-level patterns poorly represent lower-level units constituting the group (Simpson, 1951; Kievit, Frankenhuis, Waldorp, & Borsboom, 2013; also see Hamaker, 2012, for an illustrative example on the relation between typing speed and mistake frequency). To reiterate, even when psychologists deploy appropriate experimental designs, they often, if not always, only report their group-level analyses.

As a result of the current work, it is suggested that, if a research question or theory is a person-level one, and the goal of a study is to make a general claim (Hamaker, 2012), then researchers ought to choose appropriate designs and analytical procedures that allow themselves (and readers) to answer the question, "How many people's responses match the pattern(s) indicated by the mean difference(s) between conditions?"

How Poorly Can Group-Level Patterns Represent Persons?

If the evidence and arguments presented thus far are not convincing in motivating investigation of person-level responses, one ought to consider just how poorly group-level experimental effects can represent persons. To investigate this, hypothetical datasets were generated in which sets of group-level patterns fail to describe the psychological experience of any single participant. In these (N = 100) datasets, the following group-level patterns were created: 2x2 crossover interaction, 2x2 attenuation interaction, and a three-level ordinal effect. All of these patterns emerged at the group-level (and survived non-parametric tests) but with none of the participants' scores showing *all* of the relevant effects.

For example, in the attenuation interaction dataset, there are two, two-level factors (A vs B, 1 vs 2), where there are three group-level effects: condition A2 > A1, B2 > B1, and the interaction (i.e., a difference-in-differences in which the simple effect of 2 vs 1 is larger in condition A than condition B). The interaction between A/B and 2/1 is significant, $F(1,99) = 9.19$, $p = .003$. This interaction survives a Wilcoxon signed-rank test ($A_{Diff} > B_{Diff}$: $Mdn_{Diff} = 1.50$ [0.50, 2.50], $V = 2654$, $p < .001$, $r = 0.26$). Results of paired-samples t-tests revealed that A2 was significantly higher than A1, $M_{Diff} = 2.44$ [1.86, 3.02], $SD_{Diff} = 2.39$, $t(99) = 8.33$, $p < .001$, and B2 was significantly higher than B1, $M_{Diff} = 1.16$ [0.77, 1.55], $SD_{Diff} = 1.98$, $t(99) = 5.85$, $p < .001$. Moreover, these simple effects survive non-parametric Wilcoxon signed-rank tests (A2 > A1: $Mdn_{Diff} = 3.00$ [2.50, 4.00], $V = 226$, $p < .001$, $r = 0.63$; B2 > B1: $Mdn_{Diff} = 1.50$ [1.00, 2.50], $V = 428$, $p < .001$, $r = 0.48$).

However, in these data, it would be erroneous to claim that people respond (or even that a single person responds) in a way that reflects same-direction but different-in-magnitude responses. These data were created to show how a group-level attenuation effect could emerge which suggests a process that is less sensitive to one type of stimulus compared to another, even though no single participant shows this pattern (see Figure 11's "Pos, Pos, Pos" pattern). Although this is just an existence proof, it should be taken as a cautionary tale of relying on *only* group-level patterns when making inferences about psychological process.

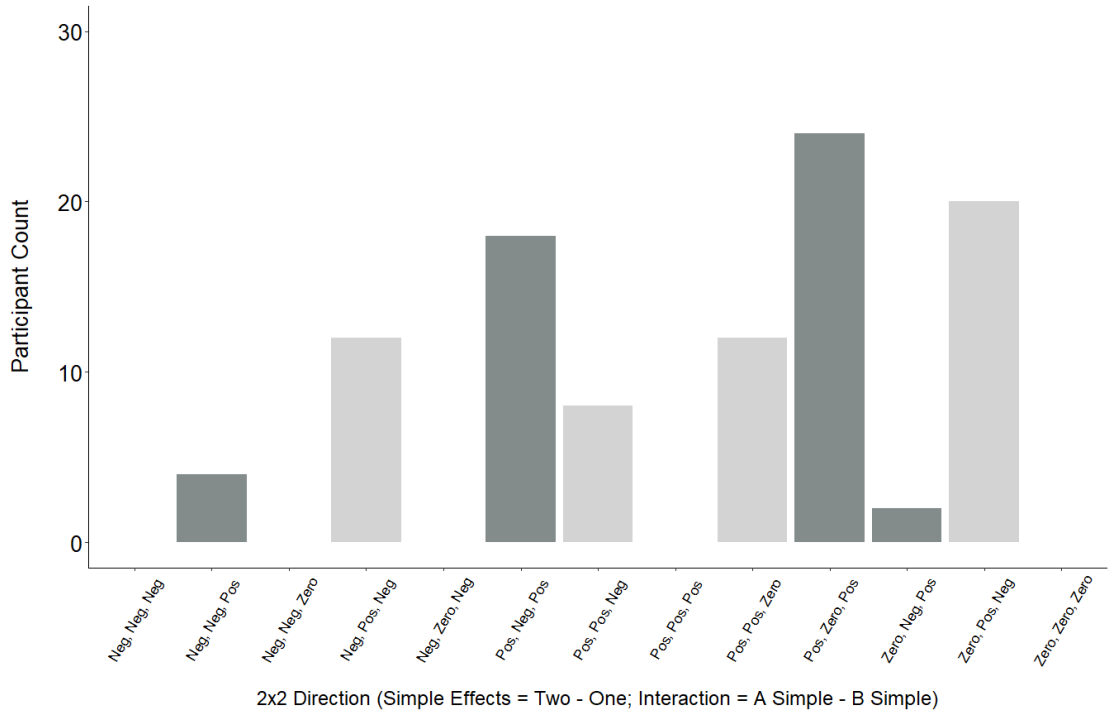

Figure 11. Person-level patterns for the attenuation interaction dataset. Pattern descriptions (e.g., Pos, Neg, Pos) communicate the A difference, B difference, and Interaction difference, respectively. The (non-existent) black bar represents the claimed group-level patterns (i.e., "Pos, Pos, Pos," which describes zero participants here). Dark grey bars represent patterns which also yielded an interaction value that contributed to the group-level interaction pattern's emergence.

Recommendations for Confronting the Group-to-Person Generalizability Problem

Given the group-to-person generalizability problem, what should experimental psychologists do? In this section, four easy-to-implement analytic strategies are proposed to aid in making person-level prevalence claims (see Table 6 for pros and cons of each andFigure 12 for these recommendations as a simple decision flowchart).

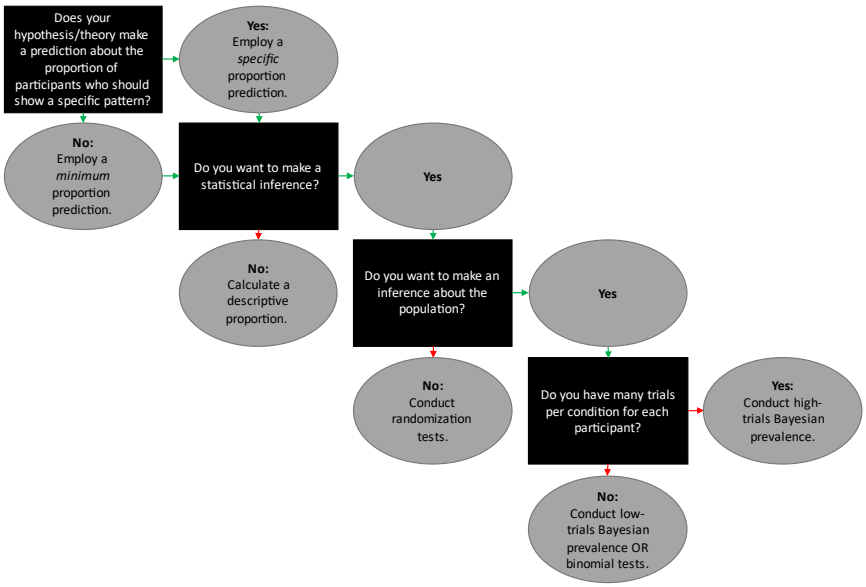

Figure 12. Decision flowchart for investigating proportions. Black boxes represent questions that researchers need to answer, whereas grey ovals represent possible decisions. Red arrows from black boxes to grey ovals indicate that there are no more decisions to be made, but green arrows indicate that there is at least another question and therefore decision to be made.

Recommendations here are consistent with those in a recent critique (Yarkoni, 2020). Specifically, it is suggested that predictions should reflect orderings of observations based on theory (e.g., A1 higher than A2 in B1, but A2 higher than A1 in B2), while specifying a proportion of participants whose responses should match predictions for the theory-derived hypothesis to survive. To conduct a "severe test" (Mayo, 2018) or corroborate a "risky prediction," (Meehl, 1990a, 1990b), the empirical proportion should be close to the theory-predicted proportion, and other theories should not predict this proportion.

Unfortunately, there may not be many psychological theories (especially in moral/social cognition which typically lack formal models) that can make such predictions (for examples and discussion of this issue see, Crockett, 2016; Hamlin et al.,

2013). However, theoretical progress can still be made (see the first black box of Figure 12). For example, researchers can make and test minimum proportion predictions (e.g., 51+%) if the goal is to make a *general* claim. To be clear, 50% is not advocated as the benchmark against which psychologists should test theory. It is also not recommended to ignore theory-inconsistent patterns, or patterns represented by a small minority of participants. Understanding if and why other patterns exist allows refinement of theory by postulating and testing whether there are substantive moderating variables (e.g., individual differences), or simple violations of auxiliary assumptions (e.g., divergent interpretations of measures; see Quintana, 2021, for a discussion). Here, methods are being suggested that simply enable researchers to identify evidence that fails to support general psychological regularities (Hamaker, 2012). Elsewhere, some have suggested even higher proportions as convincing evidence (e.g., 80%; Speelman & McGann, 2020). Though responses from the current work suggest that psychologists disagree about the exact appropriate cutoff, most people believe researchers are intending to make claims that represent a majority of their studies' participants; and that to support/challenge theory, researchers' results should be describing a majority of participants.

To investigate proportion predictions, researchers can engage in various analytic strategies. First (see the second black box of Figure 12), it must be decided whether a statistical inference is desired. If not, researchers can simply calculate the sample's descriptive proportion and report it. If, however, researchers want to make a statistical inference, they must then decide whether an inference about the population is desired (see the third black box of Figure 12). If not, researchers can conduct randomization tests, which test whether a particular pattern emerges more than physical chance (Grice, 2021;

Grice et al., 2020). If, however, researchers want to make an inference about the population, they must then decide between three methods, a choice which will depend on participants' number of trials per condition (see the fourth black box of Figure 12). If there are very few (or only single) trials per condition for each participant, researchers can decide between using binomial tests and low-trials Bayesian prevalence estimation. If researchers want to use frequentist null hypothesis significance testing, they can conduct binomial tests, which test whether a particular pattern emerges more than physical chance but simultaneously allows for an inference from the sample to the population. However, this approach requires careful thought, as the physical chance null varies across design/measure combinations. If researchers want to avoid the complexity of simulating the physical chance null, they can instead conduct low-trials Bayesian prevalence estimation, which uses the sample proportion to estimate the most likely proportion in the population, given the sample data. Importantly, the binomial test (on its own) cannot achieve the same insight as Bayesian prevalence estimation, as the binomial null value that is being tested against must equal physical chance (Grice, *personal communication*).

When researchers have many trials per condition for each participant, high-trials Bayesian prevalence estimation is possible (see https://estimate.prevalence.online/ for an easy-to-use GUI). This is achieved by first conducting typical group-level tests within each person, and second by estimating (using results from the first step) the most likely proportion of people in the population who would show the predicted pattern. High-trials Bayesian prevalence estimation, which is perhaps the gold standard for making person-level prevalence inferences, is, as the name implies, only possible for high-repetition within-subjects designs. Therefore, for future research, there must be a sea change in how

psychological data are collected. Overall, it is hoped that the evidence marshalled, and the recommendations provided, gives researchers both the motivation and tools to examine group-to-person generalizability in their own areas of interest.

Table 6. Easy-to-implement analytic strategies to aid in making person-level prevalence claims

Analytic Method	Pros	Cons
High-Trials Bayesian Prevalence Estimation	• Tests whether qualitative differences between conditions are truly non-zero, assuming measurement error averages out within each person • Allows calculation of person-level standardized effects sizes and intervals • Allows prevalence inferences from samples to populations	• Requires as many observations within each person as typical group-level methods require across persons (holding expected effect sizes constant) • Cannot be applied to all prior (e.g., low-trial) studies
Low-Trials Bayesian Prevalence Estimation	• No requirement for total number of observations within persons • Allows prevalence inferences from samples to populations • Can be applied to all prior (even low-trial) studies	• Assumes qualitative differences between conditions are truly non-zero and error-free • Does not allow calculation of person-level standardized effect sizes and intervals
Binomial Tests (against the chance null)	• No requirement for total number of observations within persons • Allows NHST inferences from samples to populations • Can be applied to all prior (even low-trial) studies	• Assumes qualitative differences between conditions are truly non-zero and error-free • Figuring out the appropriate null value can be difficult and may require simulation (as the value will vary across design/measure combinations) • Does not allow calculation of person-level standardized effect sizes and intervals
Randomization Tests (against chance)	• No requirement for total number of observations within persons • Can be applied to all prior (even low-trial) studies	• Assumes qualitative differences between conditions are truly non-zero and error-free • Does not allow calculation of person-level standardized effect sizes and intervals • Does not allow prevalence inferences from samples to populations

Potential Objections, Limitations, and Future Directions

In the approach used throughout Aim 3 (i.e., person-level investigation of data), any one participant's responses were used to create a variable that indicated a qualitative directional (e.g., positive) difference between conditions, assuming that this feature was error-free. However, especially in cases when this variable was created from single scores in each condition, it is a fair objection that this qualitative difference cannot be assumed as error-free. The reported proportion estimates may be (extremely) higher or lower depending on how much measurement error played a role in single- and few-trial designs. This problem could be compounded in all three Aims' studies by the fact that many-pointed slider scales were often used to measure constructs of interest. Therefore, it is possible that many participants who were counted as "hypothesis-inconsistent" were indeed "hypothesis-consistent," but the many-pointed sliding measure made it possible to make very small, wrong-direction distinctions between conditions when a participant's intention was to indicate a small, correct-direction distinction. To combat these two problems in future research, one analytic approach and one design-based approach are recommended, perhaps together.

First, when possible, it is suggested to use an analytic approach called high-trials Bayesian prevalence estimation (see Ince, Kay, & Schyns, 2022, and Ince, Paton, Kay, & Schyns, 2021). The first step of this approach combats within-person measurement error in the same way that typical group-level approaches combat across-person measurement error. With large sample sizes, typical group-level approaches (e.g., t-tests) allow near-accurate estimation of population-level mean differences because measurement error is assumed to average out across persons. In the first step of high-trials Bayesian Prevalence

estimation, it is required to collect enough person-level data to conduct typical group-level tests *within* each person's data. Therefore, with a large enough trial set, a t-test, for example, can be conducted to compare response scores across conditions within each person; as the logic goes for across-person measurement error, here, measurement error should average out within each person's set of high-N trials. The next step of high-trials Bayesian Prevalence simply estimates the proportion of people whose person-level tests match predictions (see Table 6 for a list of other possible, easy-to-implement, analytic solutions). Second, because the scale-point issue remains as another source of error, a design-based approach is also recommended. Specifically, when feasible, researchers should present measures/stimuli in a way that requires relative responses (e.g., "Which face is angrier?" with scales ranging from *Face A is much angrier* to *Face B is much angrier*). This may allow researchers to have more confidence about whether any one trial's difference is a true difference and whether any one trial's non-difference is a true non-difference. The number of scale points here likely matters as well, with many-pointed (unmarked and/or sliding) measures likely increasing the number of true non-distinctions being recorded as small directional distinctions. This design-based approach should alleviate concerns about scale-based error, but more targeted research is necessary to fully support this possibility.

The suggested method of high-trials Bayesian prevalence estimation may make salient typical concerns about within-subjects designs (i.e.., order/and or demand effects). However, these concerns are not reason to jettison within-subjects designs. First, if participants engage in many trials per condition, researchers need not worry about stimulus order affecting the psychological comparison of interest. This is because

researchers will no longer need to make psychological inferences based on single judgments per condition (which themselves could be subject to context effects, Birnbaum, 1999). With high trials per condition, researchers can instead focus on the *in general* differences between conditions within each person. Additionally, if participants receive pairs of stimuli and make relative judgments, then order effects disappear completely. Second, the stimulus pairing with relative judgments suggestion may justifiably lead to concerns about demand effects (i.e., when participants learn the purpose of the study and attempt to respond in ways that they expect the researchers to desire). However, it is important to note that just because participants may infer the purpose of the study, this does not mean that they will attempt to respond in ways that researchers desire. If this is a worry, researchers can investigate this possibility empirically. After completing a study's main task, researchers can first ask participants if they attempted to, or incidentally, inferred the hypothesis of the study. Then, of the participants who did infer the hypothesis, researchers can ask if they responded in a manner that mirrored their inferred hypothesis. Finally, of the participants who responded this way, researchers can ask them to explain what the hypothesis was. To encourage honest reporting, researchers can reassure participants that responses to these questions will not influence their participation credit or payment. Researchers can then use these responses to make an informed decision about which participants should not be included in analyses (or to investigate the robustness of the analyses by conducting tests with and without these participants). As an example, this method was implemented in Aim 3's "Simultaneous Judgments" experiment. Of 1,283 participants (before attention check exclusions), only 185 participants (14%) indicated that they tried to infer the hypothesis,

and only 41 participants (3%) indicated that they tried to respond in line with the inferred hypothesis. Of these 41 participants, some were able to generate a version or portion of the broad research question, but none of them were able to articulate the full hypothesis. For these reasons, it was decided that no participants would be excluded from analyses. Even when additional concerns arise about within-subjects designs, the earlier arguments regarding the downsides of between-subjects designs should not be ignored. That is, between-subjects designs will often suffer from context effects, and more importantly, they fundamentally disallow investigation of person-level processes.

Another, unrelated objection is that there are other sources of noise accounting for the group-to-person generalizability problem, beyond those tested here. For example, some participants are distracted, leading to frequencies of person-level patterns which do not represent the "true" frequencies. First, consistent with Study 3.3's experimental results, there is no reason to believe, if such noise was reduced, that most person-level patterns would conveniently shift to the group-level pattern. Second, as explained in the discussion section of Study 3.1, and as evidenced in the additional hypothetical datasets, there are simple non-method explanations for how group-level patterns can be (even perfectly) unrepresentative of persons. Therefore, rather than assuming that there are solvable methodological issues underlying the problem, it should be accepted that person-level patterns cannot be inferred from group-level analyses.

One constraint of this person-level approach is that it ignores magnitude information (e.g., participants who use two extreme ends of a measure are treated identically to participants who use two close points of a measure). However, magnitude information can be incorporated into this approach. Researchers can choose an

"imprecision value" (Grice et al., 2020), allowing only certain magnitudes to support a qualitative pattern. Additionally, researchers can plot frequencies of qualitative patterns by different imprecision values, allowing discernment between participants who show small versus large effects (see Speelman & McGann, 2020, Figure 4).

Relatedly, there are other (potentially better) methods for evaluating person-level effects in high-repetition studies that also yield magnitude information, such as person-level effect sizes and confidence intervals (see e.g., Kurz, Johnson, Kellum, & Willer, 2019, and for incorporating measurement error in N =1 designs, Schuurman, Houtveen, & Hamaker, 2015). These methods are not entirely unlike the suggested approach of using Bayesian Prevalence estimation (Ince, Kay, & Schyns, 2022; Ince et al., 2021). However, relative strengths of the Bayesian Prevalence approach are clear: it requires very little statistical knowledge, is easy to implement and interpret, and therefore, is easy to communicate. It is noteworthy that all these methods will require drastic changes in data collection practices for some subdisciplines of experimental psychology, as person-level statistical tests would be subject to the same issues that have pervaded the replicability movement (e.g., number of observations and therefore statistical precision/power).

Another limitation of this research is that only one moral judgment paradigm was used to test method-based noise explanations for the group-to-person generalizability problem. Additionally, much research in moral cognition—including the research conducted in Aims 1-2—utilizes on-the-fly measurement practices (see Flake & Fried, 2020). Future research is needed to determine whether method manipulations fail to remedy the problem in other paradigms and areas of psychology with better measurement

practices. However, as shown earlier, there are obvious non-method (and non-measurement) explanations for the problem. Therefore, a person-level approach should still be used in disciplines with better measurement standards to ensure generalizability.

Finally, the ubiquity of the group-to-person generalizability problem was not assessed. This research simply documented (and replicated) existence proofs. The complexity of the experimental designs employed and the phenomenon under investigation will be important in determining the ubiquity of the group-to-person generalizability. For example, when experiments have factors with more than two levels, or multiple factors, the problem should be more likely to occur because the number of possible person-level patterns explodes as design complexity increases. In contrast, simple binary choice designs common to developmental and comparative work may suffer less from the group-to-person generalizability problem.

Intuitively the problem seems more likely in higher-level areas like social cognition compared to lower-level areas of inquiry like perception. Presumably this is due to most people sharing basic physiological and neural perceptual mechanisms whereas higher-level cognition may be influenced more by individual differences (e.g., values and knowledge). Additionally, social psychologists in particular are often interested in phenomena that participants do not have introspective access to or are motivated to conceal, leading to the overuse of between-subjects designs (rather than the creative use of within-subjects designs). Therefore, any subdisciplines which habitually rely on between-subjects designs to make inferences about psychology may be especially prone to committing the error of assuming that group-level patterns generalize to the

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

Conclusion

Moral psychologists specifically, and psychological scientists broadly, often make claims about, and interpret others' claims as being about, person-level processes. Sometimes, however, these claims are made from experiments that disallow investigation of person-level phenomena. Even when such investigation is possible, these claims are typically derived from group-level patterns, interpreted *as if* they reveal the pervasiveness of some person-level phenomenon. The current work documented novel moral judgment effects, finding that, *on average,* people judge agents who help strangers as more morally good than agents who help family, but they also judge agents who help strangers instead of family as less morally good than agents who help family instead of strangers. Moreover, these patterns were underlied by differential obligations attributed to strangers versus family. However, implementing person-level analytic techniques suggested that the documented set of moral judgments never described most participants. Therefore, it was suggested that additional research is necessary to arrive at valid, *in general*, conclusions about the nature of relationship obligations and their impact on moral judgments. Overall, this work confirms and builds upon previous warnings that this practice can lead to serious errors in inference, as (sets of) group-level patterns need not reflect even a simple majority of sampled persons. At worst, these patterns need not reflect even a single person. Put simply, psychology is a feature of persons, not averages or distributions. Therefore, person-level design and analytic approaches should be customary for the validity and future of (moral) psychological science.

References

Bakdash, J.Z., & Marusich, L.R. (2017). Repeated measures correlation. *Frontiers in Psychology, 8,* 456.

Baron, J., & Miller, J.G. (2000). Limiting the scope of moral obligations to help: A cross-cultural investigation. *Journal of Cross-Cultural Psychology, 31*(6), 705-727.

Berry, Z., Lewis, N.A., Sowden, W.J. (2021). The double-edged sword of loyalty. *Current Directions in Psychological Science, 30*(4), 321-326.

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

Bleske-Rechek, A., Nelson, L.A., Baker, J.P., Remiker, M.W., & Brandt, S.J. (2010). Evolution and the trolley problem: People save five over one unless the one is young, genetically related, or a romantic partner. *Journal of Social, Evolutionary, and Cultural Psychology, 4*(3), 115-127.

Bloom, P. (2011). Family, community, trolley problems, and the crisis in moral psychology. *The Yale Review, 99*(2), 26-43.

Brandt, M.J., & Morgan, G.S. (2022). Between-person methods provide limited insight about within-person belief systems. *Journal of Personality and Social Psychology, 123*(3), 621-635.

Burnstein, E., Crandall, C., & Kitayama, S. (1994). Some neo-Darwinian decision rules for altruism: Weighing cues for inclusive fitness as a function of the biological importance of the decision. *Journal of Personality and Social Psychology, 67*(5), 773-789.

Clark, M., & Mills, J. (1979). Interpersonal attraction in exchange and communal relationships. *Journal of Personality and Social Psychology, 37,* 12-24.

Cottrell, C. A., Neuberg, S. L., & Li, N. P. (2007). What do people desire in others? A sociofunctional perspective on the importance of different valued characteristics. *Journal of Personality and Social Psychology, 92*, 208–231.

Curry, O.S. (2016). *Morality as cooperation: A problem-centred approach.* In T. K. Shackelford & R. D. Hansen (Eds.), *Evolutionary psychology. The evolution of morality* (p. 27–51). Springer International Publishing.

Curry, O.S., Chesters, M.J., & Van Lissa, C.J. (2019). Mapping morality with a compass: Testing the theory of 'morality-as-cooperation' with a new questionnaire. *Journal of Research in Personality, 78,* 106-124.

Curry, O.S., Mullins, D.A., & Whitehouse, H. (2019). Is it good to cooperate? Testing the theory of morality-as-cooperation in 60 societies. *Current Anthropology, 60*(1), 47-69.

Earp, B.D., McLoughlin, K.L., Monrad, J.T., Clark, M.S., & Crockett, M.J. (2021). How social relationships shape moral judgment. *Nature Communications, 12*(1), 6257.

Everett, J.A.C., Faber, N.S., Savulescu, J., & Crockett, M.J. (2018). The costs of being consequentialist: Social inference from instrumental harm and impartial beneficence. *Journal of Experimental Social Psychology, 79,* 200-216.

Faul, F., Erdfelder, E., Lang, A., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39*, 175–191.

Fisher, A.J., Medaglia, J.D., & Jeronimus, B.F. (2018). Lack of group-to-individual generalizability is a threat to human subjects research. *Proceedings of the National Academy of Sciences, 115*(27), E6106-E6115.

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

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

Goodwin, G.P. (2015). Moral character in person perception. *Psychological Science, 24,* 38-44.

Goodwin, G.P., Piazza, J., & Rozin, P. (2014). Moral character predominates in person perception and evaluation. *Journal of Personality and Social Psychology, 106*(1), 148-168.

Graham, J., Nosek, B.A., Haidt, J., Iyer, R., Koleva, S., & Ditto, P.H. (2011). Mapping the moral domain. *Journal of Personality and Social Psychology, 101*(2), 366-385.

Gray, K., Young, L., & Waytz, A. (2012). Mind perception is the essence of morality. *Psychological Inquiry, 23*(2), 101-124.

Greitemeyer, T., Rudolph, U., & Weiner, B. (2003). Whom would you rather help: An acquaintance not responsible for her plight or a responsible sibling? *The Journal of Social Psychology, 143*(3), 331-340.

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

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

Hester, N., & Gray, K. (2020). The moral psychology of raceless genderless strangers. *Perspectives on Psychological Science*, *15*(2), 216–230.

Hughes, J.S. (2017). In a moral dilemma, choose the one you love: Impartial actors are seen as less moral than partial ones. *British Journal of Social Psychology, 56,* 561-577.

Hughes, J., Creech, J.L., & Strosser, G.L. (2016). Attributions about morally unreliable characters: Relationship closeness affects moral judgments. *Basic and Applied Social Psychology, 38*(4), 173-184.

Kelley, H.H. (1967). Attribution theory in social psychology. *Nebraska Symposium on Motivation, 15,* 192-238.

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

Killen, M., & Turiel, E. (1998). Adolescents' and young adults' evaluations of helping and sacrificing for others. *Journal of Research on Adolescence, 8*(3), 355-375.

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

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

Kurzban, R., DeScioli, P., & Fein, D. (2012). Hamilton vs. Kant: Pitting adaptations for altruism against adaptations for moral judgment. *Evolution and Human Behavior, 33,* 323-333.

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

Lee, J., & Holyoak, K.J. (2020). "But he's my brother": The impact of family obligation on moral judgments and decisions. *Memory and Cognition, 48,* 158-170.

Lieberman, D., & Lobel, T. (2012). Kinship on the Kibbutz: Coresidence duration predicts altruism, personal sexual aversions and moral attitudes among communally reared peers. *Evolution and Human Behavior, 33*, 26-34.

Linke, L.H. (2012). Social closeness and decision making: Moral, attributive and emotional reactions to third party transgressions. *Current Psychology, 31,* 291-312.

MacFarquhar, L. (2015). *Strangers drowning: Grappling with impossible idealism, drastic choices, and the overpowering urge to help.* New York, New York: Penguin Press.

Madsen, E.A., Tunney, R.J., Fieldman, G., Plotkin, H.C., Dunbar, R.I.M., Richardson, J., & McFarland, D. (2007). Kinship and altruism: A cross-cultural experimental study. *British Journal of Psychology, 98,* 339-359.

Malle, B. (2021). Moral judgments. *Annual Review of Psychology, 72,* 293-318.

Marshall, J., Wynn, K., & Bloom, P. (2020). When do children and adults take social relationship into account when evaluating other peoples' actions? *Child Development, 91,* 1395-1435.

Marshall, J., Gollwitzer, A., Mermin-Bunnell, N., Shinomiya, M., Retelsdorf, J., & Bloom, P. (2022). How development and culture shape intuitions about prosocial obligations. *Journal of Experimental Psychology: General*.

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

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

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

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

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

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

Neff, K.D., Turiel, E., & Anshel, D. (2002). Reasoning about interpersonal responsibility when making judgments about scenarios depicting close personal relationships. *Psychological Reports, 90,* 723-742.

Niemi, L., Wasserman, E., & Young, L. (2018). The behavioral and neural signatures of distinct conceptions of fairness. *Social Neuroscience, 13*(4), 399-415.

Palan, S., & Schitter, C. (2018). Prolific. Ac—A subject pool for online experiments. *Journal of Behavioral and Experimental Finance, 17*, 22–27.

Passarelli, T. O., & Buchanan, T. W. (2020). How do stress and social closeness impact prosocial behavior? *Experimental Psychology, 67*(2), 123–131.

Petrinovich, L., O'Neill, P., & Jorgensen, M. (1993). An empirical study of moral intuitions: Toward an evolutionary ethics. *Journal of Personality and Social Psychology, 64*(3), 467-478.

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

Rai, T.S., & Fiske, A.P. (2011). Moral psychology is relationship regulation: Moral motives for unity, hierarchy, equality, and proportionality. *Psychological Review, 188,* 57-75.

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

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

Schein, C. (2020). The importance of context in moral judgments. *Perspectives on Psychological Science, 15*(2), 207-215.

Schein, C., & Gray, K. (2016). Moralization and harmification: The dyadic loop explains how the innocuous becomes harmful and wrong. *Psychological Inquiry, 27,* 62-65.

Schein, C., & Gray, K. (2018). The theory of dyadic morality: Reinventing moral judgment by redefining harm. *Personality and Social Psychology Review, 22*(1), 32-70.

Simpson, A., Laham, S.M., & Fiske, A.P. (2016) Wrongness in different relationships: Relational context effects on moral judgment. *The Journal of Social Psychology*, *156*(6), 594-609.

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

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

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

Sznycer, D., De Smet, D., Billingsley, J., & Lieberman, D. (2016). Coresidence duration and cues of maternal investment regulate sibling altruism across cultures. *Journal of Personality and Social Psychology, 111*(2), 159-177.

Tepe, B., & Aydinli-Karakulak, A. (2019). Beyond harmfulness and impurity: Moral wrongness as a violation of relational motivations. *Journal of Personality and Social Psychology*, *117*(2), 310–337.

Tomasello, M. (2020). The moral psychology of obligation. *Behavioral and Brain Sciences,* 1-33. doi:10.1017/S0140525X19001742

Uhlmann, E.L., Zhu, L.L., Pizarro, D.A., & Bloom, P. (2012). Blood is thicker: Moral spillover effects based on kinship. *Cognition, 124*(2), 239-243.

Waytz, A., Dungan, J., & Young, L. (2013). The whistleblower's dilemma and the fairness-loyalty tradeoff. *Journal of Experimental Social Psychology, 49,* 1027-1033.

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

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

Yudkin, D.A., Gantman, A., Hofmann, W., & Quoidbach, J. (2021). Binding moral values gain importance in the presence of close others. *Nature Communications, 12*(1), 2718.