PNAS RESEARCH ARTICLE | PSYCHOLOGICAL AND COGNITIVE SCIENCES

Exposing omitted moderators: Explaining why effect sizes differ in the social sciences

Antonia Krefeld-Schwalb[a,1], Eli Rosen Sugerman[b], and Eric J. Johnson[b]

Edited by Timothy Wilson, University of Virginia, Charlottesville, VA; received April 18, 2023; accepted January 26, 2024

Policymakers increasingly rely on behavioral science in response to global challenges, such as climate change or global health crises. But applications of behavioral science face an important problem: Interventions often exert substantially different effects across contexts and individuals. We examine this heterogeneity for different paradigms that underlie many behavioral interventions. We study the paradigms in a series of five preregistered studies across one in-person and 10 online panels, with over 11,000 respondents in total. We find substantial heterogeneity across settings and paradigms, apply techniques for modeling the heterogeneity, and introduce a framework that measures typically omitted moderators. The framework's factors (Fluid Intelligence, Attentiveness, Crystallized Intelligence, and Experience) affect the effectiveness of many text-based interventions, producing different observed effect sizes and explaining variations across samples. Moderators are associated with effect sizes through two paths, with the intensity of the manipulation and with the effect of the manipulation directly. Our results motivate observing these moderators and provide a theoretical and empirical framework for understanding and predicting varying effect sizes in the social sciences.

online data collection | choice architecture | moderators

Significance

The effectiveness of behavioral interventions often differs sizably across settings. By repeating identical paradigms across 11 online and offline samples, we study heterogeneity in a large-scale experiment. We exploit the natural variation within and across sample populations to i) purposively increase variation and ii) identify moderators. Effect sizes vary substantially and systematically. We explain the observed heterogeneity with moderators that vary between panels and interact with the effect sizes, with both the intensity of the manipulation and the direct effect of the manipulation. We propose a framework of moderators to advance theory building and ultimately, to design more effective policy applications of behavioral research.

Effective response to global challenges often requires individual behavior change. A successful vaccine achieves efficiency only if people choose to be vaccinated. Public transport will reduce carbon emissions only if citizens change how they commute. Behavioral science can help to advance change using behavioral interventions, also called nudges, behavioral insights, or choice architecture. Behavioral interventions are typically less costly than monetary incentives, intensive persuasion, or education (1). This has resulted in many calls for their increased use in policy, yet others argue that applications are premature (2–4). One challenge to identifying successful interventions is that the size of change produced by an intervention can vary across settings and populations (5–8). Even replicating an identical intervention in different laboratories can produce quite dissimilar results (9–11) and effects often differ across subsets of a population [for example, as a function of SES (12)]. This has led to a call for a "heterogeneity revolution" in applying behavioral science (13). Understanding heterogeneity is important beyond policy applications: It enables researchers to build more complete and robust theories, exposing boundary conditions and identifying additional predictors.

Consider a policy that reduces carbon emissions by defaulting utility customers into more expensive, sustainable energy. While this generally produces a large increase in the use of sustainable energy, default effects can be stickier among people with lower SES (14). This heterogeneity may result in an undesirable policy: The poorer, who are responsible for far fewer emissions, pay relatively more than the rich. This is also important for theory because it informs researchers of settings where defaults may not work as intended.

This paper i) demonstrates significant differences in the effect sizes in standard, well-replicated paradigms often used to develop interventions, ii) suggests a guiding framework for measuring variables that underlie this heterogeneity, and iii) demonstrates techniques for improving the robustness of research. We build on the key insight that the size of the effect of one variable (like a default manipulation) on another (like a choice) may depend upon a third variable (like SES) that varies but is not observed in the original setting (13). To understand heterogeneity, it is necessary to measure and model this third variable—an omitted moderator. Neglecting it limits our ability to generalize results. We focus on text-based interventions, which aim to influence decision processes using context or wording changes, and often serve as the basis for real-world behavioral interventions.

We leverage the variation that exists across different offline and online panels, ranging from widely used commercial panels to a student laboratory sample. This purposive variation allows us to study the larger question of what drives heterogeneity in effect sizes (15). We observe that effect sizes vary markedly across panels, as revealed by significant interactions between the manipulations and the panels. We show that exposing omitted moderators

Author affiliations: [a]Rotterdam School of Management, Department of Marketing Management, Erasmus University, Rotterdam 3011 LC, Netherlands; and [b]Columbia Business School, Marketing Division, Columbia University, New York City, NY 10027

Author contributions: A.K.-S., E.R.S., and E.J.J. designed research; A.K.-S. and E.R.S. performed research; A.K.-S. analyzed data; and A.K.-S., E.R.S., and E.J.J. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence may be addressed. Email: krefeldschwalb@rsm.nl.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2306281121/-/DCSupplemental.

Published March 11, 2024.

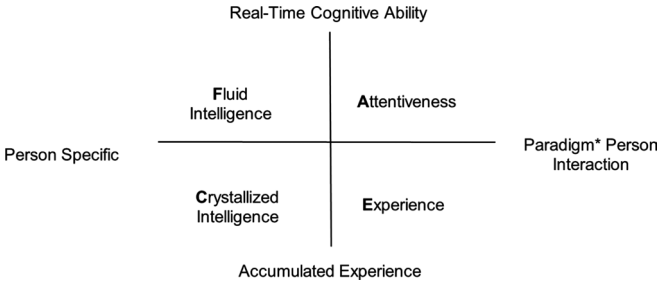

Fig. 1. Graphical illustration of the FACE factor model.

through measurement can explain this heterogeneity. While our primary focus lies in studying heterogeneity of effects sizes, understanding panel differences is also important since online panels have become increasingly common across the social sciences, including psychology, economics, sociology, and political science (16–20).

Imagine that an identical manipulation, like a default, has a large effect in one panel and an insignificant effect in another. Can differences in the characteristics of the panels themselves explain this discrepancy? By administering identical paradigms across panels, we minimize variability in study execution and reveal correlations with moderators that differ between the panels (21). Our approach can be regarded as a complementary extension to large, coordinated replication studies (10, 11, 22). However, instead of minimizing variability, we leverage differences in omitted moderators to manufacture and explain differences in effect sizes (15).

Previous studies investigating demographic differences between online (23, 24) and other panels (25, 26) have led to mixed results. Some suggest small differences between panels' demographic characteristics and effect sizes, and other research has identified differences beyond demographics that are more strongly associated with effect sizes (23, 27, 28).

Selecting which moderators to observe is challenging. There are many possible moderators, many of which may be task specific. Measuring moderators also requires time from respondents. Demographics are often used to model heterogeneity but may be indirectly related to the tasks. Instead, we focus on moderators that are more closely related to text-based interventions, particularly in an online setting. We propose measures that are brief and unobtrusively collected, such as response times. Later, we expand from this useful starting point to more sophisticated measures.

Two widely discussed cognitive constructs are likely to affect how individuals encode, process, and integrate information across paradigms: Fluid and Crystallized Intelligence. We also examine measures that are specific to respondents' interaction with the paradigm: The Attentiveness of respondents to a particular paradigm and their past Experience with that paradigm. We summarize this framework using the acronym FACE: Fluid Intelligence, Attentiveness, Crystallized Intelligence, and Experience. As illustrated in Fig. 1, these factors differ on two dimensions: whether they are a function of the person or the interaction of the person and the paradigm and whether they represent real-time cognitive abilities or serve as indicators of accumulated experience.

Many researchers have documented individual differences in concepts related to Fluid Intelligence, such as the speed of processing, numerical sophistication, numeracy, and cognitive reflection (29). It is separable from other forms of intelligence and has been shown to decrease with age. Interventions in the behavioral sciences are likely to be affected by similar variables, for instance, numerical skills. Imagine that an intervention uses numeric data to communicate the frequency of vaccine side effects. The effectiveness of this intervention might be moderated by Fluid Intelligence, because of differences in numeric skills. Similarly, Peters et al. show that numeracy affects framing and risk attitudes (30). We initially measure Fluid Intelligence using related surveys such as the Berlin Numeracy Test (31) and the Cognitive Reflection Task and variants (32, 33). We later adopt more extensively tested measures, including Ravens-like matrices and 3-D rotation tasks (34) (*SI Appendix*, sections 3 and 5).

In contrast, Crystallized Intelligence focuses on knowledge of the world and is thought to be the result of accumulated experience, mostly increasing with age (35, 36). The effect of interventions requiring text comprehension or knowledge of the world might increase with Crystallized Intelligence. Measures of Crystallized Intelligence have long played a separate role in understanding cognitive performance and can compensate for age-related decreases in Fluid Intelligence (35, 36). We again begin with simple measures, using one's age and education as a proxy for Crystallized Intelligence since they have been shown to correlate (37). Later, we adopt existing scales that measure a person's passive and active vocabulary. Some research suggests that individuals high in Crystallized Intelligence may rely more heavily on their knowledge and experience and are thus less likely to be influenced by certain interventions (38).

Attentiveness of respondents has been an enduring concern in in-person and online research. Respondents not paying attention to tasks can reduce the estimated effect sizes (23, 39–41). In survey-based experiments, attention checks are commonly used to filter for attentive respondents. All else equal, the amount of time an individual spends interacting with presented information is likely correlated with how deeply the respondent is processing the presented material. We measure Attentiveness using respondents' response times in the paradigms, which are easy and free to collect in standard survey packages.

Finally, Experience [also termed Naivete (27)], which refers to familiarity with the paradigm, is a likely moderator of many paradigms. Some panels are composed of a limited number of respondents who spend a significant amount of time doing experimental tasks. Thus, they may have experienced similar or identical paradigms on multiple occasions. Amazon's MTurk panel, for example, was documented to have only 7,300 active participants, who reported doing over 300 experimental tasks (42). There is evidence that this exposure can reduce effect sizes (27, 43, 44). Such concerns have led to the development of alternative forms of measures such as the CRT (33). We measure Experience by asking respondents directly, for each paradigm, whether they had seen these materials before.

Importantly, the FACE framework is a robust, domain-general starting point for identifying variables that have been previously neglected and may be particularly relevant for text-based interventions. Other variables, such as demographics and personality, may also be important moderators in some settings—but the FACE framework may generally be more proximal and relevant to tasks involving text-based stimuli.

The standard approach to studying heterogeneity estimates the interaction between a moderator and the manipulation. Conceptually, however, moderators are associated with effect sizes through two paths—effecting manipulation intensity and interacting with the effect of the manipulation, respectively, as depicted in Fig. 2.

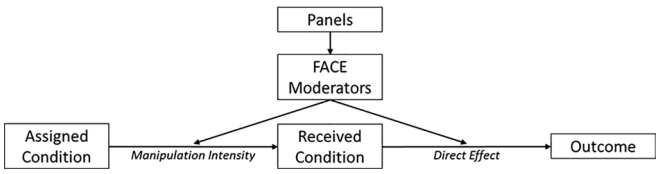

Fig. 2. Graphical illustration of the two paths of moderation.

Studies in the behavioral sciences often do not assess the intensity of the manipulation received by respondents. Without these measurements, one cannot separately observe both moderation effects. Measuring manipulation intensity allows us to measure two separable associations with moderators:

First, omitted moderators can be correlated with the manipulation intensity. For example, a respondent with low Attentiveness who speeds through the Unusual Disease problem (Table 1, row 2) problem may miss the lives lost (lives saved) wording in the condition assigned to them. They will produce a smaller effect than someone high in Attentiveness, who reads carefully and thus accurately encodes the text. Attentiveness would interact with the effect of the assigned condition on the received condition.

Second, unobserved moderators can be correlated with effects directly, holding the manipulation intensity constant. A respondent high in Fluid Intelligence might attempt a calculation for the Unusual Disease problem (multiplying outcomes by probabilities) while someone lower in Fluid Intelligence may rely more upon the "intuition" elicited by the frames. Fluid Intelligence, in this

Table 1. Text and outcome measures of the paradigms tested in studies 1 to 5

Paradigm	Condition (Base)	Condition (Treat)
Sunk cost	Imagine that your favorite football team is playing an important game. You have a ticket to the game that you have received for free from a friend. However, on the day of the game, it happens to be freezing cold. What do you do?	Imagine that your favorite football team is playing an important game. You have a ticket to the game that you have paid handsomely for. However, on the day of the game, it happens to be freezing cold. What do you do?
	DV: Definitely Stay home 1-Definitely Go 9 (Drop down menu)	
Framing (unusual disease)	Imagine that the United States is preparing for the outbreak of an unusual disease, which is expected to kill 600 people. Two alternative programs to fight the disease have been proposed. Assume that the exact scientific estimate of the consequences of the program are as follows: If Program A is adopted, 200 people will be saved. If Program B is adopted, there is 1/3 probability that 600 people will be saved and 2/3 probability that no people will be saved. Which program would you choose?	Imagine that the United States is preparing for the outbreak of an unusual disease, which is expected to kill 600 people. Two alternative programs to fight the disease have been proposed. Assume that the exact scientific estimate of the consequences of the program are as follows: If Program A is adopted, 400 people will die. If Program B is adopted, there is 1/3 probability that no people will die and 2/3 probability that 600 people will die. Which program would you choose?
	DV: Program A or Program B	
Less is better	Imagine you are about to leave the country and have received a goodbye gift from a friend. It is a wool coat, from a nearby department store. The store carries a variety of wool coats. The worst costs $100 and the best costs $1,000. The one your friend bought for you costs $110. How generous do you think your friend was?	Imagine you are about to leave the country and have received a goodbye gift from a friend. It is a wool scarf, from a nearby department store. The store carries a variety of wool scarves. The worst costs $10 and the best costs $100. The one your friend bought for you costs $90. How generous do you think your friend was?
	DV: 0 Not generous at all—6 Extremely generous (Drop down menu)	
Default*	Imagine that you just moved to a new state and must get a new driver's license. As you complete the application, you come across the following. Please read and respond as you would if you were actually presented this choice today. We are interested in your honest response. In this state, every person is considered not to be an organ donor unless they choose to be. You are therefore currently not a potential donor. DV: If this is acceptable, click here:[0] Or If you wish to change your status, click here:[1]	Imagine that you just moved to a new state and must get a new driver's license. As you complete the application, you come across the following. Please read and respond as you would if you were actually presented this choice today. We are interested in your honest response. In this state, every person is considered to be an organ donor unless they choose not to be.You are therefore currently a potential donor. DV: If this is acceptable, click here:[1] Or If you wish to change your status, click here:[0]
Trolley Problem	Frank is on a footbridge over the train tracks. He knows trains and can see that the one approaching the bridge is out of control. On the track under the bridge, there are five people; the banks are so steep that they will not be able to get off the track in time. Frank knows that the only way to stop an out-of-control train is to drop a very heavy weight into its path. But the only available, sufficiently heavy weight is a large man wearing a backpack, also watching the train from the footbridge. Frank can shove the man with the backpack onto the track in the path of the train, killing him; or he can refrain from doing this, letting the five die. Is it morally permissible for Frank to shove the man?	Denise is a passenger on a train whose driver has just shouted that the train's brakes have failed, and who then fainted of the shock. On the track ahead are five people; the banks are so steep that they will not be able to get off the track in time. The track has a side track leading off to the right, and Denise can turn the train onto it. Unfortunately, there is one person on the right-hand track. Denise can turn the train, killing the one; or she can refrain from turning the train, letting the five die. Is it morally permissible for Denise to switch the train to the side track?
	DV: Yes or No	

*In studies 1 to 4, we also implemented a neutral condition, see *SI Appendix*, Table S4.

case, would interact separately with the effect of the received condition on the outcome variable.

We expect that FACE moderators might have similar effects on the intensity of the manipulation across different paradigms. We posit that greater Attentiveness, Fluid and Crystalized Intelligence will result in a more intense manipulation. However, higher Experience should produce less elaborate representations, as respondents may not attend to the details that differentiate the present manipulation from past exposures.

The diagram is also an overview of our approach: First, we use different panels to create variation in moderators. Second, we assume that moderators are correlated with effect sizes along two paths, with the intensity of the manipulation and with effect of the manipulation on the outcome variable. Most of our studies will examine interactions between manipulations and moderators, collapsing across the two paths in the figure. Study 5, however, allows us to separately estimate both paths of moderation.

We first demonstrate heterogeneity of effects by documenting panel differences for several standard paradigms. This can be assessed by testing the interaction of the manipulation and the panel. Next, we try to account for these effect size differences across panels using the FACE factors. Our focus is on whether effects sizes decrease or increase with varying levels of FACE factors. For example, we ask whether the framing effect interacts with Attentiveness or Crystallized Intelligence. Finally, we ask whether these interactions reduce or eliminate the panel differences. To ensure robustness of the observed panel differences, we replicate most paradigms in four separate studies.

We conducted five preregistered studies with over 11,000 respondents using 10 online panels and one laboratory student panel. Each study employed well-studied experimental paradigms, that use randomized between-subjects conditions. All paradigms have either been included in a large replication project (9, 10, 22, 45) or been examined in a meta-analysis (46). This provides us with a data-driven benchmark of effect sizes for comparison. All paradigms could be implemented online, were not time-intensive and spanned a range of expected effect sizes. Respondents in all panels saw identical questionnaires. We included panels commonly used in academic online experiments (MTurk, CloudResearch, and Prolific Academic) as well as panels primarily used in commercial market research. We also included two panels labeled by Prolific as nationally representative of the United States and the United Kingdom, respectively.

In Study 1, we examine how the effect sizes of four paradigms varied across 11 panels (*SI Appendix*, section 1). We label these Defaults (47), Framing (48), Less-is-better (49), and Sunk Cost (50). Table 1 reports the text of each paradigms' conditions and outcome measures. Study 2 and Study 4 replicate the results with improved measures of the FACE factors. In Study 3, we extend our results to a paradigm from the broader social science literature, the Trolley Problem (51). In Study 5, we further improve our understanding of heterogeneity by additionally measuring manipulation intensity.

Results

Study 1 produced substantial heterogeneity in effect sizes of the four paradigms across 11 panels and 6,438 respondents, depicted in Fig. 3. The first two plots present the default and framing paradigms, which showed large differences across panels. To characterize these differences, we conducted a regression analysis predicting the responses in each paradigm with the condition, panel, and our key test—their interaction (*SI Appendix*, section 2.2). Eq. **1** illustrates this analysis:

$$y_i = \beta_0 + \delta_0 T_i + \sum_{p=1}^{P} \gamma_0 d_{p_i} + \sum_{p=1}^{P} \epsilon_0 T_i d_{pi}, \qquad [1]$$

y = outcome variable, i = respondent index, T =dummy coded condition, p = index for panel, P = vector of Panels, d_p = dummy coded panel variable.

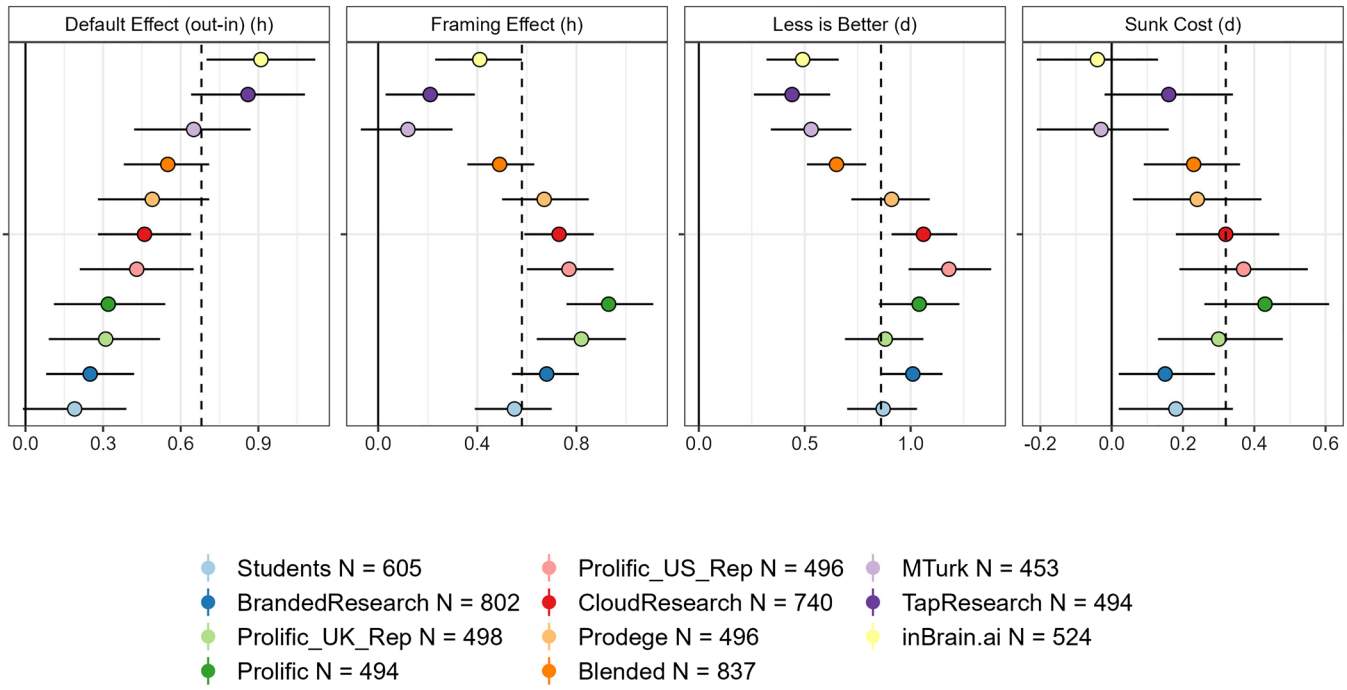

Fig. 3. Effect sizes observed in Study 1, sorted by size of the default effect observed in each panel. Error bars in the effect size plots represent 95% CI around the effect sizes. Sample sizes are reported in the legend. The dashed vertical lines represent the benchmark effect sizes from meta-analyses or a Many Labs replication study.

The analysis confirms what we see in Fig. 3: Panels produced different effect sizes for identical paradigms. Importantly, these differences varied across the paradigms. No one panel produced bigger (or smaller) effects across all paradigms, meaning that no one panel produced "better" (or "worse") results for all items. This is illustrated in Fig. 3 by the different distributions of the default and framing effects across panels. Focusing on two of the most widely used online panels, we find that the default effect was larger on MTurk (D = 0.76) than on Prolific (D = 0.29), while the framing effect was larger on Prolific (D = 0.96) than on MTurk (D = 0.31). Across all panels, default effects are negatively correlated with the effects in other paradigms [e.g., $\tau_{(\text{default, framing})}$ = −0.25]. This inversion demonstrates that differences across the panels do not affect all paradigms equally, suggesting that no one panel produces consistently larger effects. The analysis was robust when respondents who failed attention check questions were omitted (*SI Appendix*, section 2.2).

We suggest that the heterogeneity in effects across panels is due in part to differences in previously unobserved moderators more closely related with the paradigms than demographic variables. Such moderators are differently associated with effects across paradigms. To test this, we first conducted an exploratory factor analysis with oblimin factor rotation of our measured moderators to reduce their dimensionality. This produced the four factors of the FACE framework (RMSEA = 0.03, CFI = 0.99): Fluid Intelligence (cognitive reflection task and numeracy), Attentiveness (response times in the paradigms), Crystallized Intelligence (education), and Experience (with the individual paradigms). The largest correlation between factors was r = −0.368. See *SI Appendix*, section 2.3 for more information on the factor analyses.

The Fig. 4 illustrates the large differences in panel characteristics. Each radar plot presents a panel's average factor scores as red lines. The plot in the top left corner represents the average scores across all panels in Study 1. The remaining plots in the first row depict the profiles of three commonly used panels (Prolific, MTurk, and students). We see that MTurk respondents are low in Attentiveness (z = −0.23) and high in Experience (1.33). In contrast, Prolific respondents are attentive (0.25) and much lower in Experience (−0.03). The student sample is most attentive (0.29) and, naturally, because they are young and just starting higher education, have low Crystallized Intelligence (−1.37). We next assessed moderation by adding the four moderators' main effects and interactions with the condition to the previous regression models, schematically illustrated in Eq. **2**.

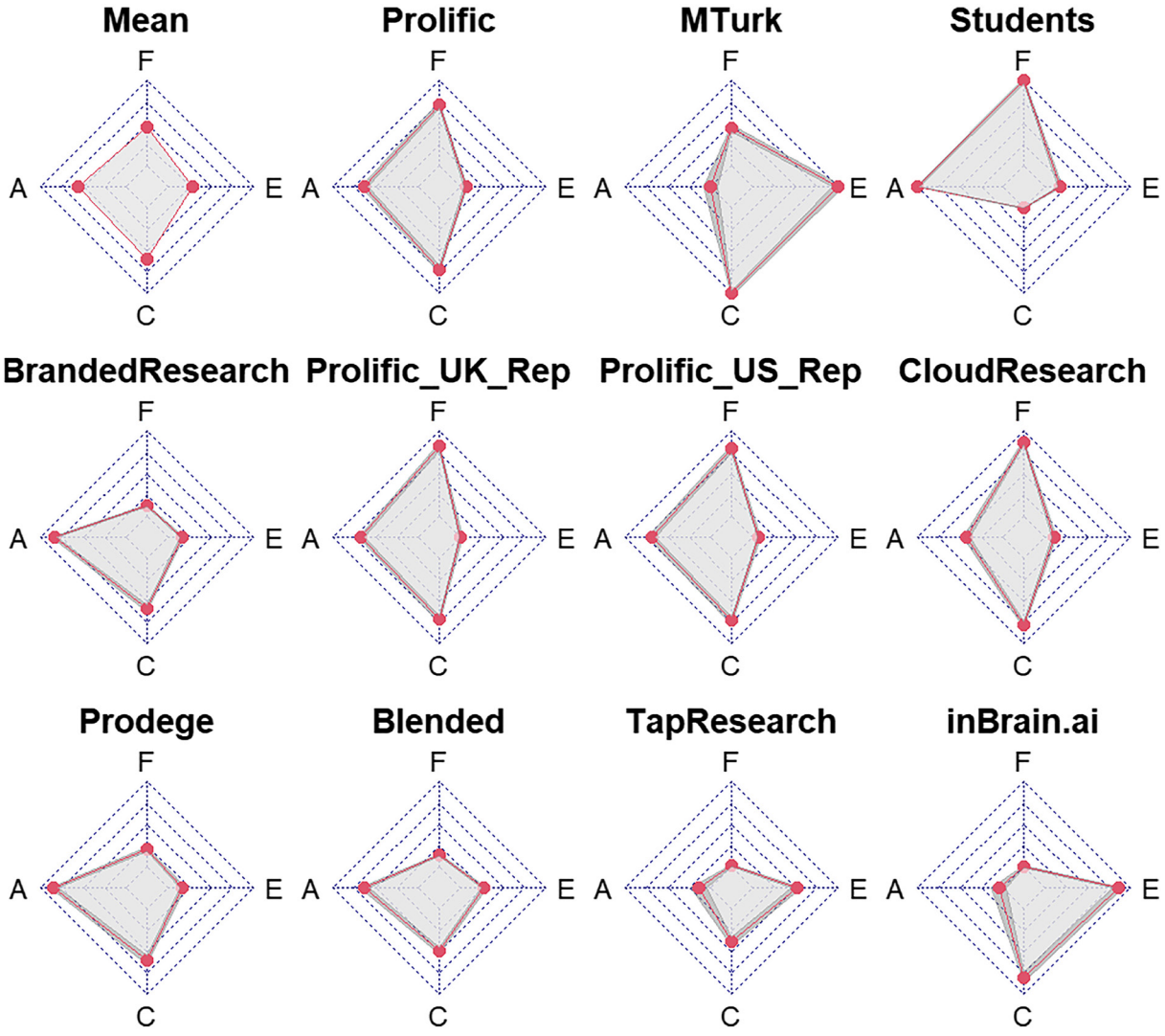

Fig. 4. Radar charts of the panels' average z-scores on the FACE factors in Study 1: F = Fluid Intelligence, A = Attentiveness, C = Crystallized Intelligence, E = Experience. The red lines illustrate the average scores in each panel, the dark gray area around the red line illustrates the 95% CI around these estimates. The inner border is the minimum, and the outer line is the maximum average score on the respective FACE factors across panels.

$$y_i = \beta_0 + \delta_0 T_i + \sum_{p=1}^{P} \gamma_0 d_{pi} + \sum_{p=1}^{P} \epsilon_0 T_i d_{pi} + \beta_1 F_i + \beta_2 A_i + \beta_3 C_i + \beta_4 E_i + \delta_1 T_i F_i + \delta_2 T_i A_i + \delta_3 T_i C_i + \delta_4 T_i E_i, \quad [2]$$

F, A, C, E = factor scores.

Adding moderators to the regression models reduced the panel main effects and interactions. After adding moderators, the panel main effects and interactions explained less variance in the outcome variable. The average explained variance of panel main effect decreased from $\eta^2 = 0.05$ to $\eta^2 = 0.04$, and from $\eta^2 = 0.007$ to $\eta^2 = 0.006$ of the panel condition interactions. Adding the moderators further increased model fit across all paradigms, accounting for model complexity (*SI Appendix*, section 2.4).

The Fig. 5 illustrates the models' conditional average treatment effects (CATEs). Treatment effects here and below refer to the difference between the conditions. Each row represents a moderator, and each column a paradigm. For example, the first cell of the second row depicts how the default effect changes with Attentiveness. One can easily see that each line (a panel) slopes downward, showing that the default effect decreases with higher Attentiveness. In contrast, the bottom cell in that column shows that default effects increase in size with Experience. Experienced (vs. naive) respondents show larger default effects. These are clear demonstrations that Attentiveness and Experience are moderators of the default effect. Slopes are very similar in each cell, indicating that the correlations with each moderator in the paradigm are similar across panels.

If we focus on the first two columns of the figure, we can see that framing and default effects are both correlated with the moderators. However, the effects are different, producing the striking negative correlation of effect sizes across panels in Fig. 3. As we have seen, the size of the default effect decreases as Attentiveness increases. In contrast, the default effect increases with self-reported Experience. The second column shows that these two moderators have the opposite effect on framing: Framing effects increase with Attentiveness and decrease with Experience. Thus, the markedly different effects of panel on the default and framing effects are explained by the moderators, which differ across panels and have opposite effects on the two paradigms. These results show that heterogeneity is a function

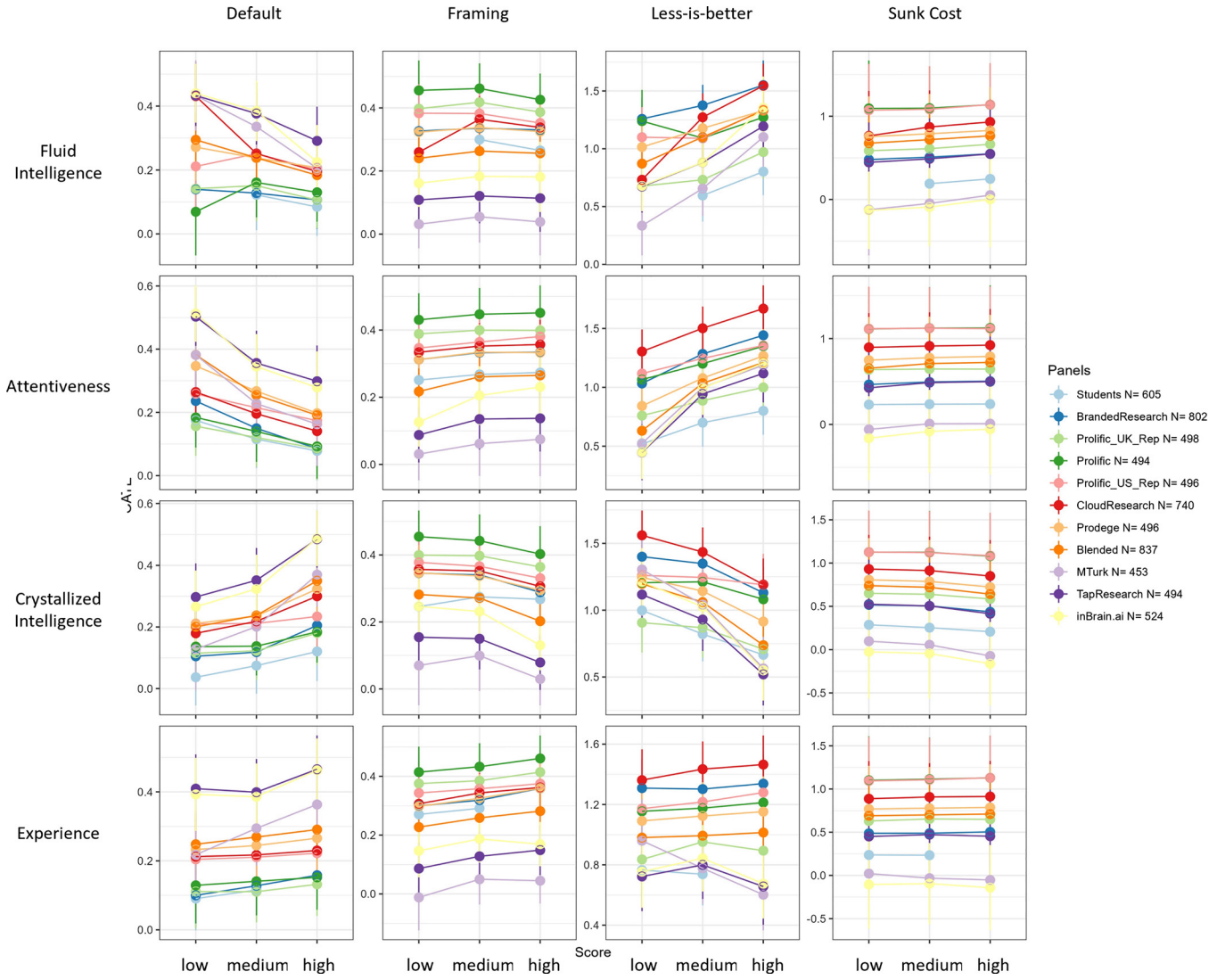

Fig. 5. Correlation with moderators. Each column represents the conditional average treatment effects in one paradigm (on the y-axis) as a function of different levels (low, medium, and high on the x-axis) of FACE moderators in rows and panel (color). The error bars illustrate 95% CI around the estimate.

of both task and panel characteristics and again demonstrates that no one panel is "better."

To test the reliability and the robustness of results from Study 1, we conducted 3 additional preregistered studies. Study 2 (N = 1,057), Study 4 (N = 1,460), and Study 5 (N = 1,460) replicated the pattern of reversal for the default and framing paradigms on the MTurk and Prolific panels (*SI Appendix*, section 1.3). To illustrate the stability of these findings, Fig. 6 plots the effect sizes in the four experimental paradigms and three panels that we used repeatedly across studies. Each respondent was unique and could not participate in more than one study.

We think that these results are unexpected. In a small, separate survey, we asked judgment and decision-making researchers (N = 67) to predict panel differences. They vastly underestimated the observed heterogeneity across panels and incorrectly predicted that students would always produce the largest effect sizes (*SI Appendix*, section 6).

Study 3 extended our analysis of heterogeneity using the Trolley Problem ((51); *SI Appendix*, section 3). In Study 4, we improved our measure of Crystallized Intelligence and replicated the same analysis as in the between-subjects paradigms of Study 1. We also added TIPI (52) scales for measuring personality traits as alternative moderators. Adding FACE factors to the panel effects helped to increase the explained variance in the outcome variables by an average of 30%, compared to an 8% increase by adding demographic variables or 21% by adding TIPI (*SI Appendix*, section 4.3).

Finally, Study 5 unpacks the two moderation paths outlined in the introduction. To test the effect on manipulation intensity, we asked respondents (at the end of the survey) which of the conditions they saw in each paradigm (received condition). We assume that the more certain respondents were about which condition had been assigned to them, the stronger the intensity of the manipulation. We regressed this measure on the assigned condition and the moderators. Their significant interaction indicates a moderation of manipulation intensity by FACE factors (as illustrated in the graphic in the introduction). The *Left* panel of Fig. 5 plots the standardized coefficients of these models.

As predicted, the manipulation intensity was similarly affected by FACE factors across the paradigms. Greater Fluid Intelligence, Attentiveness and Crystallized Intelligence was associated with increases of manipulation intensity across the paradigms (as indicated by the positive coefficients) while increased Experience correlated with reduced manipulation intensity.

In contrast, we do not predict similar effects for each paradigm when considering how FACE factors correlate with the effect of the manipulation (plotted in the right column of Fig. 7). Such predictions depend upon theories for different strategies in each paradigm. For example, Crystallized and Fluid Intelligence have opposite effects on defaults, for reasons that we leave to future research. Similarly, the negative correlation between defaults and the remaining items can be explained by a negative main effect of manipulation intensity on the outcome in the default paradigm (*SI Appendix*, Table S23). Thus, stronger manipulations in the default paradigm are associated with reduced effect sizes.

In Study 5, we also used improved measures of Crystallized and Fluid Intelligence and replicated the FACE factor structure. We further found that the FACE factors explain more heterogeneity than TIPI or demographic variables (*SI Appendix*, Fig. S8).

In sum, Study 5 helps us understand the (replicated) pattern of effect sizes seen in Figs. 3 and 6. A large part of the effect of moderators is relatively uniform across paradigms and reflects moderation through the manipulation intensity (see the left column in Fig. 7). In contrast, the size and direction of the effect of moderators on the outcome through the direct effect of the manipulation is paradigm specific (see the right column in Fig. 7). Both default and framing effects show paradigm-specific correlations with moderators, which account (in part) for the markedly different distribution of effect sizes across panels between the paradigms (Figs. 3 and 6).

Discussion

We demonstrate that basic phenomena from the decision and social sciences vary more widely across samples than previously documented (19, 25, 26). The heterogeneity we observe suggests that we generally cannot (or do not) study generally applicable effect sizes. Instead, we study effect sizes that are conditional on the distribution of moderators within a sample. The observed pattern of heterogeneity is systematic, influencing effects through two distinct paths. We see a general association of the FACE moderators with the manipulation intensity. In contrast, we observe varying associations of FACE moderators with the direct effects of the manipulations, even when controlling for manipulation intensity (Fig. 7).

These results suggest that heterogeneity is not a nuisance, but rather something to embrace in both theory and practice. Moderators, including those proposed in the FACE framework, should play a larger role in theory—both because they are unavoidable in application and because they may promote better generalizability across contexts.

Empirically, our work suggests a path toward more robust results. First, we suggest that experiments should include measures of manipulation intensity, similar to manipulation checks. Such measures allow us to confirm the received condition and observe the correlations of moderators and the manipulation intensity. Such analyses also require adding measures of formerly

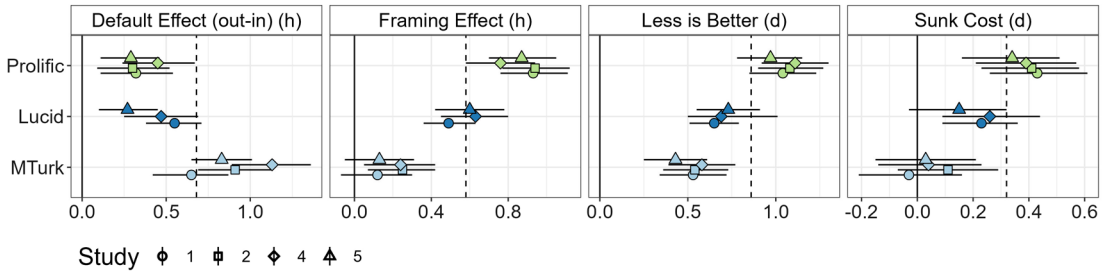

Fig. 6. Effect sizes in default, framing, less-is-better, and sunk cost observed in studies 1, 2, 4, and 5 in the online panels MTurk, Lucid, and Prolific. Error bars in the effect size plots represent 95% CI around the effect sizes. Solid vertical lines indicate zero and dashed vertical lines represent the benchmark effect sizes from meta-analyses or Many Labs replication.

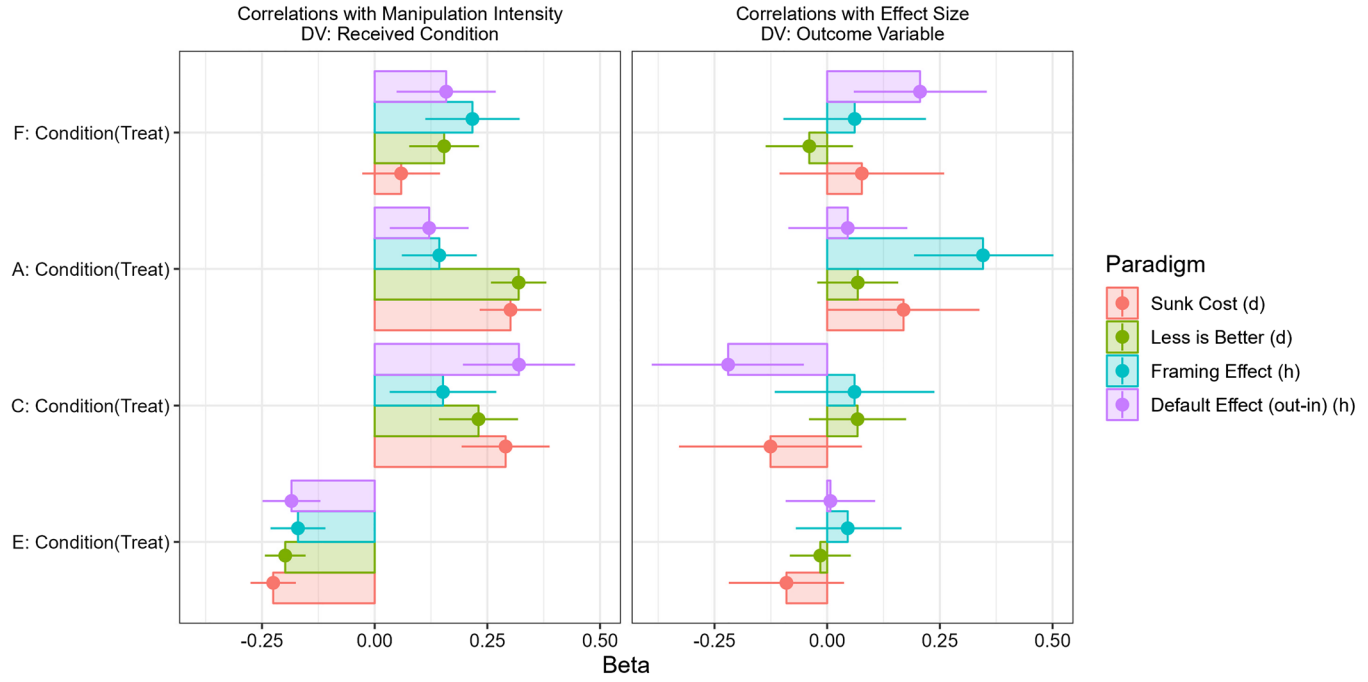

Fig. 7. Standardized regression coefficients in Study 5 on the received condition (in the *Left* panel) and outcome variables (in the *Right* panel). Moderation is tested with interaction terms of the dummy coded condition [Condition(Treat)] and the FACE moderators. Positive estimates indicate increased manipulation intensity and stronger effects and negative estimates indicate reduced manipulation intensity and weaker effects with higher levels of the moderator. Error bars illustrate 95% CI around the coefficients. The graphs plot the interactions between the treatment and the FACE moderators, the full regression tables can be found in *SI Appendix*, Tables S23 and S28.

unobserved moderators. We propose measures that are cost and resource efficient and that are particularly relevant for text-based interventions[*]. Other settings might require different moderators and measures.

Finally, the research suggests that it is important to pursue, observed and estimate purposive variation, allowing one to observe a sufficient range of the moderators. Techniques might include splitting a sample among multiple panels or ensuring that experiments, including RCTs, provide sufficient variation to ensure generalizability to other settings.

Overall, this framework serves as a starting point to help social and behavioral scientists develop more robust theory and more confidently apply behavioral science in policy. While our framework helps policy makers to choose samples that best match their target populations, our results also provide guidance for considering different interventions for different contexts in order to maximize effects. The FACE framework, in particular, may also contribute to our understanding of failed and successful replications, an issue at the forefront of the replication crisis in social science. We suggest that collecting FACE factors would be useful in future replication attempts. For example, variation in FACE factors could explain why some paradigms are more robustly replicated in some samples but not in others. While the present examination of moderators is correlational, future research could attempt experimental manipulation of moderators to test causal effects.

Beyond this paper, embracing heterogeneity in samples may be a starting point for understanding the relationship between online experiments and field applications. Our results should encourage researchers to consider the match between the experiment and the application. Specifically, estimating the effect of attention, experience and cognitive resources may lead toto more generalizable theories with greater contribution to policy and practice.

Materials and Methods

In Study 1a, we tested four paradigms on eight online panels (detailed information can be found in *SI Appendix*, section 1.1): The default effect [organ donation (47)], the framing effect [unusual disease (48)], the less-is-better effect (49), and the sunk cost effect (50). We collected demographic variables (gender, age, income, and political orientation), Crystallized Intelligence (education, age), Fluid Intelligence (the cognitive reflection task and numeracy (31, 32), Experience with each paradigm and online research in general, and measures of Attentiveness (response times), as well as two attention checks. In addition to panels that are ubiquitous in academic research (MTurk, Cloud Research, and Prolific), we used commercial providers that vary in quality (four distinct panels and a blended panel from Lucid Marketplace). We used a large in-person sample of students from a European university (in Study 1b) and included samples from Prolific stratified on age, race, and gender to be representative of the United States and the United Kingdom. We also replicated the survey on new samples from MTurk and Prolific, pre-registering the pattern of effect sizes (Study 2).

In Study 3, we improved our measure of Fluid Intelligence. We added the 3-D rotation items and Raven-like matrices from the International Cognitive Ability Resource (34). Here, we focused on two panels most used in academic research, Prolific and MTurk, since they showed very different patterns of moderators in Study 1. We also varied the time of day (1:00 AM, 9:00 AM, and 5:00 PM EST) and day of the week (Wednesday and Sunday) of data collection, because this had been previously described as a source of variation (53). In Study 4 we added vocabulary tests to improve measures of Crystallized Intelligence. We again distributed the survey across six data collection periods and sampled respondents from Prolific, MTurk, and Lucid Marketplace. Study 5 was distributed in the same manner as Study 4. We used the 3D-rotation tasks and matrix questions as measures of Fluid Intelligence and vocabulary test for Crystallized Intelligence. We further asked respondents for each paradigm, on a five-point Likert scale,

[*]A template for measuring FACE moderators can be found in the online repository: https://osf.io/7kqg9/.

which of the two conditions for each paradigm they received. To measure the received condition, we presented them with paraphrases of the texts of the two conditions of each paradigm and asked them which condition they saw on a five-point Likert scale ranging from "I definitely saw A" to "I definitely saw B" (*SI Appendix*, Table S22).

We report our analyses using regressions and factors based on a confirmatory factor analysis, but also examined the moderators using seemingly unrelated regressions (when appropriate) and structural equation modeling, and conducted exploratory factor analysis to identify the FACE factors in each study (*SI Appendix*, sections 2.5 and 2.6). The effects we report are robust across procedures. All data and code are available online (https://osf.io/7kqg9/).

Data, Materials, and Software Availability. Anonymized Survey data have been deposited in Open Science Framework (https://osf.io/7kqg9/) (54). All other data are included in the manuscript and/or *SI Appendix*.

ACKNOWLEDGMENTS. We would like to thank Ibitayo Fadayomi, Simon Xu, and Xuwen Hua for research assistance; Dan Schley and Jason Roos for advice; and Daniel Lakens and Don Green for helpful comments.

1. G. M. Walton, A. J. Crum, *Handbook of Wise Interventions* (Guilford Press, 2021).
2. H. IJzerman *et al.*, Use caution when applying behavioural science to policy. *Nat. Hum. Behav.* **4**, 1092–1094 (2020).
3. N. Chater, G. Loewenstein, The i-frame and the s-frame: How focusing on individual-level solutions has led behavioral public policy astray. *Behav. Brain Sci.* **46**, e147 (2022).
4. R. H. Thaler, C. R. Sunstein, *Nudge: The Final Edition* (Penguin Books, 2021).
5. K. E. Stanovich, R. F. West, Individual differences in rational thought. *J. Exp. Psychol. Gen.* **127**, 161–188 (1998).
6. B. B. McShane, J. L. Tackett, U. Böckenholt, A. Gelman, Large-scale replication projects in contemporary psychological research. *Am. Statistician* **73**, 99–105 (2019).
7. A. Gelman, The connection between varying treatment effects and the crisis of unreplicable research. *J. Manage* **41**, 632–643 (2015).
8. D. A. Kenny, C. M. Judd, The unappreciated heterogeneity of effect sizes: Implications for power, precision, planning of research, and replication. *Psychol. Methods* **24**, 578–589 (2019).
9. R. A. Klein *et al.*, Investigating variation in replicability: A "many labs" replication project. *Soc. Psychol.* **45**, 142–152 (2014).
10. R. A. Klein *et al.*, Many labs 2: Investigating variation in replicability across samples and settings. *Adv. Methods Practices Psychol. Sci.* **1**, 443–490 (2018).
11. T. D. Stanley, E. C. Carter, H. Doucouliagos, What meta-analyses reveal about the replicability of psychological research. *Psychol. Bull.* **144**, 1325–1346 (2018).
12. K. Mrkva, N. A. Posner, C. Reeck, E. J. Johnson, Do nudges reduce disparities? Choice architecture compensates for low consumer knowledge. *J. Marketing* **85**, 67–84 (2021).
13. C. J. Bryan, E. Tipton, D. S. Yeager, Behavioural science is unlikely to change the world without a heterogeneity revolution. *Nat. Hum. Behav.* **5**, 980–989 (2021).
14. C. Ghesla, M. Grieder, R. Schubert, Nudging the poor and the rich – A field Study on the distributional effects of green electricity defaults. *Energ. Econ.* **86**, 104616 (2020).
15. N. Egami, E. Hartman, Elements of external validity: Framework, design, and analysis. *Am. Polit. Sci. Rev.* **117**, 1–19 (2022).
16. C. O. L. H. Porter, R. Outlaw, J. P. Gale, T. S. Cho, The use of online panel data in management research: A review and recommendations. *J. Manage* **45**, 319–344 (2019).
17. J. Chandler, D. Shapiro, Conducting clinical research using crowdsourced convenience samples. *Annu. Rev. Clin. Psycho.* **12**, 1–29 (2015).
18. K. J. Mullinix, T. J. Leeper, J. N. Druckman, J. Freese, The generalizability of survey experiments*. *J. Exp. Political Sci.* **2**, 109–138 (2015).
19. A. Krefeld-Schwalb, B. Scheibehenne, Tighter nets for smaller fishes? Mapping the development of statistical practices in consumer research between 2008 and 2020. *Mark Lett.* **34**, 351–365 (2023).
20. R. Baker *et al.*, Research synthesis: AAPOR report on online panels. *Public Opin. Quart.* **74**, 711–781 (2010).
21. M. R. Ellefson, D. M. Oppenheimer, Is replication possible without fidelity? *Psychol. Methods* 28, 1446–1455 (2022), 10.1037/met0000473.
22. C. R. Ebersole *et al.*, Many Labs 3: Evaluating participant pool quality across the academic semester via replication. *J. Exp. Soc. Psychol.* **67**, 68–82 (2016).
23. E. Peer, D. Rothschild, A. Gordon, Z. Evernden, E. Damer, Data quality of platforms and panels for online behavioral research. *Behav. Res. Methods* **54**, 1643–1662 (2022).
24. A. Coppock, T. J. Leeper, K. J. Mullinix, Generalizability of heterogeneous treatment effect estimates across samples. *Proc. Natl. Acad. Sci. U.S.A.* **115**, 12441–12446 (2018).
25. A. Coppock, O. A. McClellan, Validating the demographic, political, psychological, and experimental results obtained from a new source of online survey respondents. *Res. Politics* **6**, 2053168018822174 (2019).
26. E. Snowberg, L. Yariv, Testing the waters: Behavior across participant pools. *Am. Econ. Rev.* **111**, 687–719 (2021).
27. J. Chandler, G. Paolacci, E. Peer, P. Mueller, K. A. Ratliff, Using nonnaive participants can reduce effect sizes. *Psychol. Sci.* **26**, 1131–1139 (2015).
28. E. Peer *et al.*, Nudge me right: Personalizing online security nudges to people's decision-making styles. *Comput. Hum. Behav.* **109**, 106347 (2020).
29. E. Peters, Beyond comprehension. *Curr. Dir. Psychol. Sci.* **21**, 31–35 (2012).
30. E. Peters *et al.*, Numeracy and decision making. *Psychol. Sci.* **17**, 407–413 (2005).
31. E. T. Cokely, M. Galesic, E. Schulz, S. Ghazal, R. Garcia-Retamero, Measuring risk literacy: The Berlin numeracy test. *Judgm. Decis. Making* **7**, 25–47 (2012).
32. S. Frederick, Cognitive reflection and decision making. *J. Econ. Perspect.* **19**, 25–42 (2005).
33. K. S. Thomson, D. M. Oppenheimer, Investigating an alternate form of the cognitive reflection test. *Judgm. Decis. Mak.* **11**, 99–113 (2016).
34. D. M. Condon, W. Revelle, The international cognitive ability resource: Development and initial validation of a public-domain measure. *Intelligence* **43**, 52–64 (2014).
35. J. L. Horn, R. B. Cattell, Age differences in fluid and crystallized intelligence. *Acta Psychol.* **26**, 107–129 (1967).
36. R. B. Cattell, Theory of fluid and crystallized intelligence: A critical experiment. *J. Educ. Psychol.* **54**, 1–22 (1963).
37. M. Lövdén, L. Fratiglioni, M. M. Glymour, U. Lindenberger, E. M. Tucker-Drob, Education and cognitive functioning across the life span. *Psychol. Sci. Public Interes.* **21**, 6–41 (2020).
38. L. Zaval, Y. Li, E. J. Johnson, E. U. Weber, Aging and decision making. Sect 2. *Behav. Mech.* 149–168 (2015).
39. G. Paolacci, J. Chandler, P. G. Ipeirotis, Running experiments on amazon mechanical turk. *Judgm Decision Making* **5**, 411–419 (2010).
40. J. K. Goodman, C. E. Cryder, A. Cheema, Data collection in a flat world: The strengths and weaknesses of mechanical turk samples. *J. Behav. Decis. Making* **26**, 213–224 (2013).
41. S. Clifford, J. Jerit, Is there a cost to convenience? An experimental comparison of data quality in laboratory and online studies. *J. Exp. Polit. Sci.* **1**, 120–131 (2014).
42. N. Stewart *et al.*, The average laboratory samples a population of 7,300 amazon mechanical turk workers. *Judgm. Decis. Making* **10**, 479–491 (2015).
43. D. G. Rand *et al.*, Social heuristics shape intuitive cooperation. *Nat. Commun.* **5**, 3677 (2014).
44. J. K. Goodman, G. Paolacci, Crowdsourcing consumer research. *J. Consum. Res.* **44**, 196–210 (2017).
45. C. R. Ebersole *et al.*, Many Labs 5: Testing pre-data-collection peer review as an intervention to increase replicability. *Adv. Methods Pract. Psychol. Sci.* **3**, 309–331 (2020).
46. J. M. Jachimowicz, S. Duncan, E. U. Weber, E. J. Johnson, When and why defaults influence decisions: A meta-analysis of default effects. *Behav. Public Policy* **3**, 159–186 (2019).
47. E. J. Johnson, D. Goldstein, Do defaults save lives? *Science* **302**, 1338–1339 (2003).
48. A. Tversky, D. Kahneman "Judgments of and by Representativeness" in *Judgment under Uncertainty: Heuristics and Biases*, D. Kahneman, P. Slovic, A. Tversky, Eds. (Cambridge University Press, 1981), pp. 84–98.
49. C. K. Hsee, Less is better: When low-value options are valued more highly than high-value options. *J. Behav. Decis. Making* **11**, 107–121 (1998).
50. D. M. Oppenheimer, T. Meyvis, N. Davidenko, Instructional manipulation checks: Detecting satisficing to increase statistical power. *J. Exp. Soc. Psychol.* **45**, 867–872 (2009).
51. M. Hauser, F. Cushman, L. Young, R.K.-X. Jin, J. Mikhail, A dissociation between moral judgments and justifications. *Mind Lang.* **9**, 522–550 (2011).
52. S. D. Gosling, P. J. Rentfrow, W. B. Swann, A very brief measure of the big-five personality domains. *J. Res. Pers.* **37**, 504–528 (2003).
53. A. A. Arechar, G. T. Kraft-Todd, D. G. Rand, Turking overtime: How participant characteristics and behavior vary over time and day on amazon mechanical turk. *J. Econ. Sci. Assoc.* **3**, 1–11 (2017).
54. A. Krefeld-Schwalb, E. J. Johnson, E. Sugerman, X. Hua, Data from "Many Panels". OSF. osf.io/7kqg9. Deposited 17 February 2024.