Exposing Omitted Moderators: Explaining Differences in Treatment Effects in the Social Sciences.

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

[a]Rotterdam School of Management, Erasmus University, [b]Columbia Business School, Columbia University

*Antonia Krefeld-Schwalb

Email: krefeldSchwalb@rsm.nl

Author Contributions: AKS, ERS, and EJJ planned and designed the studies. ERS supervised the data collection. AKS ran the analysis. AKS, ERS, and EJJ wrote the manuscript.

Competing Interest Statement: No conflict of interest.

Classification: Social Science; Psychological and Cognitive Science

Keywords: Online Data Collection; Choice architecture; Moderators

Abstract

Policymakers increasingly rely on behavioral science in response to global challenges, such as climate change or global health crises. But applying behavioral science faces an important problem. Interventions exert substantially different effects across contexts and individuals. We examine this heterogeneity for different paradigms that provide the basis for many behavioral interventions. We study the paradigms across one in-person and 10 online panels with over 11000 respondents. We propose a framework of typically omitted moderators to explain this heterogeneity. The framework's factors (Fluid Intelligence, Attentiveness, Crystallized Intelligence, and Experience) affect the effectiveness of many behavioral interventions and preference measures. The observed treatment effect depends on the distribution of moderators in each sample. Our results motivate observing these moderators and provide a theoretical and empirical framework for understanding varying effect sizes.

Introduction

Effective responses to global challenges often require individual behavior change. A successful vaccine only achieves full efficiency if enough people choose to be vaccinated. Improving public transport to reduce carbon emissions must also change commuting choices. Behavioral science can help to advance change with “behavioral interventions”, also referred to as 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, but others argue that applications are premature (2–4). One challenge to identifying successful interventions is that the magnitude 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 green energy. While they generally increase the use of green energy, default effects are stickier among people with lower SES (14). From a policy perspective, the result may be undesirable: The poorer, who are responsible for fewer emissions, pay relatively more than the rich. From a theoretical perspective, this finding is important because it informs researchers of settings where defaults may not work.

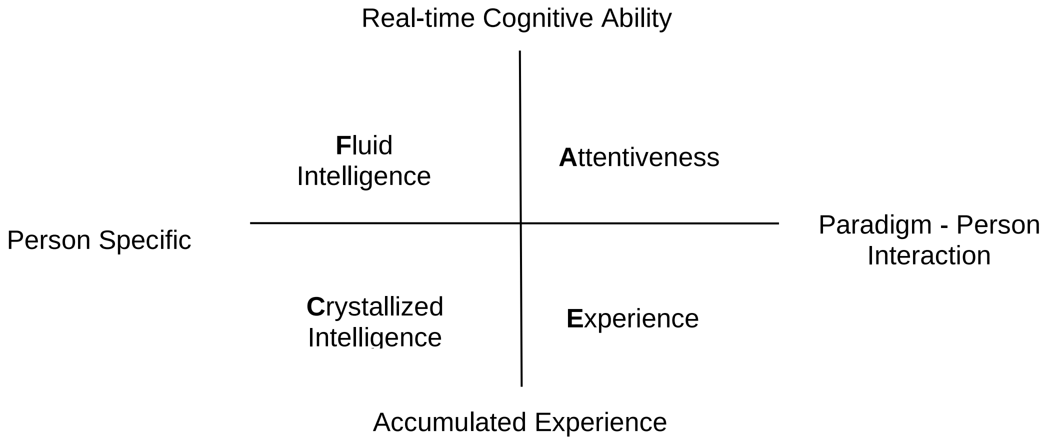
This paper (i) demonstrates the importance of heterogeneity in standard paradigms used to develop behavioral interventions, (ii) develops a guiding framework for explaining heterogeneity, and (iii) demonstrates techniques for improving the robustness of research. We build on the key insight that the effect of one variable, like a default manipulation, may depend upon a third variable (like SES) that varies, but is not observed in the original setting. To understand heterogeneity, it is necessary to measure and model this omitted moderator. Neglecting it may limit our ability to generalize results.

To do this, we leverage the variation that exists across different offline and online panels, ranging from widely used commercial panels to a student laboratory sample. We demonstrate this heterogeneity for many paradigms that are the basis for behavioral interventions. We then show how exposing omitted moderators can explain this heterogeneity and aid theory construction. While our primary focus is on studying heterogeneity in behavioral science as a whole, understanding panel differences is important. Online panels have become increasingly common across the social sciences, including psychology, economics, sociology, and political science (15–19).

Imagine that an identical manipulation, like a default, has a large effect in one panel and an insignificant effect in another. We may ask, are there characteristics of the panel that explain this discrepancy? Online panels may differ substantially in their distribution of moderators. To exploit this variation, we use different panels and measure typically omitted moderators. Performing identical paradigms across panels minimizes variability in study execution to reveal differences in moderators (20). Our data can be regarded as a complementary extension to large, coordinated replication studies (9, 10, 21). However, instead of limiting our efforts to one panel, we exploit panel differences to produce purposive variation (22), making moderators and the resulting heterogeneity more easily observable.

Studies investigating demographic differences between online (23, 24) and other panels (25, 26) have led to mixed results. Some are consistent with small differences and little heterogeneity of effect sizes, and other research has identified differences beyond demographics that are more strongly associated with effect sizes (23, 27, 28). Paradigms in the behavioral sciences have a large set of possible moderators, many of which may be specific to the paradigm. We focus on a set of fundamental moderators that are relevant across many paradigms and contexts. Measuring possible moderators is often burdensome, particularly in an online setting. To accommodate this, we propose measures that are reliable and brief, and we exploit measures such as reaction time that may be implicitly collected. These measures provide a starting point that we later complement with more sophisticated measures.

Two widely discussed cognitive constructs are likely to affect how individuals encode, process, and integrate information in the paradigms: *Fluid* and *Crystallized* Intelligence. We also examine measures that are specific to the respondents' interaction with the paradigm: The *Attentiveness* of respondents to a particular paradigm and their past *Experience* with that intervention. We summarize this framework using the acronym *FACE,* representing: *Fluid Intelligence, Attentiveness, Crystallized Intelligence, and Experience.* As illustrated below, the four factors differ on two dimensions: 1) whether they are a function of the person, or the interaction of the paradigm and the person and 2) whether they represent real-time cognitive abilities or whether they are 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 such an intervention might be moderated by Fluid Intelligence in general, and by numeric skills in particular. For example, Peters et al. show that numeracy affects framing and risk attitudes (30). To measure Fluid Intelligence, we initially use related surveys such as the Berlin Numeracy Test (31) and the Cognitive Reflection Task and its variant (32, 33), and later moved to more recently developed items that have been more extensively tested as measures of fluid intelligence, including Raven-like matrices and 3-D rotation tasks (34) (see the Supporting Information 3 and 4).

In contrast, Crystallized Intelligence focuses on knowledge of the world and is thought to be the result of accumulated experience. Accordingly, it has been found to increase with age, at least until ones' 60s (35, 36). The effect of interventions requiring comprehension of text or knowledge of the world might be increased by Crystallized Intelligence. Measures of Crystallized Intelligence have long played a separate role in understanding cognitive performance and can compensate for decreases in Fluid Intelligence (35, 36). A common measure of Crystallized Intelligence is a person's passive and active vocabulary, which we adopt in line with Li et al. In other cases, we use one's age and highest level of education as a proxy for Crystallized Intelligence, two measures that have been shown to correlate with Crystallized Intelligence (38). 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 (37).

Attentiveness of respondents has been an enduring concern in research, both in-person and online, when respondents are not directly monitored. Respondents not paying attention to tasks can reduce the size of measured effects (23, 39–41). In survey-based experiments, attention checks are commonly used to filter for attentive participants. 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, easily measured in standard survey packages.

Finally, *Experience* (i.e. Naivete (27)), refers to familiarity with the intervention and is a likely moderator of behavioral interventions. Some panels are composed of a limited number of participants who spend a significant amount of time doing experimental tasks and thus may have seen similar or identical paradigms on multiple occasions. MTurk, for example, is thought to have only 7300 active panelists, who report doing over 300 experimental tasks (42). There is evidence that this exposure can reduce the size of observed treatments (27, 43, 44). Familiarity with information has several influences on the processing of the information. The effects of repeated questions may lead to simple recall of past answers, the use of simplified heuristics, or a lack of attention. Such concerns have led to the development of alternative forms of tests such as the CRT (33). We measured Experience by asking respondents directly, for each paradigm, whether they had seen these questions before, on a three-point scale.

In all studies, we verify the FACE framework by using a multigroup factor analysis, assuming strong measurement invariance and allowing factors to correlate (see Supporting Information 2.3). Of course, the FACE framework is just an initial proposal to explain unobserved heterogeneity using variables that have been somewhat neglected. Other variables such as demographics and personality may well affect outcomes and treatments. While we believe that the FACE framework is a good, domain-general starting point for identifying variables that have been previously neglected, we will examine the effect of other potential moderators, such as demographics and the Big 5, using the TIPI scale.

We first demonstrate the effect of heterogeneity by documenting panel differences for several standard paradigms. This can be tested by looking for differences in the effect size as a function of the panel. Next, we try to account for the differences across panels using the FACE factors. Our focus will be asking if the treatments are more or less effective, as the FACE factor scores change. To give a concrete statistical example, we ask if the framing effect interacts with Attentiveness or Crystallized Intelligence. Finally, we ask if these interactions reduce or eliminate the panel differences. Several studies address the replicability of the observed panel differences and extend our analysis to preference elicitation.

We conducted 5 preregistered studies with over 11000 respondents using 10 online panels and one laboratory student panel. Each study employed several well-studied paradigms from the literature that have either been the subject of a large replication project (9, 10, 21, 45) or been examined in a meta-analysis (46, 47). This provides us with a data-driven effect size standard for comparison. These paradigms could be implemented online, were not time-intensive, and spanned a range of expected effect sizes. All panels were presented with identical questionnaires. We included panels commonly used in academic online experiments (MTurk, CloudResearch, and Prolific Academic) and panels primarily used in market research. We also included two panels labeled by Prolific as nationally representative of the US and UK.

In Study 1, we examined how the effect sizes of five paradigms changed across 11 panels (see Supporting Information 1). We label these Local Warming (48), Defaults (49), Framing (50), Less-is-better (51), and Sunk Cost (52). Study 2 and Study 4 replicated the results with improved measures of the FACE factors. In Study 3, we extended our results to two paradigms from the broader social science literature, False Consensus (53) and the Trolley Problem (54)) as well as self-report and choice measures for three economic preferences, Time, Risk, and Reciprocity (55)). In Study 5 we further improved our understanding of heterogeneity with additional measures.

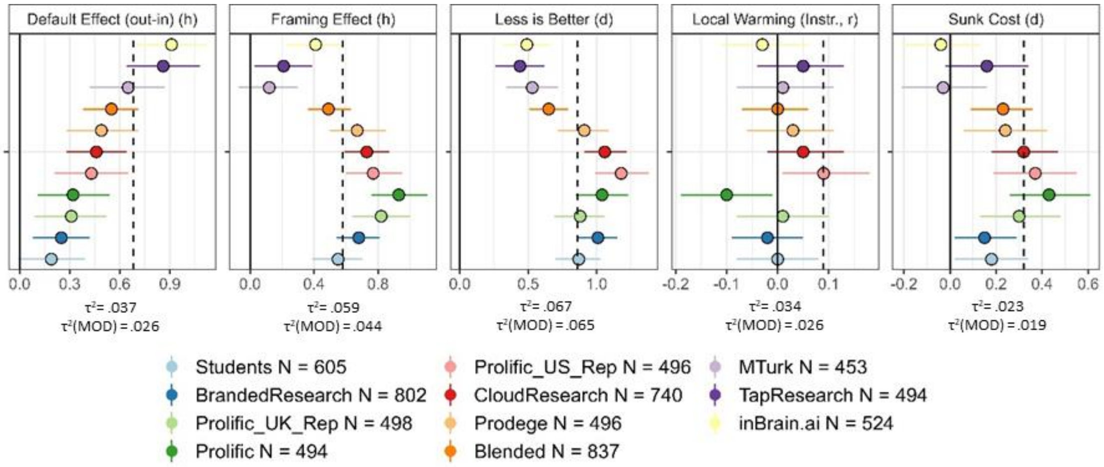
Results

Study 1 produced substantial heterogeneity in effect sizes of the five paradigms across 11 panels and 6438 participants, depicted in Figure 1. 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 treatment, panel, and their interaction (see Supporting Information 2.2). Equation 1 illustrates this analysis:

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

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

Fig. 1. 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 statistics below each plot summarize the heterogeneity of conditional treatment effects without (τ2) and with (τ2(MOD)), moderators in the regression models.



The analysis confirms what we see in Figure 1: 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 Figure 1 by the different distribution of the default and framing effects across panels. Focusing on two of the most commonly 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 heterogeneity in effect sizes results from an interaction of the paradigm and the panel. The analysis was robust when respondents who failed one or two attention check questions were omitted (64% passed both checks; see Supporting Information 2.2). The only paradigm that did not show heterogeneity across panels was local warming, which also failed to replicate when averaged across all panels.[*]

[*] While local warming has been shown to be a significant but small effect, the size of the effect depends upon the variability in temperature at the time and location in the study. When this study was run such variability was low. To remove this source of heterogeneity, we dropped this paradigm in subsequent studies.

We argue that this heterogeneity in effects is due in part to differences in previously unobserved moderators across panels. To test this, we conducted an exploratory factor analysis with oblimin factor rotation of our measured moderators. This produced the four factors of the **FACE** framework (RMSEA = 0.03, CFI = 0.99), suggesting the factors **F**luid Intelligence (cognitive reflection task and numeracy), **A**ttentiveness (response times in the paradigms), **C**rystallized Intelligence (education), and **E**xperience (with the individual paradigms). The largest correlation between factors was *r* = -0.368. See Supporting Information 2.3 for more information on the factor analyses.

Fig. 2. 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 light gray area illustrates the average scores across panels. The inner border is the minimum and the outer line is the maximum average score on the respective FACE factors across panels.

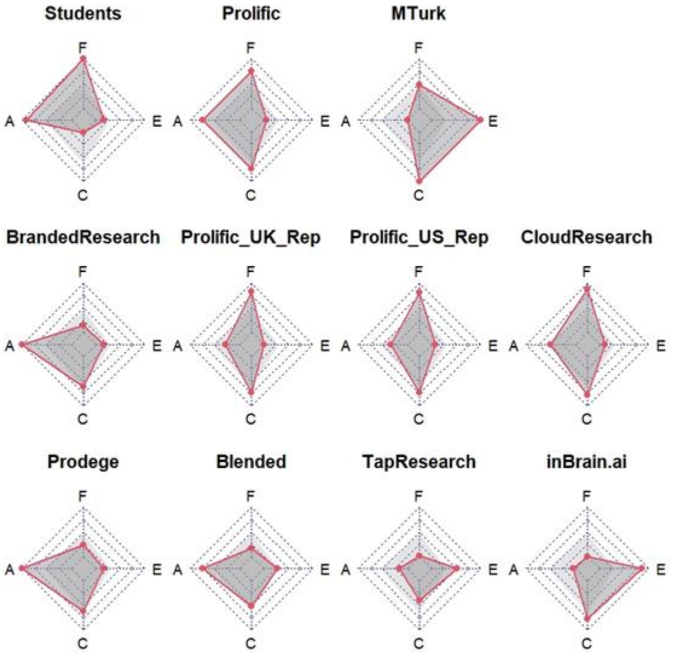


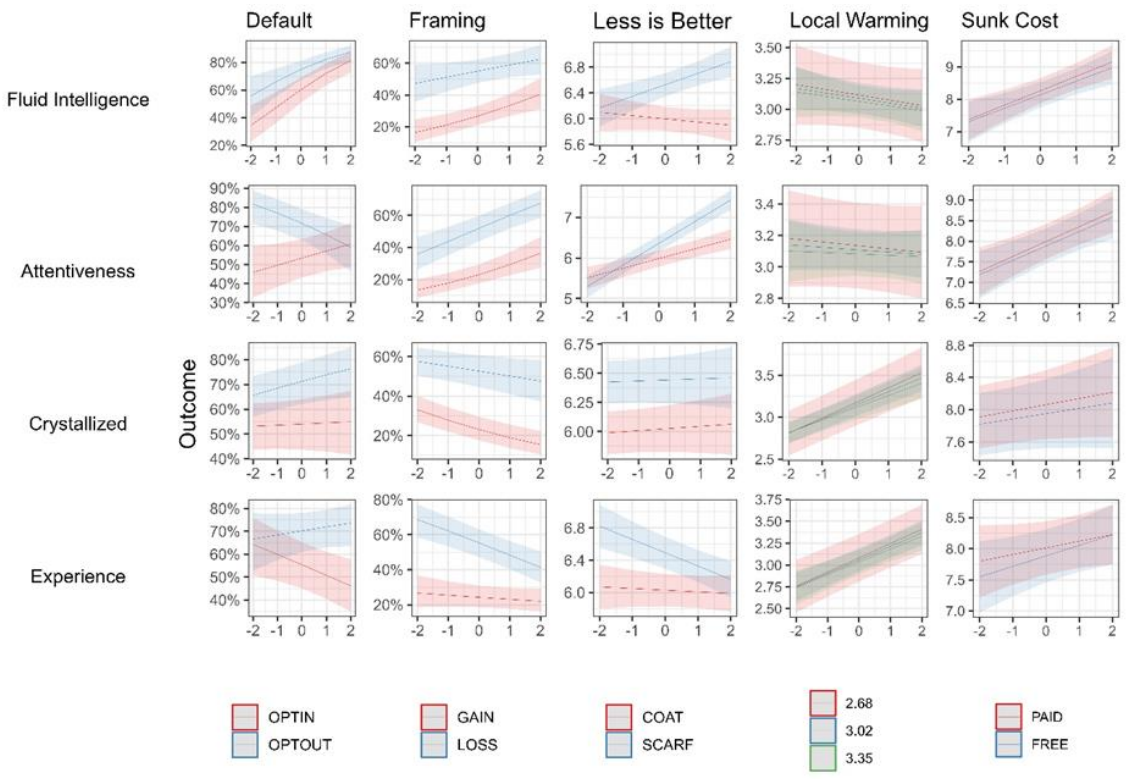
Figure 2 illustrates the large differences in panel characteristics. Each radar plot presents a panels' average factor scores as red lines. The light gray areas represent the average score across all panels in Study 1. For example, the first row depicts the profile of three commonly used panels (student, MTurk, and Prolific). We see that MTurk participants are low in Attentiveness (z = -0.23) and high in Experience (1.33). In contrast, Prolific participants 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 school, have low in Crystallized Intelligence (-1.37). We next assessed moderation by adding the four moderators' main effects and interactions with the treatments to the previous regression models, schematically illustrated in Equation 2.

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

F, A, C, E = factor scores.

Adding moderators to the regression models reduced the panel main effects by 58% and the interaction effects with the treatment by 29%, on average; and heterogeneity of treatment effects (τ^2) by 18%, on average. Adding the moderators further increased model fit across all paradigms, accounting for model complexity (see Supporting Information 2.4). Figure 3 illustrates the models, with rows and columns representing moderators and paradigms, respectively. Looking at the leftmost cell of the second row, the relationship between Attentiveness and the default effect, the two non-parallel lines indicate that default effects grow smaller with higher Attentiveness. In contrast, the last cell in that column shows that default effects increase in size with Experience. Experienced (vs. naive) respondents show larger default effects. Such interactions are clear demonstrations that Attentiveness and Experience are moderators of the default effect. The graphs also show main effects of treatment, the distance between the lines, and the main effect of the moderator on the outcome, shown by the slopes.

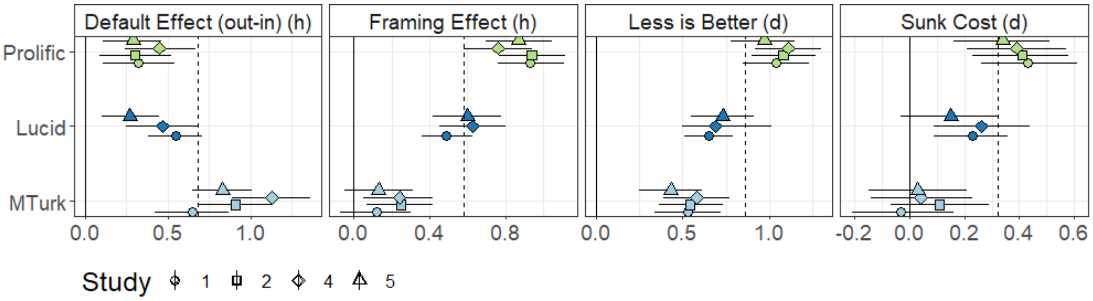
Fig. 3. Effects of moderators. Each column represents the predictions of one paradigm (on the y-axis) as a function of FACE moderators (on the x-axis). In the local warming paradigm, the IV (concern about climate change) was split into three equal intervals to illustrate an eventual interaction.



We focus on differences in the framing and default effects to illustrate the importance of these moderators. Both paradigms show a striking negative correlation of effect sizes across panels in Figure 1. 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 complex, and a function of both task and panel characteristics.

To test the reliability and the robustness of Study 1's results, we first ran Study 2 (N = 1057), which replicated the reversal pattern of results for the default and framing paradigms for the MTurk and Prolific panels (see Supporting Information 1.3). To illustrate the stability of our findings, Figure 4 plots the effect sizes in the four experimental paradigms and panels that we used repeatedly across studies. Note that respondents could participate in the research only once (e.g., respondents in Study 1 could not participate in Studies 2, 3 or 4).

Fig. 4. Effect sizes in Defaults, 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. The dashed vertical lines represent the benchmark effect sizes from meta-analyses or a Many Labs replication study.



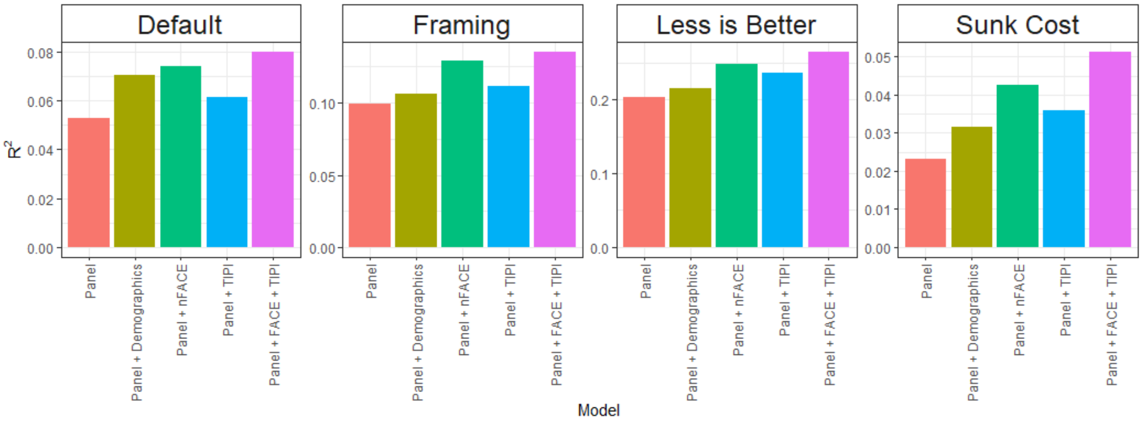
Next, Study 3 extended the analysis of heterogeneity to two paradigms from social psychology, the trolley problem (54) and the false consensus effect (53), as well as measures central to economics—preferences for risk, time and reciprocity. The two panels used, Prolific and MTurk, showed significant differences in preferences, the correlation between choice and self-reported preferences, and in the trolley problem. As in other studies, the relationship of choice-based and self-report measures of preferences is surprisingly weak and varied between panels (Prolific: r_{risk} = .223, r_{time} =.191; MTurk: r_{risk} = .173, r_{time} =.030). Despite new measures of Fluid Intelligence, we reproduced the FACE framework and explained panel differences by introducing the improved FACE factors as moderators (see Supporting Information 3 and Table S16). After adding FACE moderators, we observed no significant difference in the correlation of preference measures between panels. Across all paradigms in Study 3, the FACE framework accounted for 63% of the variance in panel effects.

In Study 4 (N = 1460), we improved our measure of Crystallized Intelligence for the FACE framework and replicated the same analysis on the between-subjects paradigms of Study

1. We also added TIPI scales for measuring personality traits as alternative moderators. After adding FACE moderators, panel main effects and interactions were reduced by 68%. Moreover, FACE factors helped to explain on average 30% more variance compared to demographic variables (8%) and TIPI (21%).

Finally, in Study 5 (N = 1460) we used both the improved measure of Crystallized and Fluid Intelligence. We again replicated the FACE factor structure and found that the FACE factors explain more heterogeneity than TIPI or demographic variables. Figure 5 shows the fit of the models. Adding FACE moderators reduced panel main effects and interactions by 80%.

Fig. 5. Explained variance of the regression models predicting the outcome variables in Study 5 with different moderators. Each of the models contains main effects of the treatment, the panel, and interaction terms of the panel and the treatment. All moderator variables are added with main effects and interactions with the treatment. As demographic variables, we added the gender, age, and education of the respondent. TIPI contains the personality traits extraversion, agreeableness, openness, conscientiousness, and neuroticism, and FACE contains the four factors Fluid Intelligence, Attentiveness, Crystallized Intelligence and Experience.



Discussion

We demonstrate that basic phenomena from the decision, social, and economic sciences vary more widely across samples than previously documented (18, 25, 26). The amount of heterogeneity we observe suggests that we generally cannot (or do not) study treatment effects. Instead, we study *conditional* treatment effects, conditional on the sample selected. Although the observed pattern of heterogeneity is complex, and more complex than is commonly assumed, it is systematic. We propose a framework to understand the complex pattern. The FACE framework serves as a starting point to help social and behavioral scientists develop more robust theory and more confidently apply behavioral science in policy.

A researcher might be concerned that participants pay attention to stimuli and respond by adding attention checks to eliminate respondents. We think this solved the wrong problem for two reasons: First, our results suggest that there is more to heterogeneity than attention. Experience with the paradigm and the Fluid and Crystallized strengths of the respondent are also important. Second, these moderators do not just affect how strongly the manipulation affects the respondent, but also fundamentally change their response.

To explore this, we introduced a measure in Study 5 of how well the respondents understood the manipulation by asking them to identify what they had read (see Supporting Information). While the FACE variables affect the strength or Dosage of the treatment, they also affect the effect of the manipulation on the outcome. About 25% of the average explained variance in the outcome was explained by this measure. As one would expect, respondents with higher Crystallized and Fluid Intelligence and those who paid more attention seemed better able to identify their treatment, while greater Experience reduced identification. However, that is not the whole story. There were still significant effects of the FACE factors and their interactions on the outcomes, explaining about 26% of the average explained variance in the outcome. This suggests that the heterogeneous effects seen in Figure 1 are not just due to how strongly the intervention affects respondents. An important question for future research is understanding how variables like FACE affect these and a broader set of paradigms.

The FACE framework 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. It would be great to observe that FACE factors could explain why some paradigms are more robustly replicated in some samples, but not in others.

Our approach bears similarity to past attempts at generalizing from an imperfect sample, such as multilevel regression with poststratification (56). Poststratification is typically based on observed, true, and stable values; while the true distributions of the moderator variables in the FACE framework are unknown and can vary over time (e.g., Attentiveness and Experience). Our results suggest that considering these FACE factors may be at least as relevant.

We make two methodological contributions that could be applied in future research. To advance the investigation of moderators, we propose a technique to create purposive variation by exploiting the existing variety of online panels. We encourage researchers to employ a similar approach, while measuring FACE factors. Future research could further improve the proposed measures and explore other important moderators such as need for cognition or political affiliation particularly when generalizing to other tasks. The FACE framework helps to evaluate how effects vary across different parts of society and affect implementations of interventions such as defaults. When applied widely, the framework offers the potential to develop a theory on the structure of paradigms and their interaction with FACE.

Methods and Materials

In Study 1a, we tested five behavioral interventions on eight online panels (detailed information can be found in Supporting Information 1.1). These included one regression with an instrumental variable (local warming (48)) and four between-subjects experimental paradigms, including the default effect (organ donation (49)), the framing effect (unusual disease (50)), the less-is-better effect (51), and the sunk cost effect (52). 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). To include in-person data collection, we used a large sample of students from a European university in Study 1b and added samples often termed representative by providers, from the US and UK, stratified on age, race, and gender, from Prolific. 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. Here, we focused on two panels most used in academic research, Prolific and MTurk, since they showed quite different patterns of moderation 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 (57). In Study 4 we used only the between-subjects paradigms of Study 1 and added vocabulary tests to improve the measures of Crystallized Intelligence. We again distributed the survey across four data collection periods, and sampled participants from Prolific, MTurk and Lucid Marketplace. Study 5 was distributed in the same manner as Study 4. We used the 3D rotations 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, which of the two treatments for each paradigm they saw.

We report our analyses using regressions and factors based on a confirmatory factor analysis, but also examined the effects of moderators using seemingly unrelated regressions (when appropriate) and structural equation modeling in Study 1, and conducted exploratory factor analysis to identify the FACE factors in each study (sections 2.5 and 2.6 Supporting Information). The effects we report are robust across procedures. All data and code is available online (https://osf.io/7kqg9/).

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.

References

1. G. M. Walton, A. J. Crum, *Handbook of Wise Interventions* (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*, 1–60 (2022).

4. R. H. Thaler, C. R. Sunstein, *Nudge: The Final Edition* (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. *Social Psychology* 45, 142–152 (2014).

10. R. A. Klein, et al., Many labs 2: Investigating variation in replicability across samples and settings. *Advances in Methods and Practices in Psychological Science* 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. 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).

16. J. Chandler, D. Shapiro, Conducting Clinical Research Using Crowdsourced Convenience Samples. *Annu Rev Clin Psycho* 12, 1–29 (2015).

17. K. J. Mullinix, T. J. Leeper, J. N. Druckman, J. Freese, The Generalizability of Survey Experiments*. *J Exp Political Sci* 2, 109–138 (2015).

18. A. Krefeld-Schwalb, T. Pachur, B. Scheibehenne, Structural Parameter Interdependencies in Computational Models of Cognition. *Psychol Rev* 129, 313–339 (2022).

19. R. Baker, et al., Research Synthesis: AAPOR Report on Online Panels. *Public Opin Quart* 74, 711–781 (2010).

20. M. R. Ellefson, D. M. Oppenheimer, Is Replication Possible Without Fidelity? Psychol Methods (2022) https:/doi.org/10.1037/met0000473.

21. 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).

22. N. Egami, E. Hartman, Elements of External Validity: Framework, Design, and Analysis. *Am Polit Sci Rev*, 1–19 (2022).

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 National Acad Sci* 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. *Judgement and Decision 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. L. Zaval, Y. Li, E. J. Johnson, E. U. Weber, Aging and Decision Making. *Sect. 2: Behav. Mech.*, 149–168 (2015).

38. 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).

39. G. Paolacci, J. Chandler, P. G. Ipeirotis, Running experiments on Amazon Mechanical Turk. *Judgment and 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 Political Sci 1, 120–131 (2014).

42. N. Stewart, et al., The average laboratory samples a population of 7,300 Amazon Mechanical Turk workers. Judgment and Decision Making 10, 479–491 (2015).

43. D. G. Rand, et al., Social heuristics shape intuitive cooperation. *Nature Communications* 5 (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 Psychological Sci* 3, 309–331 (2020).

46. E. R. Sugerman, Y. Li, E. J. Johnson, Local warming is real: A meta-analysis of the effect of recent temperature on climate change beliefs. *Curr Opin Behav Sci* 42, 121–126 (2021).

47. J. M. Jachimowicz S. Duncan, E. U. Weber, E. J. Johnson, When and why defaults influence decisions: a meta-analysis of default effects. *Behavioural Public Policy* 3, 159–186 (2019).

48. Y. Li, E. J. Johnson, L. Zaval, Local Warming. Psychol Sci 22, 454–459 (2010).

49. E. J. Johnson, D. Goldstein, Do Defaults Save Lives? Science 302, 1338–1339 (2003).

50. A. Tversky, D. Kahneman, “Judgments of and by Representativeness” (1981).

51. 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).

52. D. M. Oppenheimer, T. Meyvis, N. Davidenko, Instructional manipulation checks: Detecting satisficing to increase statistical power. J Exp Soc Psychol 45, 867–872 (2009).

53. L. Ross, D. Greene, P. House, The “false consensus effect”: An egocentric bias in social perception and attribution processes. J Exp Soc Psychol 13, 279–301 (1977).

54. M. Hauser, F. Cushman, L. Young, R. K.-X. Jin, J. Mikhail, A Dissociation Between Moral Judgments and Justifications. *Mind Lang* 22, 1–21 (2007).

55. T. Dohmen, et al., Individual Risk Attitudes: Measurement, Determinants, And Behavioral Consequences. *J Eur Econ Assoc* 9, 522–550 (2011).

56. A. Gelman, J. Lax, J. Phillips, J. Gabry, R. Trangucci, Using Multilevel Regression and Poststratification to Estimate Dynamic Public Opinion (2018).

57. 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).