Empirical Research Paper

Recruitment Strategies Bias Sampling and Shape Replicability

Personality and Social Psychology Bulletin
1–19

DOI: 10.1177/01461672241293504
journals.sagepub.com/home/pspb


Thomas I. Vaughan-Johnston[1], Faizan Imtiaz[2], Gabriella Avila Patro[3], Samantha Xiao Shang[3], Leandre Fabrigar[3], and Li-Jun Ji[3]

Abstract

Replicating psychological research has become a central concern for psychologists. Although attention has been paid to the possibility of heterogeneous populations driving replication success/failure, the heterogeneous recruitment strategies researchers use to draw samples from those populations are often overlooked. Yet recruitment strategies may bias the participants who show up and shape replication results. We examine this idea through several unique paradigms (sampling North American university students, N_{total} = 1,009). First, subtle manipulations of recruitment strategies (i.e., mentioning cash, expedient credit, fun, or a study narrative) were differentially appealing to individuals varying on experiential versus reward-based motivations (Experiment 1). Second, employing different recruitment strategies biased the motivational styles of actual participant show-ups, and sometimes even shaped the success of several replication studies (Experiment 2–3). We conclude that recruitment strategies may sometimes alter the degree of successful replication.

Keywords

materialism, metascience, recruitment, replication

Received April 4, 2022; revision accepted September 2, 2024

Replication plays an important role in psychology. Recent replication failures have drawn psychologists' attention to research design and replicability concerns (e.g., Ebersole et al., 2016; Klein et al., 2018). Clearly, many research design choices contribute to replication results (see Fabrigar et al., 2020), including operationalizations (Fabrigar & Wegener, 2016; Flake et al., 2022; Schwarz & Clore, 2016), the presence of manipulation checks (Hauser & Schwarz, 2015), and population selection (e.g., Brandt et al., 2014; Klein et al., 2018). However, an understudied yet potentially important component of research design is recruitment strategies, our focus in the present work.

Although original researchers often detail the use of incentives (e.g., study credit and cash), they rarely disclose how research was "pitched" to potential participants via recruitment strategies. For instance, were rewards such as monetary compensation or course credit emphasized in recruitment notices? Was the study characterized as fun? Interesting? Was the study's purpose or even hypothesis disclosed? Seminal handbooks of psychological research design often give only cursory attention to recruitment strategies (e.g., Reis & Judd, 2014). Yet inconsistent recruitment strategies used across experiments may create a subtle (and typically unreported) source of variance: both within data sets collected by a particular laboratory, and between different laboratories that use distinct recruitment tactics.

Strikingly, replication and original literature generally overlook recruitment. Although it is periodically acknowledged that standardizing recruitment strategies would be part of an ideal "replication recipe" (e.g., Brandt et al., 2014) or that equality of original/replication recruitment methods is an assumption that must be held to expect successful replication (Steiner et al., 2019), in practice this is given minimal attention. Submission guidelines at major psychological journals usually target only a narrow range of methodological considerations (Fabrigar et al., 2019), and in our review of submission guidelines, we found that almost no "method reporting" checklists mention recruitment materials. Examining replication articles specifically, the Many Labs 2 project (ML2; Klein et al., 2018) standardized a protocol for running participants but simply required sites "to collect data from at least 80 participants" without standardizing recruitment techniques (ML2 Coordinating Proposal; https://osf.io/uazdm/). Furthermore, in replication efforts where multiple

[1]Cardiff University, UK
[2]Towson University, MD, USA
[3]Queen's University, Kingston, Ontario, Canada

Corresponding Author:
Thomas I. Vaughan-Johnston, School of Psychology, Cardiff University, Cardiff CF10 3AT, UK.
Email: thomasvaughanjohnston@gmail.com

studies are replicated in a single session, it may be impossible to match the recruitment methods of all original studies. For instance, ML2 included a study that originally had compensated in-lab participants with cash (Miyamoto & Kitayama, 2002, Study 1), and another study where participants were in a park and not compensated (Knobe, 2003); it seems impossible that ML2 could have matched both studies' recruitment strategies. This pattern is typical, suggesting that replication studies may often mismatch the recruitment strategies that were used in the original research. We are not "blaming" replication researchers. When original researchers have not disclosed their recruitment methods (and they seldom do), replicators have no mechanism to match their recruitment strategies to the original study. We are also not "blaming" original researchers—sharing recruitment materials is not normative—but we propose that perhaps it should become normative. In the present work, we track the potential consequences of inconsistencies.

Recruitment Strategies and Biased Sampling

Recruitment strategies require attention because (a) recruitment strategies might be differentially appealing to people varying on multiple individual differences, (b) this differential appeal might lead to actual differences in joining studies, and (c) those selection biases may contribute to differences between the results of original and replication research. Recruitment strategies are often designed as social influence attempts: They persuade people to engage in research,[1] yet it has long been acknowledged that different people find different social influence messages to be persuasive (e.g., Teeny et al., 2021). This suggests one mechanism by which recruitment strategy differences could lead original and replication results to differ: Different recruitment strategies could lead to distinct biases in the types of people who are sufficiently convinced to show up and participate. The resulting bias would be problematic if those individual differences covertly influence a study's effects, particularly if researchers are unaware this is happening. For example, Carnahan and McFarland (2007) found that when using materials from the Stanford Prison Experiment (SPE), including versus excluding references to "prison" in their sampling materials led to increased aggressiveness, authoritarianism, Machiavellianism, and narcissism among show-ups; that is, the very people likely predisposed to antisocial behaviors. How might this affect a hypothetical replication study of SPE that simply omitted "prison" references from its recruitment materials?

We believe that this issue has implications for a far broader range of situations than the SPE. Indeed, we suspect that numerous individual difference variables may become biased due to what recruitment materials are circulated, leading to potential consequences for intra-lab and inter-lab replication. For the present experiments, we focused on one specific type of individual difference variable. We used three principles to select which dimensions of individual difference to scrutinize, which led us to settle on extrinsic/reward-focused and intrinsic/experiential motivations (Amabile et al., 1994). First, we considered how recruitment strategies are likely to be deployed in the research world, aiming for ecological validity. We considered that psychologists' recruitment strategies are very likely to draw from psychological theory and appeal to fundamental motivation frameworks, a central example of which is the reward/experience motivation distinction. Second, we considered the recruitment strategies deployed in our own university participant pool (see Note 1) and considered what motivations were commonly appealed to: These could typically be considered reward-focused (e.g., emphasizing credit/money) and/or experiential (e.g., the study being fun/interesting). Third, we wanted to select individual differences that would have substantial implications if common recruitment strategies indeed shape their distribution among show-ups. Broad distinctions between reward-based and experiential-based motivation are central to areas of psychology including judgment/decision-making, education, and consumer behavior (e.g., Cerasoli et al., 2014; Deci et al., 1999; Hidi & Harackiewicz, 2000). Thus, if recruitment strategies shift the distribution of these traits showing up in participants, this has broad implications for psychology.

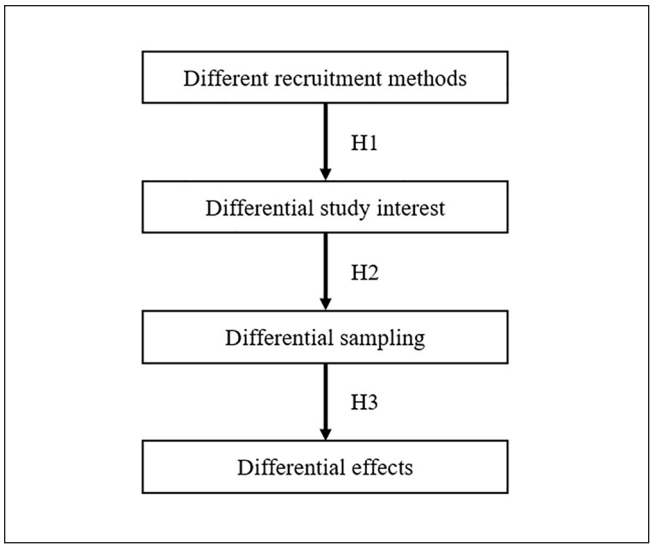

Figure 1. Conceptual Model of How Recruitment Methods May Influence Study Effects.

Overview of the Present Research

We report three experiments that targeted several critical objectives in this program of work, as depicted in Figure 1. Experiment 1 tests the first hypothesis, which is:

> **Hypothesis 1:** Different recruitment materials can appeal to various individual differences.

Experiments 2 and 3 test the next two steps in the conceptual model. First, because different recruitment materials are appealing to different people, this can produce changes in the sort of people who will show up to participate in research. This is a bolder hypothesis than hypothesis 1 because many factors will influence participants' tendency to sign up for studies other than the degree of personality match implied by recruitment strategies, including the convenience of the timing and location of the study, baseline interest in completing the study, topic area of the study (e.g., a cognitive versus social study), and many other considerations. Nonetheless, we predict:

> **Hypothesis 2:** Different recruitment materials can produce samples that are different on various individual difference variables.

Finally, Experiments 2 and 3 also test whether the outcome of psychology studies sometimes changes in the conditions where this bias is created. This is a bolder claim again than Hypothesis 2, upon which Hypothesis 3 relies. First, the variables affected by the recruitment strategy have to actually moderate the particular psychological effect targeted in the research. Second, recruitment strategies could conceivably introduce multiple biases to the characteristics of participants who show up, and so we will only detect moderation if the preponderance of these variables bias the effect in a uniform direction. Nonetheless, Experiments 2 and 3 each test whether:

> **Hypothesis 3:** Different recruitment materials can alter the effects tested in psychological research.

In all experiments, we focused on reward-focused versus experience-focused motivation. Thus, we showed both how recruitment strategies can appeal differently to people who are higher in reward-focused and/or experience-focused motivation (testing Hypothesis 1), draw in samples of participants biased concerning these traits (testing Hypothesis 2), and also how this biasing process can have consequences for the degree to which psychological effects will replicate (testing Hypothesis 3).

Open Practices Statement

The anonymized data and code for all studies can be found at: https://osf.io/xz5an/. Experiment 3 was preregistered at https://osf.io/g4wq6. The complete materials including recruitment documents are included in the Supplemental Material associated with this article (SOM-1 and SOM-2).

Experiment 1

We wanted first to establish if there is concrete reason to be concerned that subtle alterations in recruitment strategy wording can make studies more appealing to people based on their underlying motivations (Hypothesis 1). In Experiment 1, we tested whether different recruitment strategies or strategies would have different appeals to potential participants depending on their reward/experiential-focused motivation.

Method

Participants. In all, 258 undergraduate psychology students participated in the study in the laboratory for partial course credit, but 15 participants had missing data and thus were excluded. The final sample included 243 participants (80% women, $M_{age} = 18.6$). We used a time-based stopping rule. A post hoc power analysis based on G*Power's analysis of variance (ANOVA; repeated measures, within-between interaction) suggested we had 80% power to detect effect sizes of $r >= .23$ to .27 depending on average inter-measure correlation (we tested $r_{ave} = .10$ to .50). We targeted Western undergraduates throughout the studies because they represent a population predominantly studied in psychology, adding to the ecological validity of our work.

Procedure and Materials. We had participants complete the study in two ostensibly unrelated parts, with filler materials in between to disguise the relationship between the parts. Part 1 consisted of a motivation scale; in Part 2, we had participants rate the pseudo-studies. We cleared all studies by our research ethics board.

Motivation. We assessed participant trait motivation scores using the 30-item Work Preferences Inventory (WPI, Amabile et al., 1994) which assesses several reward-based and experience-based motivational drives. The WPI has good test–retest reliability across 5 months (rs $> .70$).

Pseudo-Studies. We created 16 pseudo-studies that were designed to simulate real studies, including superficial qualities common in our departmental research participation pool (see Table 1). We started with basic study descriptions (e.g., a pattern detection study), into which we incorporated four different strategies in the recruitment strategy: either highlighting getting credit faster than the usual rate of one credit per research hour (*Expedient Credit*); mentioning financial compensation or not (*Money*); alluding to an interesting-sounding study topic (*Narrative*); and implying that the study is enjoyable (*Fun*). The first two strategies were designed to appeal to reward-based motivations, and the latter two to experience-based motivations.

Participants were told that these studies were to be run in the next term, and rated how appealing each study seemed on a scale from 1 (*Not at all appealing*) to 9 (*Extremely appealing*). The pseudo-study stimuli were created in a 2 (Expedient Credit: No vs. Yes) × 2 (Money: No vs. Yes) × 2 (Narrative: No vs. Yes) × 2 (Fun: No vs. Yes) set of combinations, with each participant reading and rating all combinations of

Table 1. Example Study Advertisements (Experiment 1).

Condition	Description
0000	TREEHOUSE is a study about what opinions people draw about others based on the clothes that those people wear.
0001	Come join CAT, this enjoyable study involving musical learning. You will have fun in CAT!
0010	CABLE is a study for credit where we would like to learn about your attitudes and opinions, and self-esteem may play a role in making you confident in your thoughts.
0011	SAFE is a great study with a great prize: credit! This study seeks to unlock the secrets of brain activity in specific regions by simply monitoring your eye movement. Exceptionally fun and engaging study.
0100	In BRINE we use a video-game like task. You can earn cash for arriving 5 minutes early for the study because the study has precise timing.
0101	Earn $5 while you complete this study. MEADOW is a study about culture. Participants tell us that the study is quite enjoyable. ☺ Join today!
0110	AQUAMAN is unique: We want to see how you learn based on patterns of synchrony in the light flashes. Finally, you get entry to a cash draw as a nod of appreciation for your time.
0111	HAIRCUT is a study which gives you 1 credit. On top of that credit, you get entry to a cash draw for participating. The study is a mock courtroom experience. We are interested in learning how you perceive members of a jury and how objective you think they would be. Although there is some work, people say that it is quite a blast and really fun.
1000	ROBIN always gives credit a bit under time and is a study about personality.
1001	CARDS is a simple study about pattern detection. It is really quick, running just 40 minutes but still offers a credit.
1010	In DECISION you will listen to some music. You always get your credit pretty fast, 45 minutes or less. You can earn $5 if you are punctual because we have to start on time!
1011	WISH is a really fun study where you get to play games! The study investigates your attitudes toward winning and losing in games, focusing on your "play style" andwhether that can predict what games you like. We believe that play style may be a brand new concept that psychologists haven't thought about. People tend to get their credit a bit before the hour is up, too.
1100	In FISHSTICK you will be completing some nonspatial reasoning problems. You always get your credit even though the study can often take much less than the full hour. You can earn a few bucks by completing trials accurately.
1101	CLOUD is a quick study that still gives credit. You look at pictures and it can be quite fun because you get to play detective! Some participants finish in just 30 minutes but you always get 1.0 credit. The first 100 signups get a $5 bonus. ☺
1110	In WATCH we will ask you to monitor some clocks and report on time. We are interested to learn about your subjective experience of time. Ever notice how "time flies" sometimes? We want to find out why that happens. You earn $5 on top of credit. Additionally, the study can give you an extra half-credit if you stay just five minutes over the hour.
1111	BUBBLE is a study for a credit. In BUBBLE we want to know about your childhood, and what big events helped make you who you are today. In particular we investigate a theory about how standout memories can often come from situations that seem innocuous but mean a lot to the person. It's great fun and interactive, and people say they enjoy this new way of looking at their life. You can earn $3 if you sign up within the first 100 signups. Not a bad deal, especially as the study runs a bit short, only 45 minutes.

Note. The "Condition" column refers to the strategies used (1) versus not used (0) in each ad. The first value refers to the presence/absence of expedient credit, the second value to the money strategy, the third value to the narrative strategy, and the fourth value to the fun strategy.

factors among the studies, although each of the 16 studies was given a different basic description in terms of study name and topic. To avoid having the pseudo-study strategies be confounded by the subject matter of a particular study, we created 16 types of study topics (e.g., "HAIRCUT": a mock courtroom experience; "CARDS": a pattern detection study). These 16 topics were rotated across the set of strategies using 16 between-participant set conditions, decoupling subject matter from strategy type. Hence, the main effect of "cash" being rated as more appealing would not be accidentally driven by its being disproportionately associated with a more compelling-sounding piece of subject matter. An example set condition is displayed in Table 1. The presentation order was randomized within study sets. In total, we created ($2 \times 2 \times 2 \times 2 = 16$) $\times$ 16 set conditions = 256 pseudo-studies. A given participant, being in only one of the set conditions, only interacted with 16 of these pseudo-studies.

Results and Discussion

Motivation. We found that Amabile et al.'s (1994) original two-factor subscales (i.e., reward-based and experience-based) were modestly reliable (α = .67 and .73, respectively). Amabile et al. (1994) also examined a four-factor WPI. Thus, we ran analyses in all experiments twice: first, using the full set of items in the original two-factor model; and second, using a factor structure that improved the psychometric performance (which turned out to be four-factor).

Table 2. Multilevel Models Examined in Experiments 1 and 2, With Fit Indices and Model-Level Effect Sizes (Experiment 1)

Model	Fixed effects included	AIC	BIC	$R^2_{(m)}$ / $R^2_{(c)}$
Null Model	None	16841.9	16860.7	.000 / .177
Model 1	Cash, Credit, Fun, Narrative	16724.1	16767.9	.029 / .208
Model 2	As Model 1, plus Reward-based Motivation, Experience-based Motivation, and all Level 1–2 interaction terms.	16662.4	16768.7	.046 / .213
Model 3	As Model 2, but using fit-optimized, four-factor Reward/ Experience motivation.	16766.7	16841.8	.044 / .213

Note. Minimized AIC and BIC scores are optimal. $R^2_{(m)}$ is an estimate of the variability explained by the fixed effects (marginal R^2), and can be understood as an effect size estimate like other R^2 statistics. $R^2_{(c)}$ refers to the variability explained by random effects *plus* fixed effects (conditional R^2).

To do this, we conducted exploratory factor analysis using maximum likelihood extraction and direct oblimin rotation for multiple-factor solutions (Fabrigar et al., 1999), retaining only items with loadings >=.50 and without cross-loadings over half of the primary loading's magnitude. This produced four factors. Two were reward-based: *Outward* captures social validation goals (i.e., seeking recognition through accomplishment), and *Compensation* captures material goals (i.e., wanting grades or money). An additional two were experience-based: *Challenge* captures growth goals (i.e., wanting to learn from experience), and *Enjoyment* captures hedonic pleasure (i.e., using activities to enjoy oneself). These resemble the four "secondary factors" identified by Amabile et al. (1994).

Attraction to Study Strategies. Because all participants read and rated all 16 studies, 3,888 observations (ratings of each pseudo-study) were nested within 243 individual participants. Consequently, we used multilevel modeling (MLM) in nlme (Pinheiro et al., 2024). Fixed effects in MLM are comparable to standard regression except that MLM can permit defined coefficients to differ across levels of a higher-level organizing unit (Hayes, 2006). In this case, intercepts were set as random (permitted to vary) across participants. Qualities of the individual pseudo-studies (i.e., the presence/absence of strategies) were Level 1 variables (centered within each participant), and participant attributes (i.e., their motivation scores) were Level 2 variables (centered around the grand mean). Our key hypotheses were tested via cross-level interaction terms, which tested whether Level 1 variables (i.e., strategies) affected study attractiveness differently across levels of Level 2 variables (i.e., motivation scores).

Multilevel model statistics appear in Table 2. We ran four models. First, we tested a null model using no fixed effects to provide a baseline for comparison with respect to fit statistics. Second, we tested Model 1, which included only the Level 1 variables (i.e., the study strategy factors). This had the additional benefit of permitting a manipulation check, as all four reinforcer types were expected to increase the appeal of a pseudo-study. Third, we tested Model 2, which added Level 2 main effects (reward-based and experience-based motivation) to the model as well as interaction terms pairing reward-based and experience-based motivation with all four strategy conditions. Finally, Model 3 resembled Model 2 but substituted the fit-optimized four-factor WPI for the two-factor WPI factor solution. By appraising the relative fit of these four models, we avoided excessive multiple testing of many fixed coefficients by only examining the parameters of the best-fitting model.

As Table 2 indicates, fit improved sequentially from the null model to Model 1, and from Model 1 to Model 2. Akaike Information Criterion (AIC) values decreased from the null model to Model 1 and then Model 2; and $R^2_{(c)}$, representing the amount of variance in ratings of study appeal explained by the overall model, increased from .18 to .21 across the models. However, Model 3's additional complexity compared to Model 2 (i.e., moving from two moderators to four moderators) worsened fit according to all indices except $R^2_{(c)}$. In short, Model 2 balances fit and parsimony, so we examined its fixed predictors (the main and interaction terms of the model).

First, we note that Model 2 shows the main effects whereby experience-based motivation related to more interest in the pseudo-studies overall, but reward-based motivation was not related to more or less overall interest in doing studies.

However, these main effects were qualified by several key interactions, plotted in Figures 2A–C (complete results displayed in Table 3). First, we noted an interaction of reward-based motivation × expedient credit. We broke down simple slopes at one standard deviation above/below the mean of each motivation score. At low reward-based motivation, expedient credit was perceived as appealing, $B = .22$ [.04, .39], $t(3618) = 2.40$, $p = .016$, but at high reward-based motivation, it was even more appealing, $B = .54$ [.36, .71], $t(3618) = 5.97$, $p < .0001$. This provides initial support for hypothesis 1 that recruitment strategies may be differentially appealing based on potential participants' individual differences, even in this ecologically valid condition in which the target (incentivizing) information is mixed in with much irrelevant information.

Second, we found an interaction of experience-based motivation × money. Breaking down the simple slopes, we found that at low experience-based motivation, cash had a large impact on study appeal, $B = .56$ [.38, .73], $t(3618) = 6.20$, $p < .0001$. At high levels of experience-based

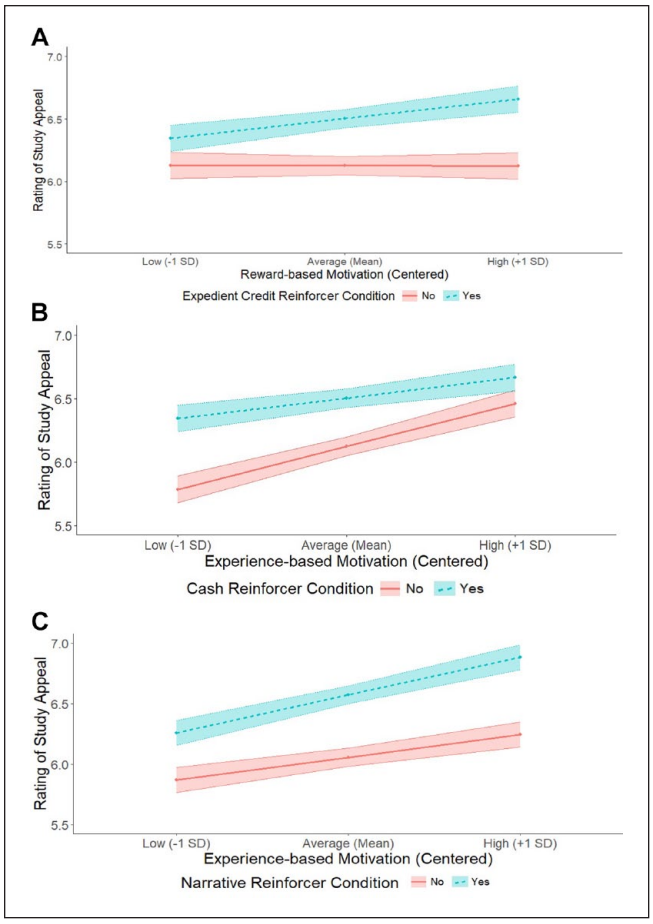

Figure 2. Appeal of Study Reinforcers Depend on Participants' Motivations. A: Expedient Credit Strategy Depends on Participants' Reward-based Motivation. B: Appeal of Cash Strategy Depends on Participants' Experience-based Motivation. C: Appeal of Narrative Strategy Depends on Participants' Experience-based Motivation.

motivation, cash showed a smaller impact on attraction, $B = .20$ [.03, .38], $t(3618) = 2.26$, $p = .024$. This inverts the shape of the pattern above, suggesting that studies that mention cash in their descriptions may be less appealing to participants who are high in dispositional experience-based motivation.

Third, we found a marginal ($p = .051$) interaction of experience-based motivation × narrative strategy. At low experience-based motivation, giving a narrative increased study appeal compared to not giving a narrative, $B = .39$ [.21, .56], $t(3618) = 4.31$, $p < .0001$. However, at higher levels of experience-based motivation, the narrative strategy had a larger impact on appealing, $B = .64$ [.46, .82], $t(3618) = 7.09$, $p < .0001$. This would indicate that recruitment strategies that include (vs. omit) some explanation of the study's purpose may draw in more experientially motivated participants, but results should be interpreted with caution.

Overall, small changes to how we characterized our studies via recruitment strategies affected the sort of participants who were attracted, supporting hypothesis 1. If this differential degree of appeal of recruitment strategies translates into participants' likelihood of showing up to studies (hypothesis 2), this may have implications for replication research (hypothesis 3). We tested each of these claims in Experiments 2 and 3.

Experiment 2

We have established that small variations in recruitment strategy wording can alter who is attracted to the study (Experiment 1). In Experiment 2 and 3, we randomly exposed participants either to a more reward-focused or experience-focused recruitment strategy, thus deliberately biasing our sampling in one of two ways. We then sought to replicate three effects from the psychological literature in each of Experiments 2 and 3 (see Table 4), examining whether replication outcomes depended on the type of recruitment strategy employed. For both Experiments 2 and 3, we note that people presumably participate in research for many reasons other than recruitment strategy wording (e.g., the convenience of the study dates/times for their schedule; willingness to do research in-person versus online; physical location of the laboratory). Furthermore, a substantial number of participants will enroll in the study without closely reading the materials, further diluting the maximum amount of biasing effect we could obtain. Nonetheless, we proposed that relatively subtle changes to recruitment materials can subtly bias the motivations of those recruited (Hypothesis 2), despite the substantial amount of additional, irrelevant variables that doubtlessly affect joining a study. Furthermore, we proposed that this might even alter the outcomes of replication studies (Hypothesis 3), assuming that the preponderance of biases introduced by the materials moderated an original effect in a unified direction.

We also considered that other differences in samples could emerge due to distinct recruitment strategies. For example, one might reason that the rewards-based message might draw participants who would complete the study less conscientiously or at least more efficiently (given that their motivation is focused on rewards and compensation) than those drawn by an experience-based message (for whom participation may be based on valuing the experience itself). Our experiment's duration was open-ended; participants could complete and terminate the study as quickly as they wanted. Therefore, we examined how long each participant spent on the study.

Method

Participants. Participants were 376 first-year psychology students (82% women, 16% men, and 1% non-binary) participating online for course credit. This sample size was larger than all original studies (1.6 times larger than Fishbach et al., 1.8 times larger than Kim et al., and 10.4 times larger than Oesch and Murnighan). Participants were recruited via one of two recruitment strategies and only received this type of message content.

Table 3. Fixed Effects in Multilevel Model Predicting Study Appealing From Reward/Experience Motivation and Recruitment Strategies (Experiment 1).

Predictor	*B* [$CI_{95\%}$]	*t*-test and *p*-value
Experience-based motivation	$B = .81$ [.27, 1.35]	$t(239) = 2.95$, $p = .004$
Reward-based motivation	$B = .04$ [−.49, .57]	$t(239) = .15$, $p = .879$
Expedient credit	$B = .38$ [.25, .50],	$t(3618) = 5.95$, $p < .0001$
Money	$B = .38$ [.26, .51]	$t(3618) = 6.01$, $p < .0001$
Narrative	$B = .51$ [.39, .64],	$t(3618) = 8.10$, $p < .0001$
Fun	$B = .16$ [.03, .28]	$t(3618) = 2.46$, $p = .014$
Experience-based × Expedient credit	$B = -.01$ [−.38, .36]	$t(3618) = -.03$, $p = .975$
Reward-based × Expedient credit	$B = .47$ [.10, .83]	$t(3618) = 2.51$, $p = .012$
Experience-based × Money	$B = -.52$ [−.89, −.15]	$t(3618) = -2.77$, $p = .006$
Reward-based × Money	$B = -.03$ [−.39, .34]	$t(3618) = -.14$, $p = .887$
Experience-based × Narrative	$B = .37$ [−.001, .74]	$t(3618) = 1.96$, $p = .051$
Reward-based × Narrative	$B = .13$ [−.24, .50]	$t(3618) = .70$, $p = .487$
Experience-based × Fun	$B = .02$ [−.35, .39]	$t(3618) = .10$, $p = .921$
Reward-based × Fun	$B = -.21$ [−.57, .16]	$t(3618) = -1.11$, $p = .267$

Note. Model effect size $R^2_{(c)}$, capturing variance collected across random and fixed effects, is .21.

Table 4. Complete List of Original Effects Replicated in Experiments 2 and 3, With Overview of Key Results.

Original effect authors	Description of the selected effect	Original observations / effect size	Replication observations / effect size	Replication success?	Changes from recruitment strategy?
Effects examined in Experiment 2					
Fishbach et al. (2004)	When people express two means to an end (versus just one), they devalue the means (with respect to enjoyment and importance).	$N = 227$ $\eta_p^2 = .02$; $\eta_p^2 = .02$	$N = 369$ $\eta_p^2 = .00$; $\eta_p^2 = .00$	No	N/A
Kim et al. (2017, Study 4)	People seeing themselves as relatively lacking in discretionary income become resentful, form intention to engage in luxury spending.	$N = 164$ *IE* = .51 [.30, .76]	$N = 294$ *IE* = .67 [.37, 1.01]	Yes	Yes
Oesch & Murnighan (2003)	People make more selfish hypothetical judgments about distributing money rewards as the absolute magnitude of money increases, despite proportion remaining the same.	$N = 35$ $\eta_p^2 = .34$	$N = 363$ $\eta_p^2 = .01$	Yes	No
Effects examined in Experiment 3					
Black & Davidai (2020, Study 3)	Luxury versus charitable spending of target wealthy person results in less favorable moral judgments of them, and less internal / more external attributions of the wealthy in general.	$N = 194$ $d = -1.69$; $d = -.30$ / $-.06$ [a]	$N = 358$ $d = -1.73$ $d = -.11$ / $+.20$	Partial	Yes
Kumar & Gilovich (2015)	More anticipated regret at prospect of not telling others about experiential purchase compared with at prospect of not telling others about materialistic purchase.	$N = 100$ $d = +.49$	$N = 383$ $d = +.51$	Yes	No
Peetz & Buehler (2009, Study 1)	Anticipating spending less money "next week" compared to how much they had just spent "this week."	$N = 31$ $\$_{diff} = +31.70$	$N = 380$ $\$_{diff} = +27.29$	Yes	No

[a]Although Black & Davidai's Study 3 found a $d = -.06$ "non-effect," others of their conceptually similar studies such as Study 1a found a positive effect of luxury > charitable spending on external attributions, $d = +.44$.

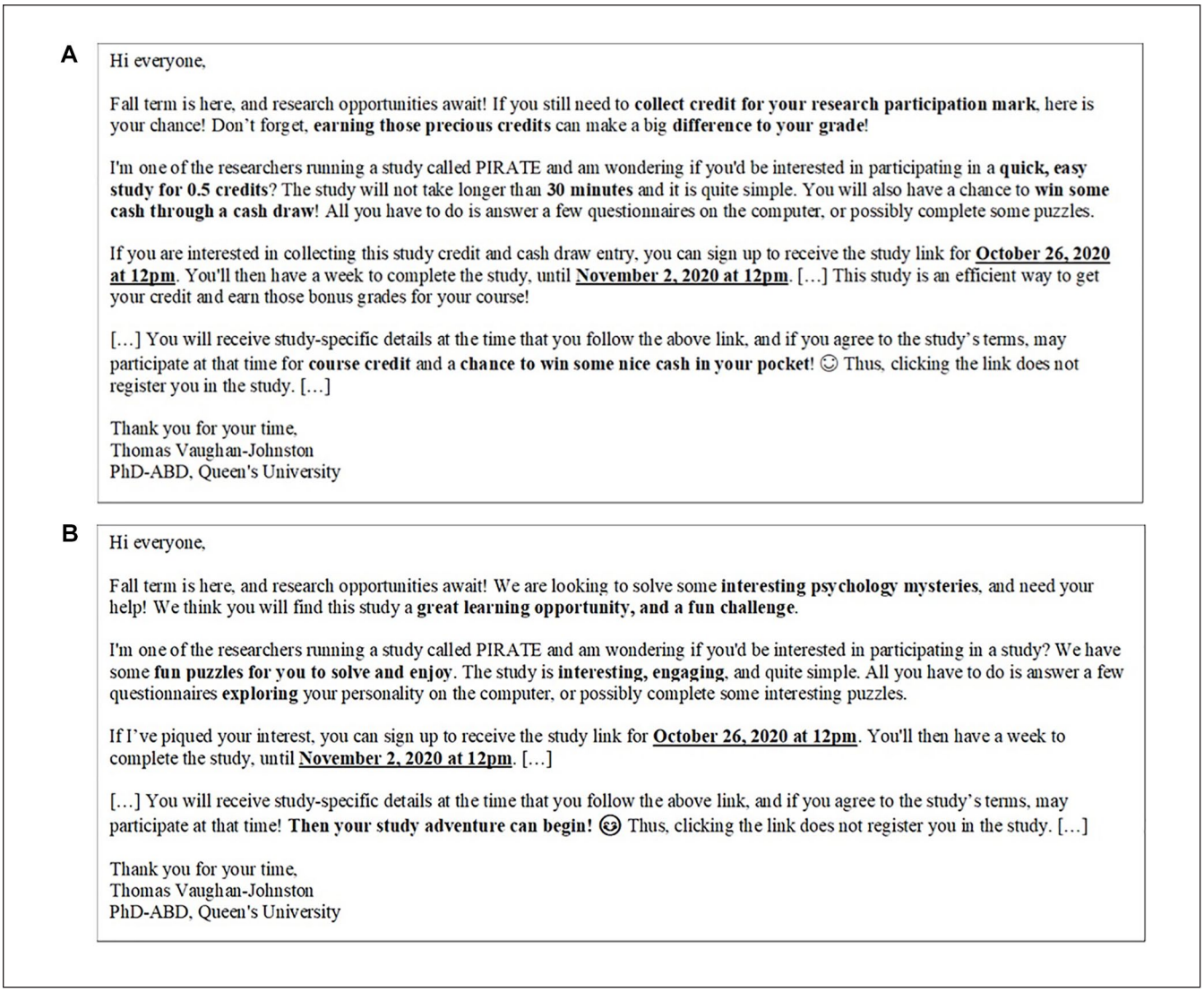

Figure 3. Stimuli Used in Experiment 2. A: Reward Motivation Advertisement. B: Experience Motivation Advertisement.

Procedure. We designed reward/experience-focused advertisements, utilizing the principles that appeared more appealing to reward/experience-motivated individuals in Experiment 1. Verbatim stimuli appear in SOM-2. The reward strategy focused on how the study allowed participants to collect course credit and emphasized the availability of a cash draw (see Figure 3A). In contrast, the experience-based ad emphasized the importance of participants' contribution, and implied that the study would be experientially satisfying (see Figure 3B).

Participants who agreed to participate were sent an individualized Qualtrics link so that their responses could be linked to their recruitment strategy condition. All participants then completed an identical set of procedures. Participants first filled out the WPI, before completing our three procedures, in a fixed order: (a) Fishbach et al. (2004), (b) Kim et al. (2017), and then (c) Oesch and Murnighan (2003). We later examine whether this order had any effects on replicability: it did not. More detail about the original studies, and our predictions about why reward- versus experience-focused motivation would shape their replicability, is provided in SOM-3. We did not change the manipulations or measures, other than by reducing materials in Oesch and Murnighan (2003) to the subset of materials that showed significant effects: the "Like" condition.

Fishbach et al. (2004, Study 3). Participants recorded a current goal (e.g., "earn my degree," or "being healthier"). Participants in the One Action condition were then asked to "state exactly **ONE** activity that you have been doing" to accomplish their goal, before rating "this action in terms of how enjoyable you think this action is" (*enjoyment*) and "in terms of how important you think this action is" (*importance*), each on 1 (*not at all*) to 7 (*extremely*) scales. In the Two Actions condition, participants acted similarly but recorded two relevant actions in separate boxes. They then rated the enjoyment and importance of both actions.

Kim et al. (2017, Study 4). We told participants that we were examining trends in the discretionary incomes of students and staff at our university and that they would complete a computerized procedure that would provide feedback about how their discretionary income compared with other people with similar "personal profiles" to themselves. Ostensibly to facilitate this matching process, we asked participants to complete a measure about their financial beliefs (Callan et al., 2011) and a Big Five personality measure (Gosling et al., 2003). They also completed demographic variables. When this feedback was entered, the computer told participants that their Comparative Discretionary Income (CDI) Index score would be produced, defining this as a person's "discretionary income relative to the discretionary income of similar others." Participants were then shown a .GIF image showing an animated circle and the words "Please wait while we process your data. . ." Seven seconds later, participants received feedback either that their "CDI Index Score" was "−$514" (Negative CDI) or "+$89" (Positive CDI). These values were designed to convert Kim et al.'s (2017) values of "−£313" and "+£54" into our local currency, based on the contemporary conversion rate. The feedback was displayed using a distinct font and background to make the feedback seem like the output of a computer process. This feedback was accompanied by information explaining "How to interpret your StatsPlus™ CDI Index Score," taken directly from Kim et al. (2017). In the original study, participants were asked to record their CDI Index Score on paper; because our replication was online, we stated that "The Index Score does not automatically save into the data file for research using it (the CDI calculation is a plug-in to Qualtrics). Please record your CDI Index Score into the box below to make it available to the researcher for use in his or her study." This was designed to enhance the cover story ("calculation is a plug-in to Qualtrics") and to ensure participants knew their score.

Participants then rated how dissatisfied, resentful, and satisfied (reverse-coded) they felt about the feedback (each item 1 = *Not at all*, 7 = *Very*; $\alpha = .78$). Participants then rated across five items their focus on discretionary spending (desire to increase it, importance of it, how much they wanted, how motivated they were to obtain it, how much they felt they needed it; all rated 1 = *Not at all*, 7 = *Extremely*; $\alpha = .94$). They then rated similarly-worded items but about charitable spending (same anchors; $\alpha = .95$). Following the original, we calculated *materialism* as the difference score between these measures: participants' interest in discretionary spending minus their interest in charitable spending.

Oesch and Murnighan (2003, Study 1). Participants were asked to "imagine someone you really like" and that this person would be their "partner" in a series of hypothetical binary choices between ways of distributing money (between themselves and their partner). Participants then completed 54 trials (in randomized order) of monetary judgments. These consisted of a 3 (Payout Ratio: Small vs. Medium vs. Large) × 3 (Payout Magnitude: Small vs. Medium vs. Large) within-participant matrix, with six trials per payout condition. This is slightly different than the original, as we explain in SOM-3. The complete set of choices is in SOM-1 with the verbatim materials, but example choices (from the Small Payoff condition) are "For you, $1. For your partner, $10" versus "For you, $10. For your partner, $1." Some choices benefited or hurt the partner with no consequence for the participant; others were dilemmas such as the provided example (benefiting either oneself or one's partner but not both). The Medium Payoff condition multiplied all values by 100, and the Large Payoff condition multiplied the values by an additional factor of 10. The Payout Ratio indicates how steep the self/other discrepancies were (e.g., a 10% difference being a Small Payout Ratio, 50% Medium, 90% Large). Choices were coded so that higher values represent more prosociality. The original study contained payout conditions in which participants imagined dealing with someone neutral or someone they disliked, but these conditions did not produce significant effects in the original so we dropped them.

Results and Discussion

Motivation. We first wanted to determine if our recruitment strategies brought in people differing across levels of motivation, representing hypothesis 2. Like Experiment 1, we analyzed the data using both the full, two-factor approach, and also using an fit-optimized four-factor approach using only those items loading most highly on each common factor. Thus, we had six (overlapping) dependent variables: two when viewing reward/experience motivation as two-factor, and four when viewing reward/experience motivation as four-factor.

Two-Factor Model. Recruitment strategy was not related to either of the full (non fit-optimized) measures of reward/experience motivation, $ps > .180$, $ds < |.14|$.

Four-Factor Model. Using the fit-optimized, four-factor model of motivation items as dependent variables, we found one effect: Participants recruited via our rewards-based strategy ($M = 2.70$, $SD = .73$) were more motivated by outperforming / "beating" others (Outward subscale) compared with those in the experiential strategy condition ($M = 2.54$, $SD = .68$), $t(369) = 2.27$, $p = .024$, $d = .24$ [.03, .44]. Thus, despite all of the other considerations weighing on participants' willingness to show up for a study (and despite participants receiving hundreds of other recruitment messages from other researchers in our participant pool alongside our own stimuli), small wording differences in our recruitment materials were sufficient to produce a detectable personality bias in who showed up to participate. Specifically, recruitment strategies emphasizing money and credit prompted an increased proportion of competitive participants who might be expected to behave differently in our replication studies than those recruited via the experiential message. There were no effects on the other three factors, $ps > .492$, $ds < |.08|$.

Participation Time. Participants in the rewards-based recruitment strategy condition completed the study faster (M = 22.6 min, SD = 10.9 min) than those in the experience-based condition (M = 24.4 min, SD = 10.2 min), $t(320) = 2.00$, $p = .046$, $d = .22$ [.00, .44].[2] Thus, not only did rewards-based condition participants tend to be motivated differently, but there was some indication of their also being *less* motivated in that they worked through the study more rapidly than participants recruited via experience-based messages.

Replications

Replication of Fishbach et al. (2004, Study 1).

Before conducting our analysis, we eliminated seven participants for providing invalid goals (one just said "life," the rest were blank), leaving $n = 369$. We then performed our analyses, attempting to mirror Fishbach et al.'s (2004) analytic plan, using ANOVA models and re-analyzing the models with each dependent variable added as a covariate in tests of the other dependent variable (e.g., controlling for enjoyment when analyzing importance).

Enjoyment. Enjoyment was rated similarly highly in both Action Number conditions, $F(1, 367) = .05$, $p = .831$, $\eta_p^2 = .00$, with moderate enjoyment ratings in the One Action (M = 4.43, SD = 1.63) and the Two Action conditions (M = 4.47, SD = 1.52). Controlling for importance made no meaningful difference, $F(1, 366) = .04$, $p = .838$, $\eta_p^2 = .00$. Thus, we did not replicate Fishbach et al. (2004).

Importance. Importance was rated similarly highly in both Action Number conditions, $F(1, 367) = .59$, $p = .442$, $\eta_p^2 = .00$, with high importance ratings both in the One Action (M = 6.44, SD = .81) and the Two Action conditions (M = 6.51, SD = .88). Controlling for importance made no meaningful difference, $F(1, 366) = .59$, $p = .444$, $\eta_p^2 = .00$. Thus, we did not replicate Fishbach et al. (2004).

Replication of Kim et al. (2017)

Understanding of the CDI Index. We followed the original study's procedure for removing two types of participants. The original removed seven participants (out of 164; i.e., 4.3%) for failing to understand the CDI feedback. In contrast, a substantial number of our participants did not understand the CDI Index (n = 56, 14.9%), leaving $n = 320$. These participants openly stated they did not understand (e.g., "Idk," "not too sure," "I have no idea") or provided an incorrect definition (e.g., "How much debt my family has"). Of those participants who we retained from this first step, an additional 4 (1.3%; compare Kim et al.'s 1.8%) were suspicious of feedback (e.g., "I was being bamboozled," "it was fake"). Eliminating these left 316 participants.

Replication Analysis. The core original analysis was an indirect effect from CDI Feedback condition to materialism (interest in discretionary spending minus charitable spending) via resentment, $IE = .51$ [.30, .76], significance determined by the bootstrapped indirect effect not overlapping with zero. This means that getting negative (vs. positive) feedback made people resentful, and greater resentment about the feedback predicted an increased desire for materialistic spending. We successfully replicated this effect, finding a significant a-path from CDI feedback condition to resentment, $B = 1.44$ [1.18, 1.71], $t(292) = 10.61$, $p < .001$, and a significant b-path from resentment to materialism, $B = .46$ [.28, .64], $t(291) = 4.98$, $p < .001$, with an overall significant mediation effect, $IE = .67$ [.37, 1.00]. This replicates Kim et al.'s (2017) core finding with a similar effect size.

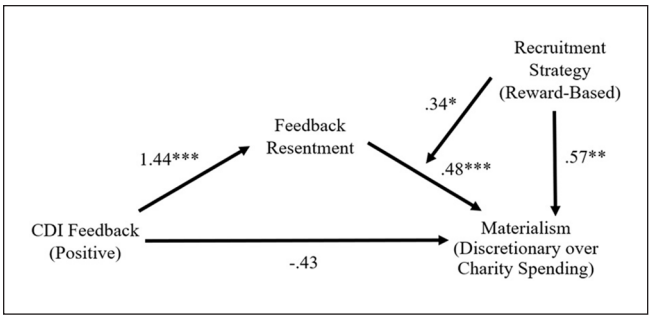

Figure 4. Replication of Kim et al. (2017) Personal Relative Deprivation Effect Depends on Recruitment Strategies: Moderated Mediation Model (Experiment 2).

Moderation by Recruitment Strategy. Kim et al.'s key model demonstrated that negative CDI feedback conditions increased resentment, which prompted participants to desire more discretionary income. However, recruitment strategies could disrupt this in one of two ways. (a) The a-path of this indirect effect could vary based on recruitment strategy, such that receiving CDI feedback produces larger effects on resentment for participants brought to the study via the reward (versus experience) message. Alternatively, (b) resentment might lead to increased materialism more strongly for participants brought to the study via the reward (versus experience) message. We reasoned that *either or both* effects might be greater for the more competitive-minded individuals brought in by our reward-based recruitment strategy.

To test (a), we used Hayes' Model 7, which assesses whether a variable moderates the a-path in a mediation model. The CDI Feedback × recruitment strategy condition interaction was non-significant, $B = -.06$ [−.60, .48], $t(290) = -.22$, $p = .824$, and the moderated mediation index was non-significant, $B = -.03$ [−.29, .23], providing no evidence that our reward-based message group was more resentful of negative CDI feedback than was our experience-based message group.

To test the second possibility, we used Hayes' Model 14, which assesses whether a variable moderates the b-path in a

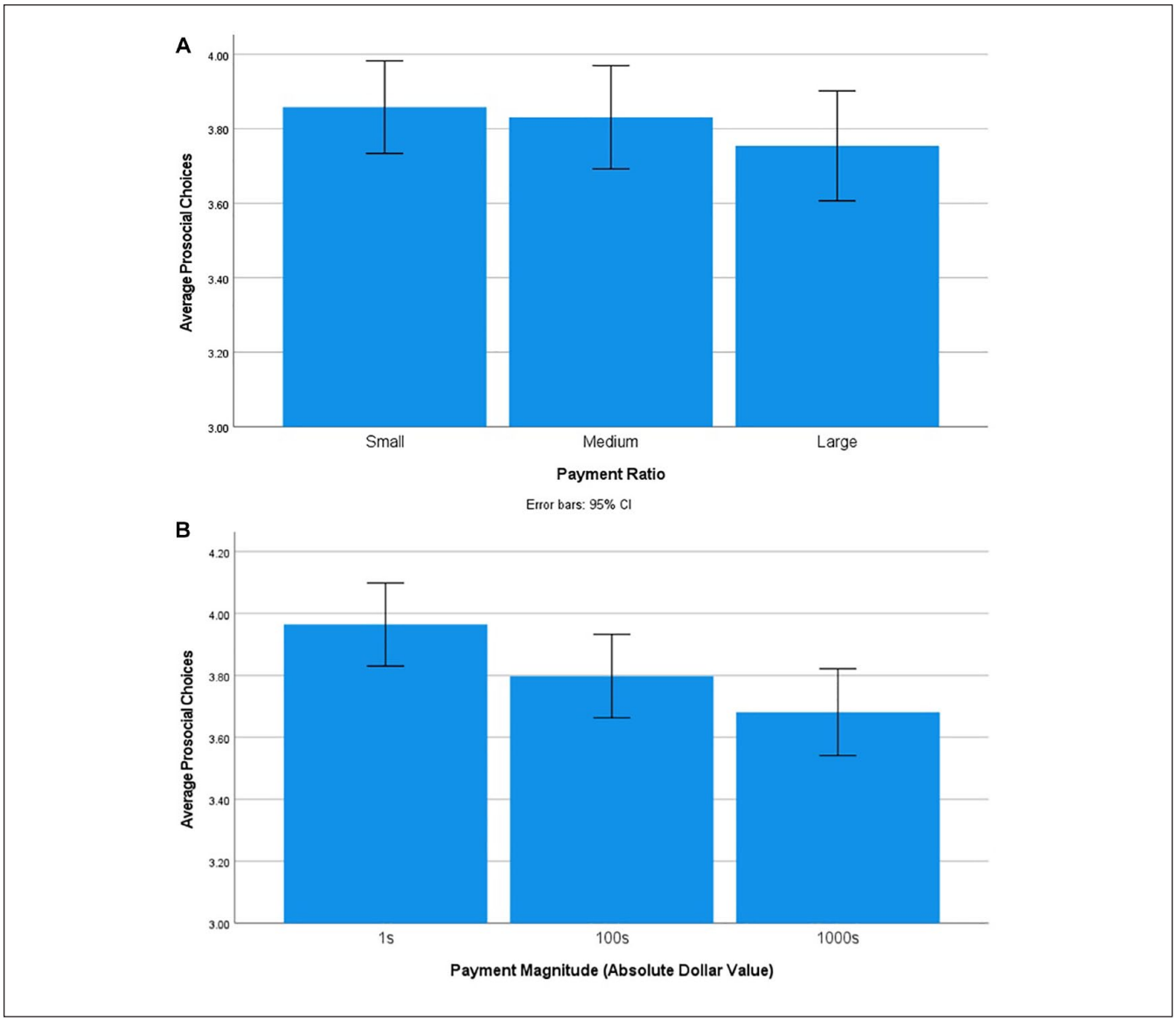

Figure 5. Effects from Replication of Oesch and Murnighan (2003). A: Prosocial Choices in Hypothetical Distributions of Money Depends on Ratio. B: Prosocial Choices in Hypothetical Distributions of Money Depends on Absolute Dollar Amount.

mediation model. The resentment X recruitment strategy condition interaction was significant, $B = .34$ [.03, .65], $t(289) = 2.17$, $p = .031$, and the moderated mediation index was significant, $B = .49$ [.02, .97], suggesting that resentment was a better predictor of materialistic desires in one of our recruitment strategy conditions. Indeed, resentment predicted materialism about twice as strongly for reward-message participants, $B = .65$ [.42, .89], $t(289) = 5.43$, $p < .001$, as for experience-message participants, $B = .31$ [.08, .55], $t(289) = 2.59$, $p = .010$. Relatedly, the indirect effect (from CDI feedback to materialism via resentment) was more than double the magnitude for reward-message participants, $IE = .94$ [.58, 1.33] than for experience-message participants, $IE = .45$ [.06, .88]. See Figure 4 for a visual display of this moderated mediation effect. This supports Hypothesis 3 by showing that different recruitment strategies not only produced biased samples but that different strategies also produced tangible impacts on our replication of Kim et al.'s finding.

In addition, we identified a main effect of recruitment strategy on materialism. This reveals that collapsing across CDI Feedback, reward-based recruitment strategy drew in a more materialistic sample than the experience-based recruitment strategy, $B = .57$ [.15, .99], $t(289) = 2.67$, $p = .008$, $d = .31$. This is a third piece of evidence supporting hypothesis 2 that recruitment messages produced biased samples.

Replication of Oesch and Murnighan (2003, Study 1). To replicate Oesch and Murnighan (2003), we should find that selfishness increases as the absolute payout magnitude

increases. We conducted a repeated-measures ANOVA with selfishness as the dependent variable, and Payout Ratio as the within-participant manipulation, finding a significant effect, $F(2, 728) = 3.39$, $p = .034$, $\eta_p^2 = .01$. This effect is much smaller than the original's effect size ($\eta_{p\,original}^2 = .36$). We displayed these results in Figure 5A, which also shows how the decrement in prosociality did not emerge between small- and medium-ratio conditions, $M_{diff} = .03$, $p = .418$, but only between medium-ratio and high-ratio conditions, $M_{diff} = .08$, $p = .026$, with the small-ratio and high-ratio conditions also differing significantly, $M_{diff} = .11$, $p = .043$. Alternatively considered, we found that people made 11.5 prosocial choices in the small-ratio condition, 11.4 in the medium-ratio condition, and 11.2 in the high-ratio condition. Once Oesch and Murnighan's numbers are adjusted for having more trials than our replication (but just examining their Like condition), they found an average 18.7 prosocial choices in the small ratio, 14.6 in the medium ratio, and 10.5 in the high ratio condition, suggesting that payment ratio more powerfully dictated their participants' level of generosity. However, it should be emphasized that Oesch and Murnighan's theory only made a directional hypothesis and did not specify a specific effect size, so our findings are consistent theoretically and in direction, just not in magnitude.

Unexpectedly, we also found an effect of Magnitude, $F(2, 728) = 25.76$, $p < .001$, $\eta_p^2 = .07$, which Oesch and Murnighan did not. We displayed these results in Figure 5B, which captures how prosociality decreased from the low-magnitude ("1s") to medium-magnitude ("100s") trials, $M_{diff} = .17$, $p < .001$ and decreased from medium-magnitude to large-magnitude ("1000s") trials, $M_{diff} = .12$, $p = .001$. Less generosity was also detected in the large- compared to the low-magnitude trials, $M_{diff} = .28$, $p < .001$. Although this was not specifically identified by the original authors, we would argue that it is conceptually consistent with the original authors' notion that "friendships [lead] to choices that are similar to. . .politeness rituals and reciprocity norms. . .but [do] not consistently reinforce equality norms" (p. 59). That is, rather than equality preferences toward liked others being fixed and invariant, our results (like Oesch and Murnighan's) suggest that prosociality depends on several contextual features: the relative ratio of self/other payouts (original and replication finding) but also the absolute dollar-value magnitude of payouts (replication finding only). For instance, our participants were less generous when distributing thousands versus only hundreds of dollars even in hypothetical judgments toward people they liked, showing the contextual fragility of sharing.

Moderation by Recruitment Strategy. We re-ran the ANOVA model as a mixed model with the recruitment strategy condition as a between-participant factor. The recruitment strategy did not moderate the effect of payout ratio, $F(2, 726) = .89$, $p = .412$, $\eta_p^2 = .002$, nor did it moderate the effect of payout magnitude, $F(2, 726) = .49$, $p = .614$, $\eta_p^2 = .001$. This does not support the idea that the Oesch and Murnighan (2003) effect was altered by our recruitment strategies.

Interference Effects. No interference effects of earlier manipulations were detected on the Kim et al. (2017) replication, all ps $> .372$; nor on Oesch and Murnighan (2003)'s replication, all ps $> .261$.

Experiment 3

Experiment 2 demonstrated that the use of heterogeneous recruitment materials led to biases in three different variables (higher reward-based motivation, higher materialism, and faster completion time among those drawn by rewards versus experiences), robustly supporting hypothesis 2. This happened despite the miscellany of noise variables working against creating such a bias in an actual recruitment situation (e.g., people not reading the messages closely, signing up for other reasons such as timing convenience, etc.). More impressively, we were even able to show differences in a replicated study's effect caused by recruiting for that study differently (Hypothesis 3).

In Experiment 3, we wanted to replicate our effects on Hypothesis 2, showing that we could once again produce biases in the actual personalities of participants merely by tweaking how we phrased our messages. We then wanted to extend our findings by targeting three different original effects distinct from those in Experiment 2. Furthermore, given our very clear expectations of how the recruitment materials should bias participant personality, informed by Experiment 2, and therefore how our effects' replications might be moderated, we preregistered Experiment 3's sampling plan, hypotheses, and analytic plans (https://osf.io/g4wq6).

Method

Participants. Per our preregistration, we collected an online sample ($N = 187$) and an in-person sample ($N = 390$) following a time-based stopping rule across two years (see SOM-4 for additional details). However, following our preregistration, we relegated the online sample to the supplement because it did not produce comparable sample-biasing properties compared to Experiment 2 (see SOM-4 for a complete report) and so would not be expected to moderate replication results. The in-person sample size was larger than all original studies (1.8x larger than Black and Davidai, 3.8x larger than Kumar and Gilovich, and 12.2x larger than Peetz and Buehler). Participants were primarily women (84%; 15% men, 1% other), young adults ($M_{age} = 18.3$, $SD_{age} = 1.1$), and primarily White / European American (76%), with 10% self-describing East Asian, 2% Indian, 2% Black, 2% Latinx, 7% other, and 2% missing. On a 1 (*extremely liberal*) to 4 (*moderate*) to 7 (*extremely conservative*) scale, our participants were somewhat liberal ($M = 3.1$, $SD = 1.2$).

Procedure. The replication studies were completed in a fixed order: Black and Davidai (2020), Kumar and Gilovich (2015), and then Peetz and Buehler (2009). We later examine the possibility of order effects, but as Klein et al. (2018)

demonstrated, such effects are usually very small across many psychological effects. Participants then filled out the WPI and materialism measure. We did not change the original studies' manipulations or measures, other than localizing the manipulation and measures in Black and Davidai by changing location names.

Black and Davidai (2020, Study 3). Participants were introduced to a bogus news story ostensibly from the "Money Matters" website. They were assigned to learn that a "recently decreased multi-millionaire" had given their money either to the "Cece Cares Foundation" (*charitable giving* condition) or to "Cece the Dog" (*luxurious spending* condition). References in each condition were changed from "Southern California" (in the United States) to "Ontario" (in Canada) to localize the manipulation. Next, participants answered randomized-order scales that addressed attributions about how people become rich: via internal means (three items, e.g., "hard work and initiative," $\alpha = .84$) or via external means (four items, e.g., "Good luck, being in the right place at the right time," $\alpha = .60$). Items were rated from 1 (*not at all important*) to 7 (*extremely important*). In addition, they rated the moral character of the multi-millionaire described in the story on 11 items (e.g., "a principled person," "a fair person," $\alpha = .93$) on scales ranging from 1 (*not at all*) to 5 (*extremely*). Finally, as in the original study, an attention check question was given: participants answered which of the two passages they recalled reading.

Kumar and Gilovich (2015, Study 1a). Participants were assigned to either list the most significant "experiential purchase" (*experiential condition*) or the most significant "materialistic purchase" (*materialistic condition*) they made in the last 5 years. Brief definitions were provided for each: experiential was defined as "spending money with the primary intention of acquiring a life experience—an event or series of events that you personally encounter or live through." Materialistic was defined as "spending money with the primary intention of acquiring a material possession—a tangible object that you obtain and keep in your possession." Next, participants were asked to focus on the "portion of happiness that comes from talking about purchases." Participants were asked to imagine that a friend or relative requested they not talk to anyone about the previously listed purchase. It was clarified that they would still have the possession or memory but had agreed to not tell anyone else about it. The dependent variable was a single item: "How much does this idea bother you?" rated from 1 (*not at all bothered*) to 5 (*moderately bothered*) to 9 (*extremely bothered*).

Peetz and Buehler (2009, Study 1). This simple study consisted of just two questions, as we only conducted the first timepoint of the study. First, participants were asked to consider how much money they expected to spend next week (defined as "the next seven days; all expenses included except things that occur only once a month such as rent"). Second, they were asked to consider how much money they spent last week (defined as "the past seven days" with identical caveats). The budget fallacy is calculated by comparing these values: the amount expected to spend "next week," and the amount actually spent "last week."

Results and Discussion

Note that we preregistered that we would use one-tailed tests because we specified directional hypotheses in the preregistration.

Materialism. Materialism was significantly higher in the reward-based ad group ($M = 1.78$, $SD = 2.12$) versus the experience-based strategy group ($M = 1.42$, $SD = 1.84$), $t(380) = 1.76$, $p_{\text{one-tailed}} = .040$, $d = .18$ [−.02, .38]. Thus, we successfully replicated the bias effect: simply employing a reward-focused versus challenge/fun-focused message led one of our subsamples to be more materialistic than the other, again supporting hypothesis 2.

Motivations. We again found evidence for a four-factor structure that produced the same factors as in our prior experiments: two EM factors (*EM-Outward* and *EM-Compensation*) and two IM factors (*IM-Challenge* and *IM-Enjoyment*). However, the recruitment strategy condition did not substantially bias any of these factors; ts $< |1.36|$, ps$_{\text{one-tailed}} > .087$.

Participation Time. Replicating Experiment 2, participants in the rewards-based recruitment strategy condition completed the study faster ($M = 18.1$ min, $SD = 9.4$ min) than those in the experience-based condition ($M = 22.0$ min, $SD = 12.0$ min), $t(388) = 3.54$, $p < .001$, $d = .36$ [.16, .56]. This continues to support a similar picture for Hypothesis 2: rewards-based messages not only bring in more materialistic and perhaps rewards-oriented people, but also draw in people who give the study less of their time.

Replications

Black and Davidai (2020, Study 3). Before conducting any analyses, we removed 7% of participants for failing an attention check. We successfully replicated Black and Davidai's (B&D's) effects of spending conditions on moral judgments of the spender, $t(356) = 16.31$, $p_{\text{one-tailed}} < .001$, $d = 1.73$ [1.48, 1.97], with more negative judgments of the luxurious ($M = 3.65$, $SD = .50$) versus charitable spender ($M = 2.65$, $SD = .65$). Unlike B&D, we did not find any effect of spending condition on internal attributions, $t(356) = 1.05$, $p_{\text{one-tailed}} = .147$, $d = .11$ [-.10, .32], although the means fell consistently with B&D: internal attributions being elevated given the charitable ($M = 5.04$, $SD = 1.26$) versus luxurious spender ($M = 4.90$, $SD = 1.41$). Interestingly, however, we found an effect theoretically consistent with B&D and present in others of their studies, but not detected in the particu-

Figure 6. Replication of Black and Davidai (2020, Study 3) Effect of Target Spending Behavior on Perceived Target Morality Depends on Recruitment Strategies (Experiment 3).

lar study we replicated: spending condition affected *external* attributions at least by the standards of a one-tailed test, $t(356) = -1.89$, $p_{\text{one-tailed}} = .030$, $d = -.20$ [−.41, .01], with more external attributions given the luxurious ($M = 4.82$, $SD = .91$) versus charitable spender ($M = 4.62$, $SD = 1.10$).

Moderation by Recruitment Strategy. Our key hypothesis was that the above effects may differ according to the recruitment strategies employed. Starting with moral judgments of the spender, recruitment strategy condition indeed interacted with spending condition to affect morality, $F(1, 354) = 3.25$, $p_{\text{one-tailed}} = .036$, $\eta_p^2 = .01$.[3] As we preregistered, the effect of spending condition on moral judgment was larger given the reward-based recruitment strategy condition, $M_{\text{diff}} = 1.11$ [.95, 1.28], $SE = .09$, $p < .001$, and smaller given the experience-based recruitment strategy condition, $M_{\text{diff}} = .89$ [.72, 1.07], $SE = .09$, $p < .001$. This is reflected in Figure 6, in which the decreased liking of a target who spends luxuriously (right bars) rather than generously (left bars) is larger within the reward-based condition (larger gap between the blue triangles) relative to the experience-based message condition (smaller gap between the red squares). We did not expect recruitment strategy condition to have a main effect, nor did it, $F(1, 354) = 3.16$, $p_{\text{two-tailed}} = .076$, $\eta_p^2 = .01$.

We turn next to the attribution effects. There was no significant interaction of recruitment strategy condition × spending condition on internal attributions, $F(1, 354) = 2.64$, $p_{\text{one-tailed}} = .053$, $\eta p^2 = .01$, nor for external attributions, $F(1, 354) = 1.23$, $p_{\text{one-tailed}} = .134$, $\eta p^2 = .00$.

Kumar and Gilovich (2015, Study 1a). We successfully replicated Kumar and Gilovich's effects. Like the original, participants were more bothered by the idea of not sharing an experiential purchase ($M = 4.87$, $SD = 2.28$) than the idea of not sharing a materialistic purchase ($M = 3.73$, $SD = 2.23$), $t(381) = 4.98$, $p_{\text{one-tailed}} < .001$, $d = .51$ [.31, .71].

Moderation by Recruitment Strategy. The interaction term with message type was non-significant, $F(1, 379) = .37$, $p_{\text{one-tailed}} = .273$, $\eta p^2 = .00$. Specifically, the experiential/materialist purchase gap in storytelling value was similar given our reward-based message, $M_{\text{diff}} = 1.29$, $F(1, 379) = 15.78$, $p < .001$, $\eta p^2 = .04$, and our experience-based message, $M_{\text{diff}} = 1.01$, $F(1, 379) = 9.54$, $p = .002$, $\eta p^2 = .03$. We did not hypothesize that recruitment strategy condition would have a main effect, nor did it, $F(1, 379) = 2.52$, $p_{\text{two-tailed}} = .113$, $\eta_p^2 = .01$.

Peetz and Buehler (2009, Study 1). Before we ran analyses, we noticed a single very large outlier of US$11,050 for one participant's "last week" spend, which we removed. A paired-samples *t*-test revealed that we replicated Peetz and Buehler's finding: people anticipated spending less money "next week" ($M = \$97.94$, $SD = 120.77$) compared with "last week" ($M = \$125.23$, $SD = 146.82$), $t(379) = -3.18$, $p_{\text{one-tailed}} = .001$, $d = -.16$ [−.26, −.06], translating to $M_{\text{diff}} = -\$27.29$ or a projected 21.8% reduction in spending from last week to next week.

Moderation by Recruitment Strategy. The interaction term with message type was non-significant, $F(1, 378) = .00$,

$p_{\text{one-tailed}} = .478$, $\eta p^2 = .00$. Specifically, the spending fallacy effect was similar given our reward-based message, $M_{\text{diff}} = \$27.75$, $F(1, 378) = 5.29$, $p = .022$, $\eta p^2 = .01$, and given our experience-based message, $M_{\text{diff}} = \$26.81$, $F(1, 378) = 4.79$, $p = .029$, $\eta p^2 = .01$. We did not hypothesize that recruitment strategy condition would have a main effect, nor did it, $F(1, 378) = 1.82$, $p_{\text{two-tailed}} = .179$, $\eta_p{}^2 = .01$.

Interference Effects. Because the three studies were presented in a set order and were experimental designs, we tested if assignment to condition for any experiment interfered with experiments run later in the sequence (either as main effects or interacting with the subsequent studies' factors). No such interference effects manifested across six such tests (i.e., Black & Davidai main/interaction effects on the later two studies; Kumar & Gilovich main/interaction effect on Peetz & Buehler were all $p > .498$).

Deviations From the Preregistration. In the preregistration, we said that we would exclude inattentive participants in our replication of Black and Davidai (2020) if at least 10% of participants were inattentive. We had anticipated near-zero inattention as we have (successfully) replicated this experiment in the past with near-zero inattention. In fact, 7% of the present sample were inattentive but we excluded them anyway because we reasoned it was difficult to justify keeping participants who misunderstood their condition assignment. We report results with and without this exclusion.

For our replication of Peetz and Buehler (2009), we did not preregister removing outliers, but the presence of this outlier inflated the standard error. In any case, removing a positive "last week" condition observation would if anything have worked *against* confirming the original hypothesis ("last week > next week").

We did not preregister the participation time analysis because we detected this in Experiment 2 after preregistering and running Experiment 3.

General Discussion

Psychologists often consider direct replication to be valuable (Brandt et al., 2014; Zwaan et al., 2018; but see Fabrigar & Wegener, 2016; Luttrell et al., 2017, for some qualifications), making it imperative to evaluate the many factors required for establishing direct replication (Brandt et al., 2014; Steiner et al., 2019). Although much replication literature has been dedicated to examining diverse factors crucial for successful replication (e.g., construct validity of operationalizations, Flake et al., 2022; statistical power, Anderson & Maxwell, 2017), the role of recruitment strategies has generally been neglected. Our three experiments reveal that people are drawn to participate in studies more when those studies are advertised in a way conforming to their motivational predispositions (Hypothesis 1), and that this creates a bias in the type of participant who participates (Hypothesis 2). Experiments 2 and 3 then each show that this can even have consequences for replicating studies: at least one psychological effect per experiment was enhanced or dampened by this biased sampling (hypothesis 3). The implication is that replication effect sizes (and thus the likelihood of successful replication) vary according to the recruitment tactics that one employs.

We also ran an additional, correlational study, reported as a supplementary experiment. It attempted to replicate the same three original studies as the main text Experiment 3. However, rather than using biased samples drawn through recruitment materials, we ran the studies as normal but asked participants at the end of the study which future pseudo-studies (similar to Experiment 1) they found appealing. We used this as a proxy for actually biasing the samples, reasoning that if people who endorsed reward-focused or experience-focused studies behaved differently in the studies, it provided evidence that employing corresponding recruitment strategies could affect replicability. As SOM-6 notes, we again replicated all three original studies. We also found that our replication of effects from all three studies depended on participants' interest in reward-focused or experience-focused studies.

Contributions

Given that psychology is grappling with a replication crisis, the mere fact that we repeatedly obtained successful replications of key effects from Black and Davidai (2020), Kumar and Gilovich (2015), Peetz and Buehler (2009), Kim et al. (2017), and Oesch and Murnighan (2003) is itself noteworthy. Although some of the original studies originally used Mechanical Turk participants, we nonetheless successfully replicated them with our university student population (likely to not be a trivial change; Hauser et al., 2019). The only non-replicating study (Fishbach et al., 2004) might be explained by statistical power limitations: we did not reach Simonsohn's (2015) "2.5 *N*" guideline in this case, and the original effect size was modest.

Beyond providing numerous contributions to a growing body of direct replication studies, we showed several novel findings in the present work. First, we demonstrated that even subtly different recruitment strategies are differently appealing to people based on their dispositional motivations (Hypothesis 1; Experiment 1). Second, we were able to create subtle biases in our samples simply by using distinct recruitment strategies (Hypothesis 2; Experiments 2 and 3). Recruitment strategies mentioning rewards (vs. experiences) brought in biased samples of more reward-motivated (Experiment 2), materialistic (Experiments 2 and 3), and fast-responding people (Experiments 2 and 3) which is itself potentially noteworthy for any research relevant to these individual differences. Furthermore, we showed that two

different effects were moderated by recruitment methods (Hypothesis 3; Experiments 2 and 3). The present findings are relevant especially given that reward/experience motivation is a key psychological variable (Cerasoli et al., 2014; Deci et al., 1999; Hidi & Harackiewicz, 2000) with the potential to influence many effects, but in principle one could imagine a broad array of research conceivably influenced by variability in reward-based and materialistic motivations—judgment/decision-making paradigms involving marketing (Dowling et al., 2020), motivation in learning (Ryan & Deci, 2020), consumer behavior research (Srikant, 2013), and economic game paradigms (van Dijk & De Dreu, 2021) seem particularly likely to be influenced by these effects. Similarly, participants completing studies more rapidly, as in our reward-based motivation conditions, may be more likely to fall for judgment/heuristic effects (Kahneman, 2011), be persuaded on the peripheral versus central route in attitudes research (Petty & Cacioppo, 1986), and show more dispositional rather than situational attributions in person perception research (Gilbert et al., 1988). Future research may examine how many of these effects are borne out, possibly in the form of preregistered replications of original experiments.

Strictly, we did not reduce any original effects to undetectably small sizes ("fail to replicate") via our method; however, we were very well-powered for all experiments, so failing to replicate would have been comparatively unlikely—provided that the original findings were valid. Nonetheless, decreasing effect sizes is important, not only because reductions from the original studies' effect sizes are often characterized as inherently problematic in the replication literature (e.g., Camerer et al., 2018), but because corroding an effect size in this way jeopardizes the probability of recovering a significant effect favoring the original hypothesis. For example, Kim et al. (2017) collected 164 for their original experiment. Suppose we matched their sample size and attempted to detect their association between felt resentment and materialism (relative interest in discretionary over charitable spending). In our reward-based message condition in Experiment 2, we obtained $r = .42$ for this correlation (i.e., the *b*-path in Experiment 3's replication of Kim et al.); in our experience-based message condition, this shrunk to $r = .16$. Thus, using Kim et al.'s sample size would obtain approaching-100% power to replicate their effect if that sample were recruited given a rewards-based message, but this drops to 54% statistical power to replicate their effect if that sample were recruited given the experiential message. Alternatively considered, to obtain 80% power, we estimate needing 47 participants sampled through our rewards-focused versus 309 participants sampled through the experience-focused ad.[4]

In short, if the original effect is modestly sized or fragile, the recruitment strategy condition could potentially shape the probability of replicating that effect. Even if the original effect is substantive, in cases where a researcher cares about effect size, that effect size may depend on recruitment strategy conditions. Furthermore, in cases where a researcher might care about descriptive statistics (such as baseline interest and willingness to give money to charity), the distortion of variables due to recruitment strategy condition is potentially problematic.

These findings matter for original researchers, who might use different recruitment strategies across within-lab replications, unaware that this can matter. Our work also has important implications for replication science. First, it shows the need for original researchers to be explicit about their recruitment strategies, preferably providing advertisements verbatim. Second, replicators should be encouraged to consider using the original studies' recruitment materials to be a part of direct replication where possible. In cases where original recruitment materials cannot be recovered, several groups of participants may be recruited using different types of advertisements, with explicit tests of how this affects replicability. This should be informed by known moderators of the effect detected in previous literature (i.e., deliberately using recruitment methods that maximize the likelihood of obtaining an effect).

Two further considerations should be raised, both suggesting that simply adopting the original study's exact recruitment strategies would not be an ideal strategy. First, in cases where the original recruitment materials were clearly biased (e.g., Carnahan & McFarland, 2007, on the SPE), it is not clear that replicators should simply carry over the biased materials into their replication unless they are specifically interested in recreating a biased effect. Second, even assuming the replicator wishes to draw in a similar sample to the original study, viewing recruitment materials as a psychological stimulus raises the question of whether advertisements have psychometric equivalence (Fabrigar & Wegener, 2016) across populations. For instance, if an original study mentioned $10 cash payments, this may have a different meaning when recruiting participants from disadvantaged communities (where $10 may be substantial), or from a campus with wealthy students (where $10 might be trivial). Hence, blindly "copy-pasting" the recruitment methods of original research may be less advisable than thoughtfully probing what psychological traits were selected for by the original's methods, and then developing recruitment materials to maximize the same variable in the replicators' targeted population *if* this is desired on theoretical grounds. Examining the psychological relevance of recruitment strategies used in experiments may clarify the boundary conditions of phenomena, and guide understanding of cross-study or original/replication discrepancies.

Our results also contribute to the replication literature more broadly by showing yet another consideration that must be weighed when "directly," "exactly," or "closely" replicating an original study. Although it is comparatively common to consider how differences in measures, manipulations, populations, and contextual features may compromise the degree to which a replication matches an original

study (e.g., Gilbert et al., 2016; Stroebe & Strack, 2014), it is comparatively less common to consider recruitment strategies in the associated methodological literature (but see Brandt et al., 2014; Steiner et al., 2019), for exceptions), journal submission requirements, and actual replication practice. Yet even if the population targeted in original and replication studies are held constant on as many dimensions as possible (obviously perfect constancy would be impossible), our results show how even relatively natural alterations in recruitment materials (compare Figure 3A and B) change who actually participates. Even herculean efforts to hold populations constant can be undermined if recruitment methods re-introduce differences. In short, we provide clear empirical evidence to substantiate an underdiscussed challenge to replication.

Finally, our results may help to explain replication "failures." When a direct replication study obtains results different from an original study, psychologists consider how methodological differences may have caused the difference to occur. The term "hidden moderator" is sometimes employed but is conceptually open-ended because such explanations may suggest that the methodological change shifted the underlying effect itself (e.g., Gilbert et al., 2016; Kerr et al., 2018; Van Bavel et al., 2016) or that operationalizations may not have captured the same type of variance among the original and replication populations (i.e., "psychometric [in]variance"; e.g., Fabrigar & Wegener, 2016; Stroebe & Strack, 2014). Potentially, the sample-biasing effect we identify in the present work could be relevant to either/both explanations for non-replication. Insofar as subtle shifts in recruitment messages can draw different samples of people, those individual differences might directly shift the emergence of an effect (e.g., if experiential- versus reward-motivated people show an effect differently) or might alter the validity of the study's operationalizations (e.g., if a manipulation/measure has different meanings for experiential- versus reward-motivated people). Future research could examine each of these possibilities in detail, potentially targeting variables beyond the scope of the present work.

Limitations and Future Directions

We focused on more reward-focused versus experience-focused motivation in the present research, but follow-up research should test generalizability by examining how parallel effects could emerge for other recruitment strategy/psychological variable pairings. For instance, recruitment strategies that mention self-esteem might attract people high in narcissism (Baumeister & Vohs, 2001), and/or people who believe self-esteem is consequential (Vaughan-Johnston & Jacobson, 2020), since these traits relate to finding self-esteem unusually interesting (Vaughan-Johnston et al., 2022). Detecting such effects could have intriguing implications for the self-literature, for example. Furthermore, recruitment strategies that mention narratives might be expected to draw in participants high in the need for cognition, which could then plausibly shift replicability in the attitudes literature (see Cacioppo et al., 1996).

Beyond recruitment strategies' biasing the sampling of populations, other influences might be conjectured, such as recruitment strategies priming participants who participate shortly afterward. For instance, receiving a small gift in exchange for participation may elicit a positive mood (Isen et al., 1978) with diverse effects on psychological phenomena (Forgas & Koch, 2013). In short, the present focus on tweaking recruitment messages can be seen as a first step into examining the robust ways that recruitment strategies more broadly can subtly influence original and replication science.

Ultimately, we are not claiming that differences in recruitment strategies on their own are the major cause of many non-replications. More frequently, however, they might play a role alongside other sample-biasing factors toward shaping a study's effects. For example, a replication targeting a *somewhat* different population with a *somewhat* different recruitment strategy may end up with a *substantially* different sampling of participants. Our key point is that recruitment strategies are one of several heretofore underrecognized contributors to inter-study heterogeneity, and that a clearer understanding of their role and greater transparency regarding their usage will be a boon to original and replication scientists.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

Disclaimer

This manuscript has not been submitted to any other journal. Data for each sample is original to the present manuscript and has not been used elsewhere; all data were collected in conformity with ethical standards of the field.

ORCID iD

Thomas I. Vaughan-Johnston https://orcid.org/0000-0002-4682-481X

Supplemental material

Supplemental material is available online with this article.

Notes

1. We examined recruitment strategies from the 19 studies run via our department's summer psychology pool. In total, 14 (74%)

mentioned at least one piece of incentivizing information (credit, pay, fun/enjoyable, etc.).
2. We log-transformed completion time before conducting the t-test for this and Experiment 3 but report the untransformed participation times for interpretative ease.
3. Attention checks were not failed more often in either recruitment strategy condition, chi-square(1, $N = 390$) = .21, $p = .646$. However, attention checks failed at different rates across the two B&D spending conditions, chi-square(1, $N = 390$) = 4.90, $p = .027$, with a 11% failure rate in the charitable spending condition versus a 5% failure rate in the luxury spending condition. A likely explanation for this effect is that the luxury spending condition is simply more memorable than the charitable spending condition because the wealthy person leaving a lot of money to their dog (versus to a charity) is attention-grabbing and unusual. Without these cuts, we still replicate the same B&D effects. The moderation of B&D's attitude effect by recruitment strategy becomes $F(1, 379) = 2.35$, pone-tailed = .063, $\eta p^2 = .01$.
4. The same was not true for replicating Black and Davidai's effect of spending behavior on moral judgment (our Experiment 3), but only because this effect is very large in both of our conditions, and we calculate power of approximately 100% in either condition for this reason.

References

Amabile, T. M., Hill, K. G., Hennessey, B. A., & Tighe, E. M. (1994). The Work Preference Inventory: Assessing intrinsic and extrinsic motivational orientations. *Journal of Personality and Social Psychology*, *66*(5), 950–967.

Anderson, S. F., & Maxwell, S. E. (2017). Addressing the "replication crisis": Using original studies to design replication studies with appropriate statistical power. *Multivariate Behavioral Research*, *52*(3), 305–324.

Baumeister, R. F., & Vohs, K. D. (2001). Narcissism as addiction to esteem. *Psychological Inquiry*, *12*(4), 206–210.

Black, J. F., & Davidai, S. (2020). Do rich people "deserve" to be rich? Charitable giving, internal attributions of wealth, and judgments of economic deservingness. *Journal of Experimental Social Psychology*, *90*, 104011.

Brandt, M. J., IJzerman, H., Dijksterhuis, A., Farach, F. J., Geller, J., Giner-Sorolla, R., & Van't Veer, A. (2014). The replication recipe: What makes for a convincing replication? *Journal of Experimental Social Psychology*, *50*, 217–224.

Cacioppo, J. T., Petty, R. E., Feinstein, J. A., & Jarvis, W. B. G. (1996). Dispositional differences in cognitive motivation: The life and times of individuals varying in need for cognition. *Psychological Bulletin*, *119*(2), 197–253.

Callan, M. J., Shead, N. W., & Olson, J. M. (2011). Personal relative deprivation, delay discounting, and gambling. *Journal of Personality and Social Psychology*, *101*(5), 955–973.

Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T. H., Huber, J., Johannesson, M., Kirchler, M., Nave, G., Nosek, B. A., Pfeiffer, T., Altmejd, A., Buttrick, N., Chan, T., Chen, Y., Forsell, E., Gampa, A., Heikensten, E., Hummer, L., Imai, T., Isaksson, S., Manfredi, D., Rose, J., Wagenmakers, E. J., & Wu, H. (2018). Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. *Nature Human Behaviour*, *2*(9), 637–644.

Carnahan, T., & McFarland, S. (2007). Revisiting the Stanford Prison Experiment: Could participant self-selection have led to the cruelty? *Personality and Social Psychology Bulletin*, *33*(5), 603–614.

Cerasoli, C. P., Nicklin, J. M., & Ford, M. T. (2014). Intrinsic motivation and extrinsic incentives jointly predict performance: A 40-year meta-analysis. *Psychological Bulletin*, *140*(4), 980–1008.

Deci, E. L., Koestner, R., & Ryan, R. M. (1999). A meta-analytic review of experiments examining the effects of extrinsic rewards on intrinsic motivation. *Psychological Bulletin*, *125*(6), 627–668.

Dowling, K., Guhl, D., Klapper, D., Spann, M., Stich, L., & Yegoryan, N. (2020). Behavioral biases in marketing. *Journal of the Academy of Marketing Science*, *48*, 449–477.

Ebersole, C. R., Atherton, O. E., Belanger, A. L., Skulborstad, H. M., Allen, J. M., Banks, J. B., & Brown, E. R. (2016). Many Labs 3: Evaluating participant pool quality across the academic semester via replication. *Journal of Experimental Social Psychology*, *67*, 68–82.

Fabrigar, L. R., & Wegener, D. T. (2016). Conceptualizing and evaluating the replication of research results. *Journal of Experimental Social Psychology*, *66*, 68–80.

Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods*, *4*(3), 272–299.

Fabrigar, L. R., Wegener, D. T., & Petty, R. E. (2020). A validity-based framework for understanding replication in psychology. *Personality and Social Psychology Review*, *24*(4), 316–344.

Fabrigar, L. R., Wegener, D. T., Vaughan-Johnston, T. I., Wallace, L. E., & Petty, R. E. (2019). Designing and interpreting replication studies in psychological research. In F. Kardes, P. Herr, & N. Schwarz (Eds.), *Handbook of research methods in consumer psychology* (pp. 483–507). Routledge.

Fishbach, A., Shah, J. Y., & Kruglanski, A. W. (2004). Emotional transfer in goal systems. *Journal of Experimental Social Psychology*, *40*(6), 723–738.

Flake, J. K., Davidson, I. J., Wong, O., & Pek, J. (2022). Construct validity and the validity of replication studies: A systematic review. *American Psychologist*, *77*(4), 576.

Forgas, J. P., & Koch, A. S. (2013). Mood effects on cognition. In M. D. Robinson, E. Watkins, & E. Harmon-Jones (Eds.), *Handbook of cognition and emotion* (pp. 231–251). The Guilford Press.

Gilbert, D. T., King, G., Pettigrew, S., & Wilson, T. D. (2016). Comment on "estimating the reproducibility of psychological science." *Science*, *351*(6277), 1037–1037.

Gilbert, D. T., Pelham, B. W., & Krull, D. S. (1988). On cognitive busyness: When person perceivers meet persons perceived. *Journal of Personality and Social Psychology*, *54*(5), 733–740.

Gosling, S. D., Rentfrow, P. J., & Swann, W. B., Jr. (2003). A very brief measure of the Big-Five personality domains. *Journal of Research in Personality*, *37*(6), 504–528.

Hauser, D. J., Paolacci, G., & Chandler, J. (2019). Common concerns with MTurk as a participant pool: Evidence and solutions. In F. R. Kardes, P. M. Herr, & N. Schwarz (Eds.), *Handbook of research methods in consumer psychology* (pp. 319–337). Routledge.

Hauser, D. J., & Schwarz, N. (2015). It's a trap! Instructional manipulation checks prompt systematic thinking on "tricky" tasks. *SAGE Open*, *5*(2), 2158244015584617.

Hayes, A. F. (2006). A primer on multilevel modeling. *Human Communication Research*, *32*(4), 385–410.

Hidi, S., & Harackiewicz, J. M. (2000). Motivating the academically unmotivated: A critical issue for the 21st century. *Review of Educational Research*, *70*(2), 151–179.

Isen, A. M., Shalker, T. E., Clark, M., & Karp, L. (1978). Affect, accessibility of material in memory, and behavior: A cognitive loop? *Journal of Personality and Social Psychology*, *36*(1), 1–12.

Kahneman, D. (2011). *Thinking, fast and slow*. Farrar, Straus and Giroux.

Kerr, N. L., Ao, X., Hogg, M. A., & Zhang, J. (2018). Addressing replicability concerns via adversarial collaboration: Discovering hidden moderators of the minimal intergroup discrimination effect. *Journal of Experimental Social Psychology*, *78*, 66–76.

Kim, H., Callan, M. J., Gheorghiu, A. I., & Matthews, W. J. (2017). Social comparison, personal relative deprivation, and materialism. *British Journal of Social Psychology*, *56*(2), 373–392.

Klein, R. A., Vianello, M., Hasselman, F., Adams, B. G., Adams, R. B., Jr., Alper, S., & Batra, R. (2018). Many Labs 2: Investigating variation in replicability across samples and settings. *Advances in Methods and Practices in Psychological Science*, *1*(4), 443–490.

Knobe, J. (2003). Intentional action and side effects in ordinary language. *Analysis*, *63*(3), 190–194.

Kumar, A., & Gilovich, T. (2015). Some "thing" to talk about? Differential story utility from experiential and material purchases. *Personality and Social Psychology Bulletin*, *41*(10), 1320–1331.

Luttrell, A., Petty, R. E., & Xu, M. (2017). Replicating and fixing failed replications: The case of need for cognition and argument quality. *Journal of Experimental Social Psychology*, *69*, 178–183.

Miyamoto, Y., & Kitayama, S. (2002). Cultural variation in correspondence bias: The critical role of attitude diagnosticity of socially constrained behavior. *Journal of Personality and Social Psychology*, *83*(5), 1239–1248.

Oesch, J. M., & Murnighan, J. K. (2003). Egocentric perceptions of relationships, competence, and trustworthiness in salary allocation choices. *Social Justice Research*, *16*(1), 53–78.

Peetz, J., & Buehler, R. (2009). Is there a budget fallacy? The role of savings goals in the prediction of personal spending. *Personality and Social Psychology Bulletin*, *35*(12), 1579–1591.

Petty, R. E., & Cacioppo, J. T. (1986). The elaboration likelihood model of persuasion. *Advances in Experimental Social Psychology*, *19*, 123–205.

Pinheiro, J., Bates, D., & R Core Team. (2024). *nlme: Linear and nonlinear mixed effects models*. R package version 3.1-166. https://CRAN.R-project.org/package=nlme

Reis, H. T., & Judd, C. M. (2014). *Handbook of research methods in social and personality psychology* (2nd ed.). Cambridge University Press.

Ryan, R. M., & Deci, E. L. (2020). Intrinsic and extrinsic motivation from a self-determination theory perspective: Definitions, theory, practices, and future directions. *Contemporary Educational Psychology*, *61*, 101860.

Schwarz, N., & Clore, G. L. (2016). Evaluating psychological research requires more than attention to the N: A comment on Simonsohn's (2015) "small telescopes." *Psychological Science*, *27*(10), 1407–1409.

Simonsohn, U. (2015). Small telescopes: Detectability and the evaluation of replication results. *Psychological Science*, *26*(5), 559–569.

Srikant, M. (2013). Materialism in consumer behavior and marketing: A review. *Management & Marketing*, *8*(2), 329.

Steiner, P. M., Wong, V. C., & Anglin, K. (2019). A causal replication framework for designing and assessing replication efforts. *Open Science in Psychology*, *227*(4), 280–292.

Stroebe, W., & Strack, F. (2014). The alleged crisis and the illusion of exact replication. *Perspectives on Psychological Science*, *9*(1), 59–71.

Teeny, J. D., Siev, J. J., Briñol, P., & Petty, R. E. (2021). A review and conceptual framework for understanding personalized matching effects in persuasion. *Journal of Consumer Psychology*, *31*(2), 382–414.

Van Bavel, J. J., Mende-Siedlecki, P., Brady, W. J., & Reinero, D. A. (2016). Contextual sensitivity in scientific reproducibility. *Proceedings of the National Academy of Sciences*, *113*(23), 6454–6459.

van Dijk, E., & De Dreu, C. K. (2021). Experimental games and social decision making. *Annual Review of Psychology*, *72*, 415–438.

Vaughan-Johnston, T. I., Fowlie, D. I., & Jacobson, J. A. (2022). Facilitating scientific communication between strangers: A preregistered lost e-mail experiment. *Cyberpsychology, Behavior, and Social Networking*, *25*(7), 424–431.

Vaughan-Johnston, T. I., & Jacobson, J. A. (2020). "Need" personality constructs and preferences for different types of self-relevant feedback. *Personality and Individual Differences*, *154*, 109671.

Zwaan, R. A., Etz, A., Lucas, R. E., & Donnellan, M. B. (2018). Making replication mainstream. *Behavioral and Brain Sciences*, *41*, 1–61.