What have we learned about demand characteristics? A meta-analytic review of experimental manipulations of explicit hypothesis cues

Anonymous for peer review (NAC)1, Anonymous for peer review (MW)1, & Anonymous for peer review (MCF)2

1 Anonymous for peer review

2 Anonymous for peer review

Author note

Public significance statement: A fundamental methodological concern in research with human participants is that their responses are biased by information that convey the researcher’s hypothesis (i.e., demand characteristics). In a meta-analysis, we combined evidence from 40 studies that experimentally tested this concern by manipulating the hypothesis explicitly communicated to participants. Results suggested that explicit information about the researcher’s hypothesis produce biases in participants’ responses that can be potentially large – but are troublingly unreliable and difficult to explain. All materials, data (raw and processed), code, and pre-registrations (for Studies 1 and 2) are openly available at <https://osf.io/3hkre/?view_only=2dc92af53f194e5eab0d7aecafaf01c2>. This work was supported by the John Templeton Foundation (grant # anonymous for peer review). The funder had no role in the preparation of the manuscript or decision to publish. We thank (1) (anonymous for peer review; AC) for assistance with code review, and (2) (anonymous for peer review; JB) for assistance developing the initial literature search strategy.

The authors made the following contributions. Anonymous for peer review (NAC): Conceptualization, Data Curation, Formal Analysis, Investigation, Methodology, Project administration, Software, Supervision, Visualization, Writing - Original Draft Preparation, Writing - Review & Editing; Anonymous for peer review (MW): Data Curation, Investigation, Project administration, Software, Writing - Review & Editing; Anonymous for peer review (MCF): Formal Analysis, Investigation, Methodology, Project administration, Resources, Software, Supervision, Visualization, Writing - Review & Editing.

Correspondence concerning this article should be addressed to Anonymous for peer review (NAC), Anonymous for peer review. E-mail: Anonymous for peer review

Abstract

Demand characteristics are a fundamental methodological concern in experimental psychology. Yet, little is known about the direction, magnitude, consistency, and mechanisms underlying their effects. We conducted a three-level meta-analysis of 195 effect sizes from 40 studies that provided experimental tests of demand effects by manipulating the hypothesis explicitly communicated to participants. Results indicated that these explicit cues tend to produce small overall increases in hypothesis-consistent responding (*d* = 0.22, 95% CI [0.11, 0.33]). However, these effects were extremely heterogeneous (between-study = 0.31; within-study = 0.20), with the estimated distribution of true effects ranging from *d* = 1.98 (a massive increase in hypothesis-consistent responding) to *d* = -1.44 (a massive increase in hypothesis-*in*consistent responding). Contrary to conventional motivation accounts, we did not find evidence that demand effects were driven by post-hoc measures of participants’ motivation or opportunity to adjust their responses. We did, however, find robust evidence for accounts that emphasize the role of participants’ expectancies about the hypothesized effects. Similar findings emerged in a direct replication of one recent study included in the meta-analysis. Taken together, results underscore the importance – and challenges – of understanding and controlling for demand characteristics in experimental design.

*Keywords:* demand characteristics, expectancies, placebo, confounds, meta-analysis

What have we learned about demand characteristics? A meta-analytic review of experimental manipulations of explicit hypothesis cues

Imagine that one day a mysterious person approaches you and begins telling you about a new method they invented for understanding humans. They tell you that their method is useful for estimating causal relationships, but add that there is one issue: it can sometimes be thrown off by a *methodological artifact*. They explain that this artifact sometimes causes researchers to detect an effect that’s not real, and other times causes them to miss an effect that is real; that it sometimes biases estimates upward and other times downward. Then, they offer a confession: the artifact doesn’t always impact their conclusions, and they don’t know why. Sometimes the artifact seems to matter, other times it doesn’t – and its underlying mechanisms are poorly understood.

If the above scenario was real, the noted limitations would likely call their whole method into question. However, perhaps experimental psychologists should not be so quick to judge. After all, we too deal with a difficult-to-understand methodological artifact: *demand characteristics*.

In a seminal paper, Martin Orne (1962) argued that human subjects are perceptive to demand characteristics – “cues which convey an experimental hypothesis” – and generally use these cues to help the experimenter confirm their hypothesis (1962, p. 779). Orne initially presented evidence that demand characteristics can lead to false positives, such as patients exhibiting sham symptoms of hypnosis (Orne, 1959). However, demand characteristics can also lead to false negatives. For example, participants will ignore visual cues of depth when they believe that disregarding them is the purpose of the experiment (Hayes & King, 1967). In addition to creating inferential errors, demand characteristics can bias estimates of causal relationships. For example, the effects of facial poses on self-reported emotion can be amplified *or* attenuated depending on whether the experimenter communicates expectations of positive or nil effects (Coles, Gaertner, Frohlich, Larsen, & Basnight-Brown, 2022). Puzzlingly, though, demand characteristics do not always seem to matter. For example, in a set of large replications of classic studies in behavioral economics, direct manipulations of demand characteristics consistently failed to impact participants’ responses (Mummolo & Peterson, 2019).

As this brief review shows, demand characteristics are uncomfortably close to the mysterious methodological artifact described in the opening of the paper. Demand effects are a literal textbook methodological concern in experimental psychology (Sharpe & Whelton, 2016). However, their magnitude, direction, and consistency remain unclear. In the present paper, we use meta-analysis t/o take stock of what we have learned – if anything – about demand effects. We begin by briefly reviewing one of the most comprehensive and influential framework describing their effects.

## How do demand characteristics alter participant responses?

One of the most influential frameworks for conceptualizing demand effects was developed by Rosnow and colleagues (Rosnow & Aiken, 1973; Rosnow & Rosenthal, 1997; Strohmetz, 2008). In this framework, they unified decades of previous theorizing by positing that demand effects had three key moderators: (1) receptivity to cues, (2) motivation to provide hypothesis-consistent responses, and (3) opportunity to alter responses.

To start, Rosnow and colleagues reasoned that participants must be receptive to demand characteristics for there to be subsequent shifts in participants’ responses (see also, Orne, 1958). As an extreme example, imagine that a researcher hands an infant a sheet of paper that precisely explains the study hypothesis. Demand characteristics are certainly present, but they are not predicted to have an impact because the infant is not receptive to the cues. Even if the infant possessed the astonishing ability to read, it’s possible they would misinterpret the cues – which may be considered another form of non-receptivity (Corneille & Lush, 2022). We mention this receptivity moderator for the sake of comprehensiveness, but will not discuss it further.

If and when participants correctly interpret demand characteristics, Rosnow and colleagues theorized that subsequent changes in participants’ responses would be driven by their motivation (or lack thereof) to provide hypothesis-consistent responses. Early work on demand characteristics was marked by debates about whether participants are motivated to adjust their responses to (a) help the researcher confirm their hypothesis (Orne, 1962), (b) receive positive evaluations (Riecken, 1962; Rosenberg, 1969; Sigall, Aronson, & Van Hoose, 1970), (c) interfere with the purpose of the study (Cook et al., 1970; Masling, 1966), or (d) follow directions as closely as possible (Fillenbaun & Frey, 1970). Rosnow and colleagues advanced this line of thinking by demonstrating that participants have *multiple* shifting motivations in mind when they conceptualize their roles as subjects [Rosnow and Rosenthal (1997); see also Silverman, 1965]. For example, participants appear to be motivated to increase performance on simple tasks when told that this is the experimenter’s expectation – but not when the experimenter adds that the increase in performance will be indicative of a negative personality trait (Sigall et al., 1970). Rosnow and colleagues, thus, suggested that participants in any given context can be characterized as being overall motivated to either: (a) non-acquiesce (i.e., not change their responses based on knowledge about the hypothesis), (b) acquiesce (i.e., provide hypothesis-consistent responses), or (c) counter-acquiesce (i.e., provide hypothesis-inconsistent responses).

If participants are motivated to adjust their response, Rosnow and colleagues theorized that subsequent changes in participants’ responses would then be driven by their ability to alter the outcome of interest. As elaborated by Corneille and Lush (2022), this could occur through faking, imagination, or phenomenological control (voluntary changes experienced by the participant as involuntary).Taking this third moderator – opportunity – into account, Rosnow and colleagues concluded that demand characteristics only bias responses when participants (1) notice the cues, (2) are motivated to adjust their responses, and (3) can adjust their responses. This framework directly maps onto psychologists’ playbook for avoiding the impact of demand characteristics: use deception (reduce receptivity), incentivize honest reporting (reduce motivation), and/or deploy difficult-to-control outcome measures (reduce opportunity to adjust responses).

## When do demand characteristics impact participant responses?

Motivation

* Student vs. non-student
* Payment vs. no payment
  + Berinsky et al. (2012): Demand are particularly problematic for MTurk participants becuase they want compensation
  + de Quidt, Hauschofer, and Ruth (2017): No evidence that incentives increase demand effects
  + Orne (1969): “…effect may be reversed if payment is involved”

Opportunity

* Within vs. between
  + “Demand effects are likely to be stronger in a within designs”. Better able to ascertain demand effects. They know w

Disagreements about whether payment matters

# Methodology

The goal of the current paper is to take stock of what we have learned – if anything – about demand characteristics as a methodological artifact. Although several excellent *narrative reviews* exist (Corneille & Lush, 2022; Rosnow & Rosenthal, 1997; Sharpe & Whelton, 2016; Strohmetz, 2008), meta-analysis allows us to quantitatively evaluate the magnitude, consistency, and potential moderators of demand effects.

We defined the scope of the meta-analysis using the Population, Intervention, Comparison, Outcome framework (Schardt, Adams, Owens, Keitz, & Fontelo, 2007). Our population-of-interest was human subjects participating in non-clinical studies. We excluded clinical studies so that we could focus on research that better isolates the discipline (experimental psychology) discussed in the demand characteristics literature. Given that there is a sizable literature and number of reviews on conceptually-related placebo effects, excluding clinical studies also improved the feasibility and reduced the redundancy of our work.

Notably, the intervention-of-interest was explicit manipulations of the hypothesis communicated to participants – i.e., scenarios where a researcher tells participants about the effect of an independent variable on a dependent variable. Demand characteristics are sometimes defined as *any* cue that may impact participants’ beliefs about the purpose of the study, including instructions, rumors, and experimenter behavior (Orne, 1962). However, such a definition creates a potentially boundless conceptual space where *any* systematic change in a research design might be considered a test of demand characteristics. To bound and simplify the conceptual space, we focused on explicit manipulations of the hypothesis communicated to participants.

Our comparison-of-interest were conditions where either no hypothesis or a different hypothesis was communicated to participants. Our outcome-of-interest was the dependent variable described in the communicated hypothesis. For example, in a study that manipulated whether the intervention is described as “mood-boosting”, the outcome-of-interest would be any measure of mood.

### Literature search.

Our literature search strategy was initially developed in consultation with a librarian at (anonymous for peer review) and then later expanded based on reviewer feedback.

On January 12, 2022, we searched APA PsycInfo using broad search terms: “demand characteristics” OR “hypothesis awareness”. On April 17, 2024, we repeated the search to identify records published after the initial search. At that time, we also expanded the search to include conceptually similar terms found in the appendix of Rosnow and Rosenthal (1997) ’s book on experimental artifacts: “participant role” OR “demand effects” OR “good subject effect” OR “expectancy effect” OR “evaluative apprehension”. We also released a call for unpublished studies on the Society for Personality and Social Psychology Open Forum; Twitter; the Facebook Psychological Methods Discussion group; and the Facebook PsychMAP group.

Our search did not have language restrictions and went as far back as 1840, which yielded 1289 published and 168 records.

### Screening.

To be eligible for inclusion in the meta-analysis, the following criteria must have been met:

* The researcher manipulated what participants were told about the effect of an independent variable on a dependent variable. This criterion included both *positive demand* (participants told that the dependent variable will increase), *negative demand* (participants told that the dependent variable will decrease) and *nil demand* (participants told the dependent variable will be unaffected) conditions. Often, these conditions were compared to a *control* condition, where participants were not told about an effect of an independent variable on a dependent variable.[[1]](#footnote-24) In most cases, the effect of the independent variable was described explicitly, but there were some included studies where it was strongly implied.
* The demand characteristics manipulation was not strongly confounded with another manipulation. For example, we excluded a study by Sigall et al. (1970) because the manipulation of the stated hypothesis was confounded with a disclosure about the meaning of the behavior (i.e., that confirming the hypothesis would be indicative of an obsessive-compulsive personality disorder).
* A non-clinical population was studied.[[2]](#footnote-25)
* Information necessary for computing at least one effect size was included.

N. C. and M. W. screened records independently, reviewed potentially relevant records together, and worked together to code the information for moderator analyses and effect size computations. Disagreements were resolved through discussion. N. C. also coded the quality of each record included in the final synthesis (described later). In total, 54 studies from 39 records were eligible for inclusion. However, one record (Allen & Smith, 2012) was removed because the information provided led to implausibly large effect size estimates (e.g., = -209.16).

### Effect size index.

We used standardized mean difference scores with small-sample correction (Hedge’s ) as our effect size index (Borenstein, 2009; Cohen, 2013).

In most scenarios, we estimated the main effect of explicit demand characteristics. For example, Coles et al. (2022) manipulated whether participants were told that posing smiles would increase happiness. Here, the main effect of explicit demand characteristics can be computed by comparing happiness ratings from smiling participants who were either informed or not informed of the mood-boosting effect of smiling. In other scenarios, we estimated the *interactive* effect of explicit demand characteristics. For example, in the same Coles et al. (2022) study, participants provided happiness ratings both after smiling and scowling. Participants’ mood generally improved when smiling vs. scowling (i.e., there was a main effect of facial pose). However, the difference was more pronounced when participants were told about the mood-boosting effects of smiling. In other words, there was an interaction between facial pose and explicit demand characteristics. In this scenario, the interactive effect of explicit demand characteristics was computed by calculating a standardized difference-in-differences score.

Effect sizes were calculated so that positive values indicated an effect consistent with the communicated hypothesis. For example, if participants were told that an intervention should be mood boosting, an increase in mood would be coded as a positive effect. If, however, participants were told that the intervention should be mood *dampening*, that same increase in mood would be coded as a negative effect.

We calculated Hedge’s by applying a small sample correction to Cohen’s (for between-subject designs) and (for within-subject designs) estimates. Whenever possible, we used the *M*’s and *SD*’s reported in a paper to compute Cohen’s *d*. If these values were not reported, we used (in order of preference), (1) *t*-values, (2) descriptive statistics extracted from figures (e.g, bar charts) using the WebPlotDigitizer (Drevon, Fursa, & Malcolm, 2017), (3) *F*-values, or (4) *p*-values. In instances where this information was not provided but the significance and direction of the effect was described, we assumed *p*-values of .04 and .50 for significant and non-significant effects respectively (e.g., Kenealy, 1988). In a few instances, the outcome variable in a study was discrete (as opposed to continuous). In these cases, we approximated a Cohen’s *d* score based on a transformation of the log odds ratio (Borenstein, Hedges, Higgins, & Rothstein, 2011).

For repeated-measure comparisons, the correlation between the repeated measures is needed to calculate Cohen’s . This correlation is rarely reported, so we followed a recommendation by Borenstein (2009) and performed sensitivity analyses on an assumed correlation. We preregistered a default correlation of = .50 but performed sensitivity analysis with = .10, .30, .50, .70, and .90. These sensitivity analyses produced virtually no change in overall effect size estimates, so we do not discuss them further.

Nearly all studies (74%) contained multiple effect sizes of interest. For example, the full design in Coles et al. (2022) included a positive demand, nil demand, and control condition. Participants also completed several facial expression poses (happy, angry, and neutral) and self-reported several emotions (happiness and anger). To be comprehensive, we recorded all reported effect sizes and accounted for dependencies using three-level meta-analysis (described later).

### Potential study feature moderators.

Below, we describe study feature that may help explain variability in demand effects:

* *Control vs. non-control group comparison group.* Demand effects should presumably be additive. For example, imagine a study where the effect of a task is either (a) not described at all (a control condition), (b) described as mood-boosting (positive demand) or (c) described as mood-dampening (negative demand). Compared to the control condition, mood is typically predicted to be boosted in the positive demand condition and dampened in the negative demand condition. If this is the case, the mean difference in mood should be larger when the positive demand condition is compared to the negative demand condition (as opposed to the control condition). To test this, we coded whether comparisons were made to a control group or a different demand condition.
* *Positive, negative, or nil demand manipulation.* Instances where a demand characteristic condition was compared to a control group also allowed us to test whether participants responses shift more when the researcher hypothesizes an increase (positive demand), a decrease (negative demand), or no change in the dependent variable (nil demand).
* *Participant pool.* Whether students, non-students (e.g., MTurk workers), or a mix of students and non-students were sampled.
* *Setting.* Whether the study was conducted online or in-person.
* *Study design.* Whether demand characteristics were manipulation within- vs. between-subjects.
* *Payment.* Whether participants were paid or unpaid.

### Vignette ratings of potential theory-based moderators.

During our literature review, we found very few papers that attempted to quantify or test the theoretical moderators described by Rosnow and colleagues. Intriguingly, though, Orne (1969) suggested that participants may be particularly adept at explaining demand characteristics. We initially shared this optimism and evaluated not only their ability to predict demand effects, but also the predictive validity of their ratings of theory-relevant moderators.

For each study included in the meta-analysis after our original literature search[[3]](#footnote-29), we created vignettes that described the key details for each demand characteristic condition and dependent variable combination. For example, Standing, Verpaelst, and Ulmer (2008) had two demand characteristic manipulations (positive and negative demand) and two dependent variables (measures of verbal and spatial reasoning). Thus, we created four vignettes for this study,

In total, there were 119 vignettes. We did not create vignettes for control conditions because participants were not given information about the experimenter’s hypothesis. Because there were no explicit demand characteristics to act upon, we left motivation, belief, and opportunity values blank for this condition.

#### Vignette rating methodology.

Using a web-based Qualtrics survey, a convenience sample of raters reviewed 10 randomly selected vignettes. For each vignette, raters were asked to first identify the researcher’s hypothesis. Here, participants chose between four options that described a filler effect (usually involving an irrelevant dependent variable) or a positive, negative, or nil effect of the independent variable on the dependent variable. Afterwards, they rated the extent to which they would hypothetically (1) be motivated to adjust responses based on the researcher’s stated hypothesis (-3 = “extremely motivated to adjust responses to be inconsistent” to 3 = “extremely motivated to adjust responses to be consistent”), and (2) be able to adjust their responses on the outcome-of-interest (0 = “extremely incapable” to 4 = “extremely capable). To evaluate the potential role of beliefs (as in placebo effects), raters also indicated the extent to which they would expect the hypothesized effect to occur (-3 =”strong disbelief” to 3 = “strong belief”). Last, raters indicated the extent to which they expected other participants to adjust their responses to confirm the hypothesized effect (-3 = “extremely likely to adjust responses to be *inconsistent*” to 3 “extremely likely to adjust responses to be consistent).

Sample size was based on availability of resources. We initially collected as much data as possible (n = 192)in a single semester from undergraduates from (anonymous for peer review). Following a reviewer recommendation, we examined the reliability of those vignette ratings in a post-hoc manner. To do so, we calculated the intraclass correlation using mixed effects models. For ratings of motivation, opportunity, belief, and predictions about whether other participants would shift responses, we used the lme4 package (CITE) to fit an intercept only mixed effect model with rater and vignette as random intercepts. We then used the performance package (CITE) to calculate the intraclass correlation for the rater random intercept, which we used as an indicator of agreement between raters. The intraclass coefficient for motivation (0.23), opportunity (0.21), belief (0.14), and predicted responses from other participants (0.21) was low.

The estimated intraclass correlation from our original sample suggested that external raters do not generally agree about how other participants will respond to demand characteristics. Nonetheless, the Law of Large Numbers stipulates that these relatively imprecise ratings should converge into relatively precise estimates of the true mean at larger samples. We attempted to exploit this statistical tendency by collecting additional ratings from Prolific workers. This left us with a total of 412 raters (55.00% female; 1% male, 1% did not report gender; 44% transgender or gender non-conforming). 11% of participants reported they were White/Caucasian, 1% Asian, 1% Black/African American, and 4% Native Hawaiian or Other Pacific Islander. 0% of participants indicated that their ethnicity could not be described by any single provided category, and 0% did not report ethnicity. The average participant age was 30.10 ( = 13.82).

For each record included in the meta-analysis, we (a) removed observations where the rater did not respond or did not correctly identify the hypothesis communicated in the vignette, and then (b) averaged across raters.

##### Accounting for different demand comparisons.

As mentioned before, Hedge’s represents the standardized difference between *two* groups. Thus, for each effect size estimate, we summed the motivation, opportunity, and belief ratings for the two groups being compared. Doing so allowed us to accommodate the fact that some comparisons involved two demand characteristics conditions. For example, imagine a study where participants are told a procedure will boost mood (positive demand), told a procedure will dampen mood (negative demand), or not told about an expected effect (control). Compared to a control condition, participants who are motivated to confirm the hypothesis are theorized to have upward-biased responses in the positive demand condition and downward-biased responses in the negative demand condition. If those demand conditions are compared to each other – instead of a control condition – their effects will be additive. Summing motivation scores allowed us to accommodate this possibility, and we used the same approach for opportunity and expectation ratings.

We did not include nil-hypothesis comparisons in our analyses because our coding strategy could not accommodate the potential moderating role of motivation and belief in these conditions. For example, imagine that a participant is (a) told that an intervention will not impact mood (nil demand), and (b) is motivated to disconfirm the hypothesis. Relative to a control condition, this participant could disconfirm the hypothesis by either increasing *or* decreasing their mood report. Thus, even if motivation does moderate the effects of demand characteristics, we would not expect a systematic pattern to emerge with our coding scheme.

### Quality ratings.

Following a reviewer recommendation, we coded the quality of each record included in the meta-analysis. To do so, we used a modified version of the DOWNS AND BLACK (1998) checklist. This checklist contains 10 items designed to evaluate the reporting quality (e.g., “Are the main findings of the study clearly described?”), 3 to evaluate external validity (e.g., “Were the subjects asked to participate in the study representative of the entire population from which they were recruited?”), 7 to evaluate internal validity (e.g., ” Was an attempt made to blind those measuring the meain outcomes of the intervention?), 6 to evaluate selection bias (e.g., “Were losses of patients to follow-up taken into account?”), and 1 to evaluate statistical power.

Many of the items in the Downs and Black (1998) were difficult to evaluate or inapplicable to the literature we evaluated. We thus focused our analysis on 7 reporting quality, 1 external validity, and 3 internal validity items. Each item was coded as either a 1 (“yes”), 0 (“no”) or NA (“not applicable”). For each category, the scores were averaged.

[FOOTNOTE: Although, Downs and Black (1998) has been recommended as a measure of quality or records included in a meta-analysis, we do not personally agree with the use of the tool. Many items were not applicable to the work we were evaluating (e.g., whether distributions of principle confounders are described). Many other were difficult to evaluate (e.g., whether the study had adequate statistical power)

### Meta-analytic approach.

For our meta-analytic approach, we used three-level meta-analysis (3LMA; also referred to as “multilevel” meta-analysis). Rather than assume that there is a single true effect of demand characteristics, 3LMA assumes that there is a distribution containing *multiple true effects*. To separate variability in these true effects from mere sampling error, 3LMA models three sources of variability: sampling error of individual studies (level 1), variability within studies (level 2), and variability between studies (level 3; often referred to as “random effects”).

We fit all models using the metafor package (Viechtbauer, 2010) in R (version 4.1.2, R Core Team, 2021). We weighed effect sizes based on their inverse-variance and used cluster-robust methods for estimating variance-covariance matrices (Pustejovsky & Tipton, 2018). To estimate the overall effect size, we fit an intercept-only 3LMA model. Unless otherwise specified, we conducted moderator analyses by separately entering variables (dummy-coded or mean-centered for categorical and continuous moderators respectively) factors into a new model. Doing so allowed us to avoid issues of collinearity. To test the significance of each moderators, we used model comparison *F*-tests. To estimate effect sizes within each subgroup of the moderator, we used model-derived estimates.

#### Publication bias analyses.

Publication bias refers to the well-documented propensity for hypothesis-inconsistent findings to be disproportionately omitted from the published scientific record (Franco, Malhotra, & Simonovits, 2014). When present, publication bias can lead to inaccurate effect size estimates and inferential errors in meta-analysis. Consequently, we used three main approaches for assessing and correcting for potential publication bias in our estimation of the overall effect of demand characteristics.

First, we visually examined *funnel plots,* wherein observed effect sizes are plotted against a measure of their precision (e.g., standard error). In the absence of publication bias, the distribution typically resembles a funnel; relatively large studies estimate the effect with high precision, and effect sizes fan out in *both* directions as the studies become smaller. If, however, non-significant findings are disproportionately omitted from the scientific record (i.e., there is publication bias), the distribution is often asymmetric/sloped. Funnel plots traditionally contain one effect size per study, but many of our studies included multiple relevant effect sizes. Thus, we examined two funnel plots: one with all effect sizes and one with the dependent effect sizes aggregated[[4]](#footnote-34).

Second, we conducted precision-effect tests (Stanley & Doucouliagos, 2014). In precision-effect tests, the relationship between observed effect sizes and their standard errors – which is typically absent when there is no publication bias – is estimated and controlled for in a meta-regression model. The slope of this model is often interpreted as an estimate of publication bias, and the intercept is often interpreted as the bias-corrected overall effect. These precision-effect tests were developed and validated for meta-analyses with independent effect sizes. Nonetheless, Rodgers and Pustejovsky (2021) demonstrated that the method retains fairly good statistical properties when (1) 3LMA is used or (2) dependent effect sizes are aggregated and modeled using random-effects (i.e., two level) meta-regression. We used both approaches[[5]](#footnote-35).

Third, we deployed weight-function modeling using the weightR package (Coburn & Vevea, 2019). In weight-function modeling, weighted distribution theory is used to model biased selection based on the significance of observed effects (Vevea & Hedges, 1995). If the adjusted model provides increased fit, publication bias is a concern and the model can be used to estimate the bias-corrected overall effect size. Once again, weight-function modeling was designed for independent effect sizes. Nonetheless, it has fairly good statistical properties when non-independent effect sizes are aggregated, which we did here (Rodgers & Pustejovsky, 2021).

As a sensitivity analysis, we included publication status (published or unpublished) as a dummy-coded moderator. This moderator allowed us to estimate the difference in the magnitude of published vs. unpublished effects.

### Transparency and openness.

All materials, data, and code are openly available at <https://osf.io/3hkre/?view_only=2dc92af53f194e5eab0d7aecafaf01c2>. This link also contains the pre-registration plan and documented amendments.

For the meta-analysis, sample size was determined by the availability of relevant records (i.e., not via power analysis). For the vignette ratings, sample size was initially determined by the availability of resources (i.e., we collected as much data as possible in a single semester). However, our second wave of rater recruitment considered precision – and was designed to decrease the length of the confidence intervals to 1 for the motivation, opportunity, belief, and prediction ratings.

Ethics approval was not initially requested for the meta-analysis because no new data were collected. However, the effort to collect ratings of motivation, opportunity, belief, and prediction ratings was reviewed and approved by the (anonymous for peer review) IRB (protocol #: anonymous for peer review; protocol title: anonymous for peer review).

All code, including the script used to generate a computationally reproducible manuscript using the papaja R package (Aust & Barth, 2022), has been checked for reproducibility.

## Results

In total, we extracted 253 effect sizes from 53 studies from between the years 1964 and 2024. The average coder rating of reporting quality, internal validity, and external validity was 0.72, 0.67, 0.67, 0.67, 0.67, 0.67, 0.67, 0.67, 0.67, 0.67, 0.67, 0.67, 0.67, 0.67, 0.67, 0.67, 0.67, 0.67, 0.67, 0.67, 0.67, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0.67, 0.67, 0.50, 0.50, 0.50, 0.50, 0.50, 0.50, 0.50, 0.50, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0.67, 0.67, 0.67, 0.67, 0.67, 0.67, 0.67, 0.67, 0.67, 0.67, 0.67, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0.50, 0.50, 0.50, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0.75, 0.75, 0.75, 1, 0.67, 0.75, 1, 1, 1, 0.50, 0.50, 0.50, 0.50, 0.50, 0.50, 0.50, 0.50, 0.50, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, and 1, and 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, and 0 out of 1, respectively. The low external validity scores were driven by the reliance on non-representative sampling methods, a common practice in experimental psychology that is nonetheless identified as a threat to external validity in the DOWNS AND BLACK checklist.

Overall, results indicated that explicit manipulations of demand characteristics cause participants’ responses to shift in a manner consistent with the communicated hypothesis, = 0.20, 95% CI [0.11, 0.30], = 4.28, < .001. As a hypothetical example, if participants were told that the researcher hypothesizes that an intervention will improve mood (positive demand), they would generally report slightly improved moods; if told that the researcher hypothesizes that an intervention will worsen mood (negative demand), they would generally report slightly worsened moods.

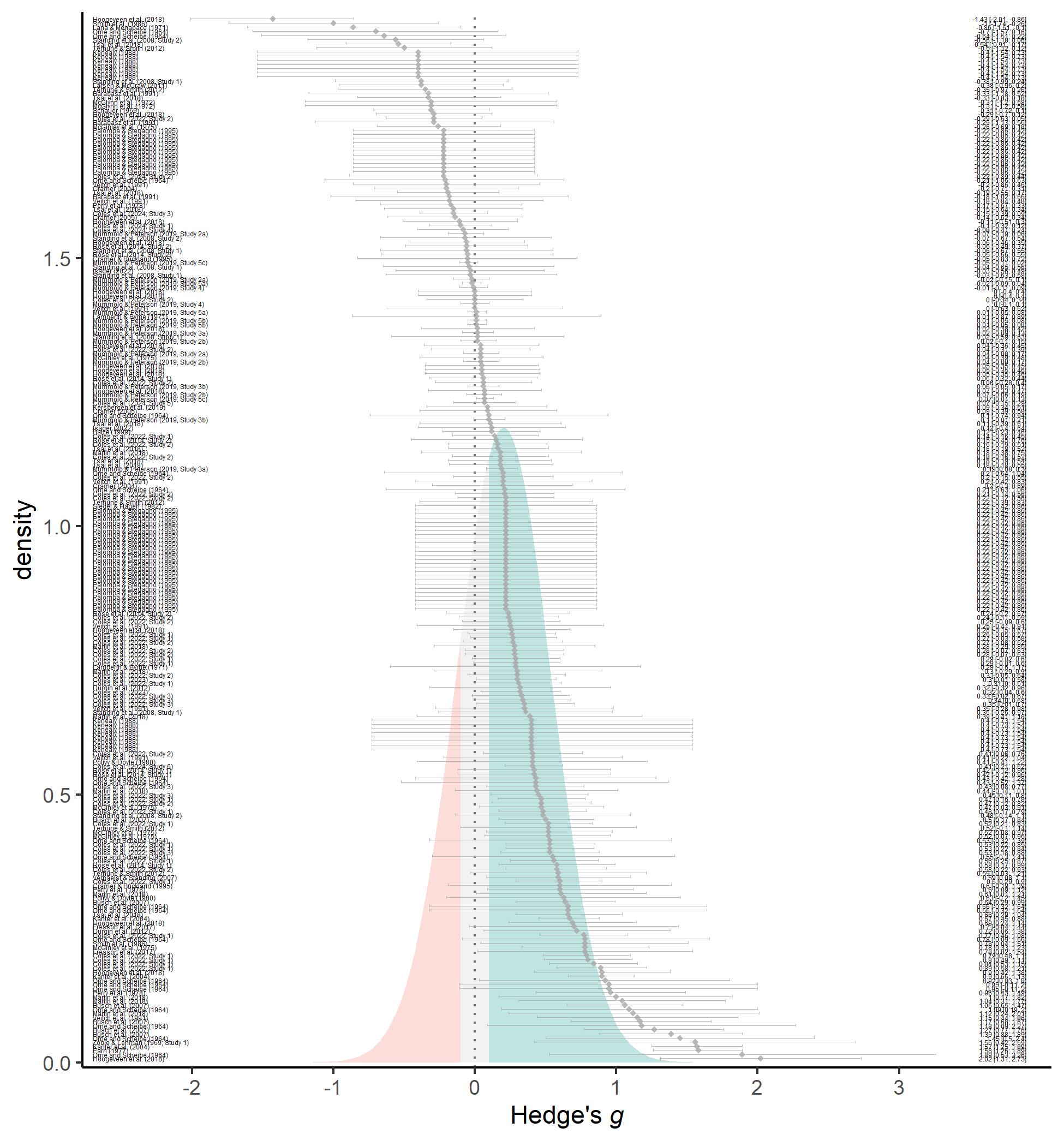


Figure 1: Forest plot of effect sizes (grey diamonds), their 95% confidence intervals (grey error bars), and their citations (left; see openly-available data for more information). The estimated effect size distribution is also shown and colored based on whether demand characteristics produce more hypothesis-consistent responding (green; g > 0.10), more hypothesis-inconsistent responding (red; g < -0.10), or negligible shifts in responding (grey; |g| < 0.10).

As a reminder, rather than assume that there is a single true effect of demand characteristics, 3LMA assumes that there are *multiple true effects.* Consistent with this assumption, observed variability in demand effects drastically exceeded what would be expected from sampling error alone (between-study = 0.29; within-study = 0.18; (252) = 978.70, < .001). 3LMA assumes that the multiple true effects form a normal distribution, which we recreated based on estimates of the average effect size and variability attributed to sources other than sampling error (between-study + within-study ). As shown in Figure 1, this estimated distribution suggests that demand effects can range from approximately = -1.33 to = 1.82 — covering the range of most conceivable effects in psychology. As a heuristic, we arbitrarily classified any effect size less than 0.10 standard deviation in either direction as “negligible”. Based on this classification, the recreated distribution suggested that demand characteristics most often produce hypothesis-consistent shifts (62%), but sometimes produce negligible shifts (20%) or shifts in the *opposite* direction of the communicated hypothesis (18%).

### Moderator analyses.

Table

| Moderator (bolded) and level | s | k | g | B1 | 95% CI | F | p |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Student Status | 53 | 253 | – | – | – | 2.44 | .282 |
| yes | 37 | 161 | 0.26 | – | [0.13, 0.4] | 15.99 | < .001 |
| no | 13 | 27 | 0.05 | – | [-0.06, 0.16] | 1.08 | .323 |
| mix | 3 | 20 | 0.05 | – | [-1, 1.09] | 0.3 | .680 |
| Payment Status | 53 | 253 | – | – | – | 0.55 | .465 |
| yes | 14 | 51 | 0.13 | – | [0, 0.26] | 4.92 | .047 |
| no | 38 | 164 | 0.19 | – | [0.08, 0.31] | 11.71 | .002 |
| Setting | 53 | 253 | – | – | – | 4.81 | .036 |
| online | 19 | 53 | 0.1 | – | [0.01, 0.19] | 5.68 | .030 |
| in-person | 34 | 189 | 0.29 | – | [0.14, 0.44] | 15.49 | < .001 |
| Design | 53 | 253 | – | – | – | 1.55 | .240 |
| within-subjects | 14 | 44 | 0.14 | – | [0.02, 0.25] | 7.07 | .020 |
| between-subjects | 45 | 209 | 0.22 | – | [0.11, 0.34] | 14.65 | < .001 |
| Comparison type | 53 | 253 | – | – | – | 1.93 | .287 |
| positive vs. control | 42 | 115 | 0.15 | – | [0.04, 0.26] | 7.14 | .011 |
| nil vs. control | 4 | 17 | 0.22 | – | [-0.14, 0.58] | 2.91 | .164 |
| nil vs. control | 4 | 17 | 0.22 | – | [-0.14, 0.58] | 2.91 | .164 |
| negative vs. control | 17 | 43 | 0.16 | – | [0.03, 0.29] | 6.4 | .021 |
| positive vs. nil | 8 | 34 | 0.36 | – | [0.02, 0.71] | 6.13 | .043 |
| positive vs. negative | 16 | 44 | 0.32 | – | [0.15, 0.5] | 15.15 | .001 |
| Publication status | 53 | 253 | – | – | – | 0.07 | .801 |
| published | 42 | 240 | 0.21 | – | [0.11, 0.31] | 18.98 | < .001 |
| unpublished | 11 | 13 | 0.17 | – | [-0.17, 0.51] | 1.26 | .289 |
| motivation to adjust responses | 48 | 202 | – | 0.01 | [-0.18, 0.19] | 0.00359794732820695 | .953 |
| opportunity to adjust responses | 48 | 202 | – | 0.04 | [-0.01, 0.08] | 3.72775668632133 | .080 |
| belief in communicated hypothesis | 48 | 202 | – | 0.06 | [-0.05, 0.17] | 1.45953104480182 | .254 |

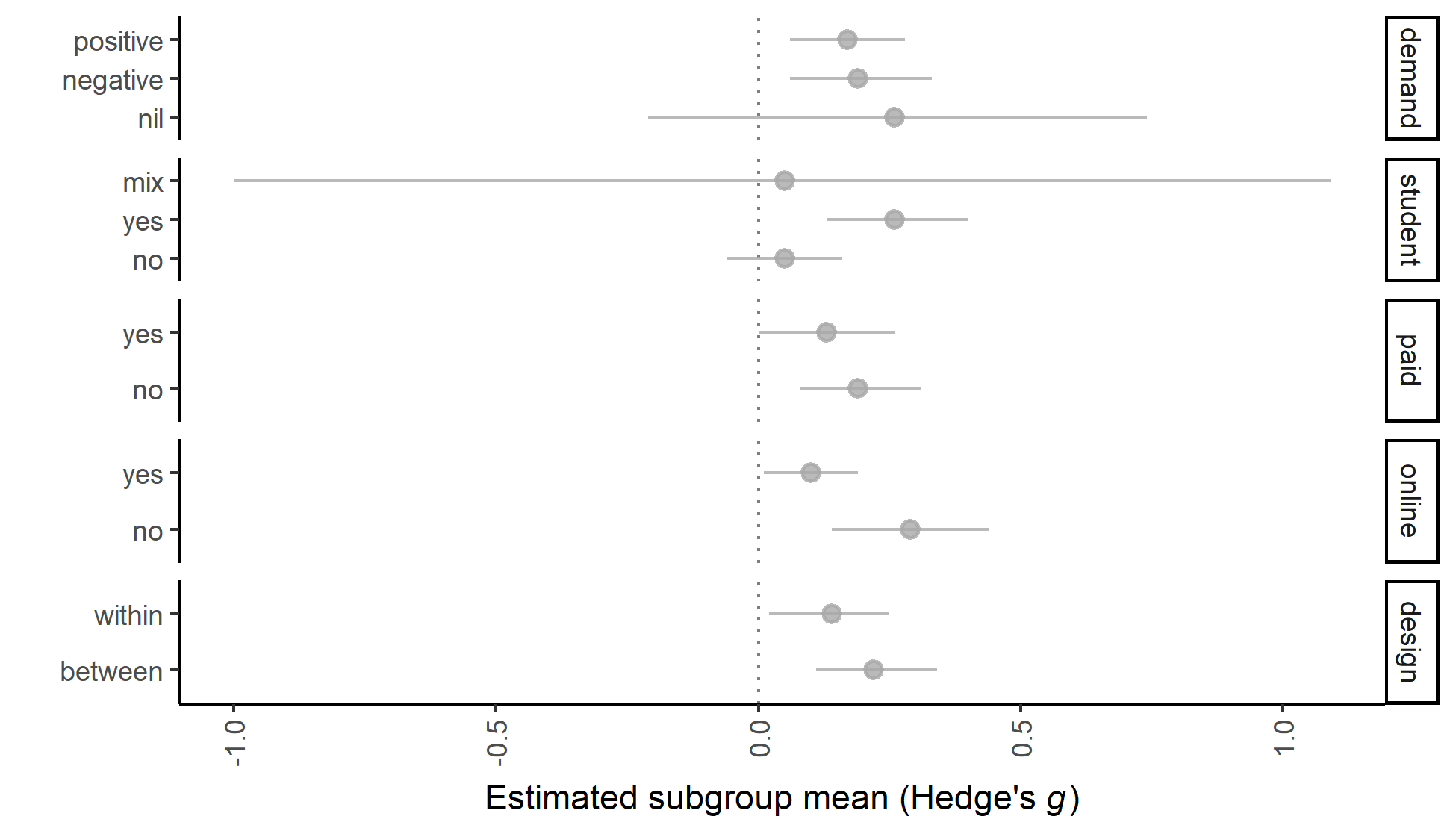


Figure 2: For each moderator, we fit a separate 3LMA and subsequently derived subgroup mean effect sizes (dots) and their 95% confidence intervals (error bars).

When variability in effect sizes exceeds what would be expected from sampling error alone, it suggests the presence of moderators. Below, we examine potential moderators

#### Study features.

Results indicated that the effects of demand characteristics tended to differ by participant pool, (2, 2.12) = 2.44, = .282. As shown in Figure 2, effects were generally positive and medium-to-large in studies with students ( = 0.26, 95% CI [0.13, 0.40], < .001), and near-zero in studies with non-students ( = 0.05, 95% CI [-0.06, 0.16], = .323) or a mix of students and non-students ( = 0.05, 95% CI [-1.00, 1.09], = .680). The effects of demand characteristics also tended to be slightly more positive for in-person ( = 0.29, 95% CI [0.14, 0.44], < .001) vs. online ( = 0.10, 95% CI [0.01, 0.19], = .030) studies; however, this difference was not significant, (1, 30.22) = 4.81, = .036 (Figure 2).

The effects of demand characteristics were additive. Compared to instances where a demand characteristic condition was compared to a control group ( = 0.15, 95% CI [0.07, 0.24], = .001), effect sizes were approximately twice as large when two demand characteristic conditions were compared ( = 0.33, 95% CI [0.18, 0.48], < .001), (1, 10.37) = 10.87, .008. Instances where a demand characteristic condition was compared to a control group allowed us to additionally test whether participants respond more strongly to positive, nil, or negative demand characteristics. Results indicated that they do, (2, 4.15) = 0.20, = .828. As shown in Figure 2, the effect of demand characteristics tended to be nearly twice as large in the nil ( = 0.26, 95% CI [-0.21, 0.74], .178) vs. positive ( = 0.17, 95% CI [0.06, 0.28], = .003), and negative demand conditions ( = 0.19, 95% CI [0.06, 0.33], = .007). In other words, participants’ responses most strongly shift when researchers communicate that they hypothesize *no* change in response

We did not find that the effects of demand characteristics tended to differ depending on whether they were manipulated within- ( = 0.22, 95% CI [0.11, 0.34], < .001) vs. between-subjects ( = 0.14, 95% CI [0.02, 0.25], = .020), (1, 10.48) = 1.55, = .240 (Figure 2). We also did not find that the effects of demand characteristics differed depending on whether participants were unpaid ( = 0.19, 95% CI [0.08, 0.31], .002) vs. paid ( = 0.13, 95% CI [0.00, 0.26], = .047), (1, 20.74) = 0.55, = .465 (Figure 2).

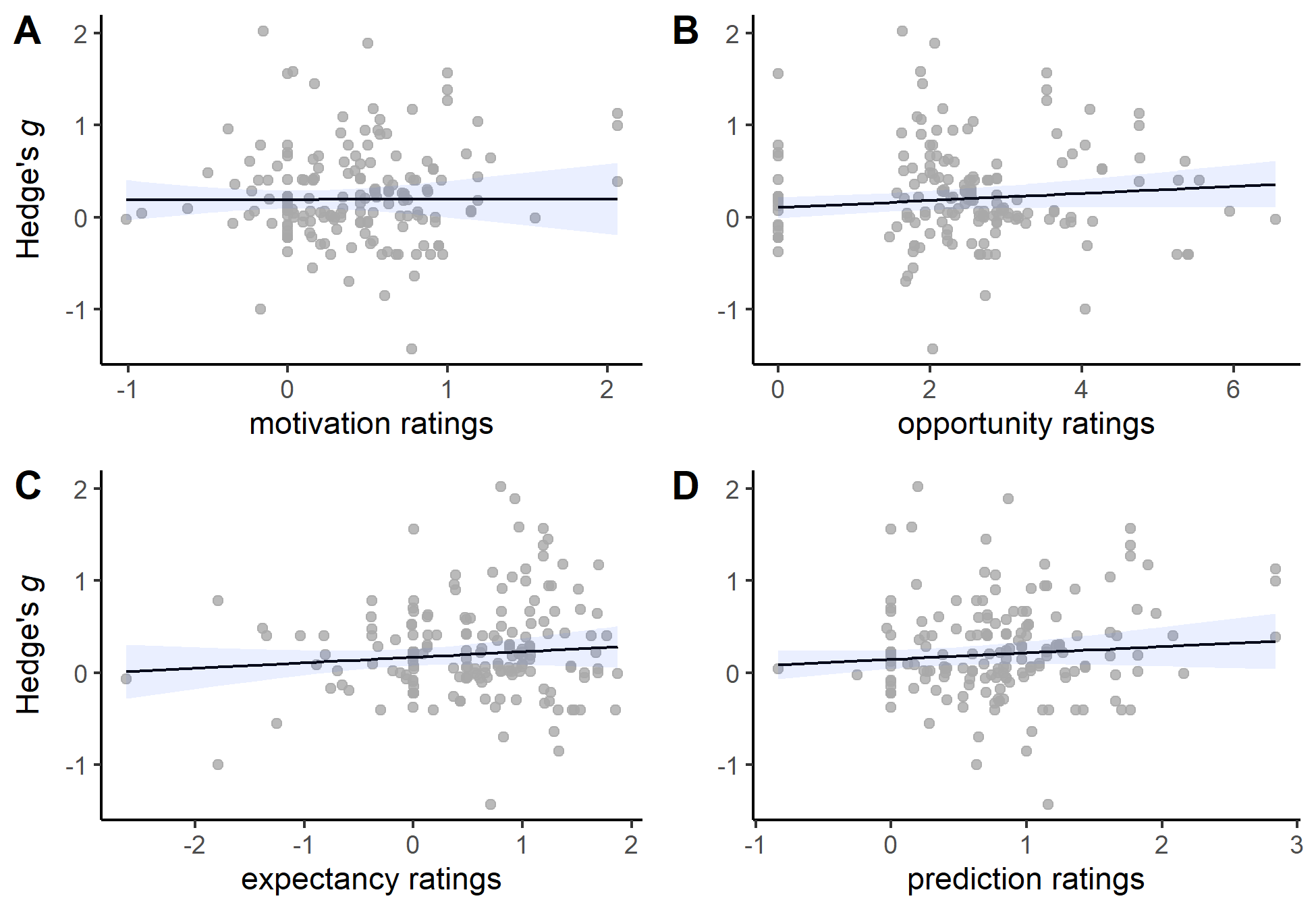


Figure 3: Scatterplots of relationships between the observed effects of demand characteristics (y-axis) and motivation (Panel A), opportunity (Panel B), expectancy (Panel C), and prediction (Panel D) ratings (x-axis). Grey dots represent jittered observations, black lines represent estimated linear relationships, and blue ribbons represent 95% confidence intervals for estimated linear relationships.

#### Theoretical moderators.

Despite the popularity and comprehensiveness of the framework developed by Rosnow and colleagues, we did not uncover significant evidence that demand effects are moderated by ratings of motivation, = 0.01, 95% CI [-0.18, 0.19], (10.87) = 0.06, = .953, or opportunity to adjust responses, = 0.04, 95% CI [-0.01, 0.08], (10.76) = 1.93, = .080. We also did not find that demand effects are moderated by ratings of belief in the hypothesized effect, = NA, 95% CI [-0.05, 0.17], (10.46) = 1.21, = .254. Even more puzzlingly, we did not find that raters were able to accurately predict the emergence of demand effects, = 0.07, 95% CI [-0.04, 0.18], (14.95) = 1.31, = .209. In other words, even equipped with theory-generated ratings scales, we failed to uncover evidence that participatns are able to predict the impact and mechanisms underlying demand characteristics.

#### Residual variability.

To evaluate how much variability in demand effects is currently accounted for by the moderators examined in Studies 1 and 2, we calculated a pseudo- statistic. We did so by comparing the sum of the variance components (between-study + within-study ) in a model containing only an intercept and (1) a model containing Study 1 moderators (type of demand characteristics manipulation, participant pool, setting, study design, and payment), (2) a model containing Study 2 moderators (motivation, opportunity, and expectations), and (3) a model containing both Study 1 and 2 moderators. To ensure comparability, all models included the same set of observations (e.g., any incomplete observations were removed). Results indicated that the Study 1 and 2 moderators respectively accounted for 27.95% and 7.18% of observed variability in demand effects. This might suggest that study features are more powerful predictors of demand effects than the theorized underlying psychological mechanisms; however, we caution against such interpretations given the imperfect ad-hoc measurement of motivations and expectations.

Including all moderators increased the pseudo- to 27.67%, indicating that the majority of observed variability in demand effects is still unaccounted.

### Estimating demand effects in specific study contexts.

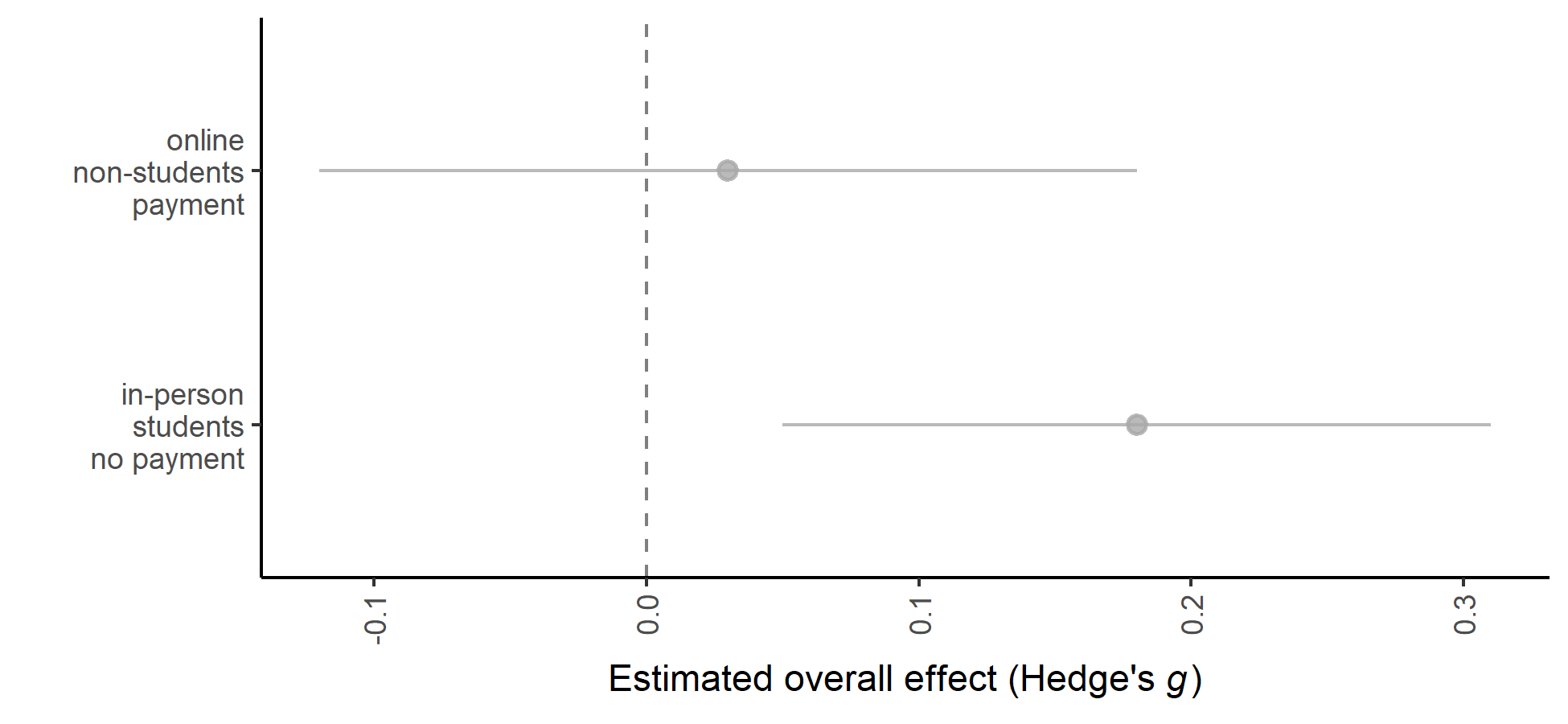


Figure 4: Estimated overall effects (dot) and 95% confidence intervals (error bars) of positive demand characteristics in a classic experimental setting (in-person studies testing positive effects with volunteer student samples) and an online worker setting (online studies testing positive effects with paid non-student samples).

Our openly-available data allow researchers to estimate the effects of demand characteristics in various study contexts. To demonstrate this functionality, we fit a 3LMA with student status, data collection medium, payment status, and type of demand characteristic comparison entered as dummy-coded factors. By changing the reference level of these dummy-coded factors, we were able to derive estimates of demand effects in two common scenarios described below.

First, we estimated the overall impact of demand characteristics in what we call a “classic experimental setting”: studies that (a) are run in-person, (b) sample students, (c) do not offer participant payment, and (d) are testing for a positive effect (i.e., positive demand). In this context, demand characteristics are estimated to produce, on average, a small increase in hypothesis-consistent responding, = 0.18, 95% CI [0.05, 0.31, = .007] (Figure 4). Second, we estimated the overall impact of demand characteristics in an “online worker experimental context”: studies that (a) are run online, (b) sample non-students, (c) offer participant payment, and (d) test for a positive effect. Here, we did not find that demand characteristics, on average, produce changes in participants’ responses, = 0.03, 95% CI [-0.12, 0.18], = 0.69 (Figure 4).

### Publication bias analyses.

Overall, publication bias analyses were inconclusive. A funnel plot containing all effect sizes indicated that publication bias favored instances where participants’ responses shifted in a hypothesis-consistent manner. However, a funnel plot where non-independent effect sizes were aggregated indicated the opposite: that publication bias favored non-significant or hypothesis-inconsistent shifts in participants’ responses.

Precision-effect tests similarly yielded opposite conclusions depending on whether we used (a) 3LMA with non-aggregated effect size estimates, or (b) two-level meta-analysis with aggregated dependent effect size estimates. On one hand, precision-effect tests with 3LMA yielded a non-significant estimate of publication bias that favored hypothesis-consistent shifts in participants’ responses, = 0.65, 95% CI [-0.02, 1.31], = .057. The bias-corrected overall effect size estimate did not significantly differ from zero, = 0.06, 95% CI [-0.12, 0.23], = .535. On the other hand, two-level precision-effect tests with aggregated dependent effect size estimates yielded an opposite pattern: that there was a slight (but not statistically significant) preference for non-significant or hypothesis-inconsistent shifts in participants’ responses, = 0.09, 95% CI [-0.81, 1.00], = .844. The bias-corrected overall effect size estimate was virtually unchanged, = 0.15, 95% CI [-0.03, 0.33], = .107. In other words, depending on how dependencies were handled, precision-effect tests yielded opposite conclusions about the direction of publication bias and the significance of the bias-corrected overall effect of demand characteristics.

Weight-function modeling suggested that better fit was achieved in a model where publication bias favored non-significant or hypothesis-inconsistent shifts in participants’ responses, (1) = 10.80, = .001. The bias-corrected overall effect size was thus upward-adjusted, = 0.41, 95% CI [0.19, 0.62], < .001. A comparison of unpublished ( = 0.17, 95% CI [-0.17, 0.51], = .289) and published ( = 0.21, 95% CI [0.11, 0.31], = < .001) studies yielded a similar pattern, although the difference was not significant, (1, 14.38) = 0.07, = .801.

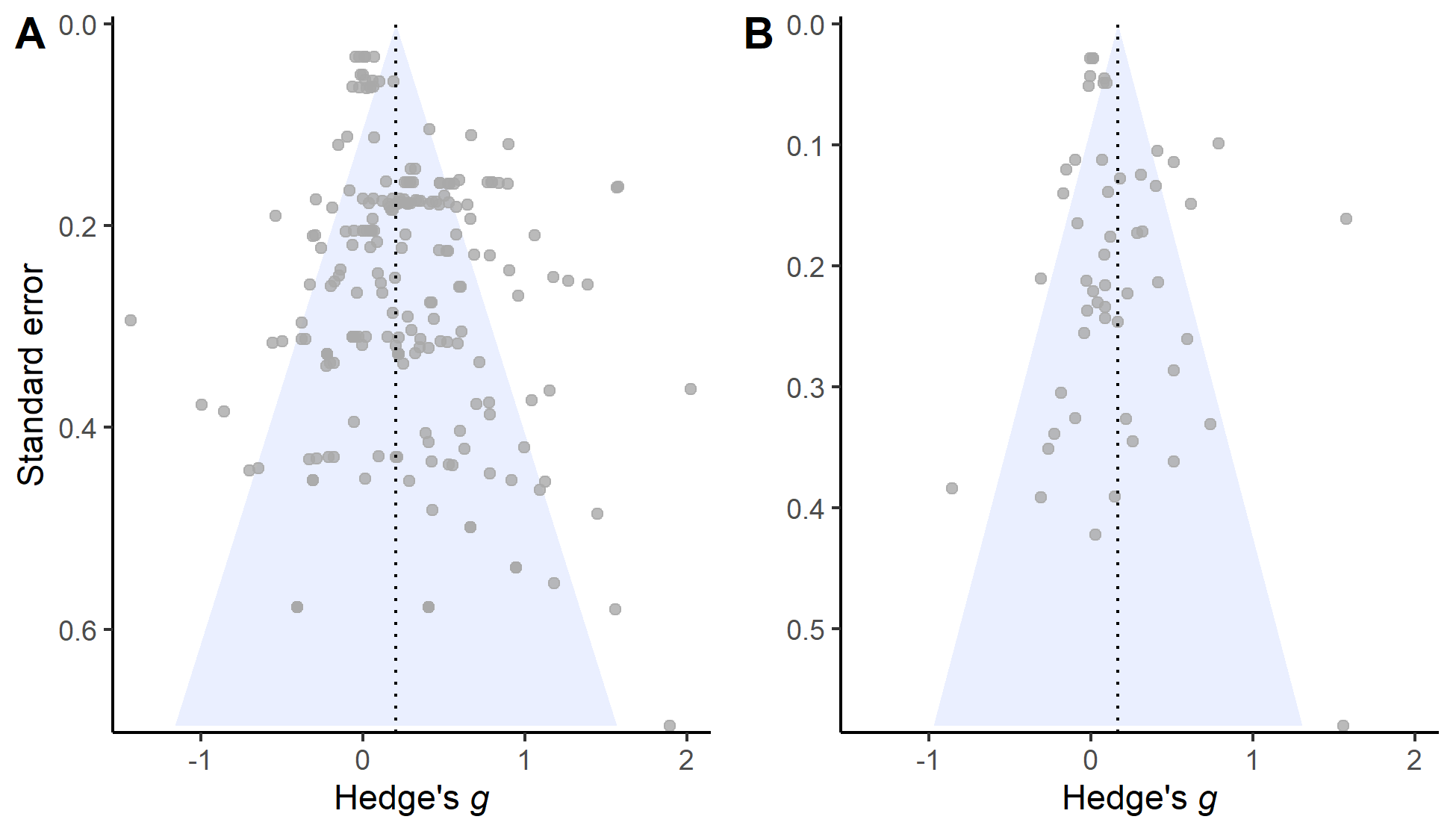


Figure 5: Raw (Panel A) or aggregated (Panel B) effect sizes plotted against their corresponding standard errors.

## Discussion

Overall, explicit manipulations of demand characteristics caused participants’ responses to shift in a manner consistent with the communicated hypothesis. However, high levels of heterogeneity were observed. As an illustration, we estimated the true distribution of demand effects, which suggested that 62% of demand characteristic manipulations produce hypothesis-consistent shifts ( > 0.10); 18% produce hypothesis-*in*consistent shifts ( < -0.10), and 20% produce negligible shifts in either direction (-0.10 < > 0.10). Moderator analyses revealed two study features that were associated with more hypothesis-consistent shifts in responses: (1) sampling student populations, and (2) communicating that the researchers hypothesizes there will be *no* shift in responses (i.e., using nil demand manipulations).

More practically, we estimated that demand characteristics produce small increases in hypothesis-consistent responding in “classic experimental settings”: in-person studies testing positive effects with unpaid student samples. In contrast, when these studies are run online with paid workers – an “online worker experimental setting” – we estimated that demand effects are near zero. However, these results are ultimately preliminary given the high heterogeneity and inconsistent evidence of the direction and impact of publication bias.

Study 1 provides preliminary insights on the magnitude, consistency, and contextual moderators of demand effects. However, it was not designed to examine mechanisms often discussed in motivation and expectancy accounts of demand characteristics. For example, consider our finding that demand characteristics tend to produce more hypothesis-consistent shifts in responses when students (vs. workers) are sampled. If this is true, it may occur because students are more motivated to help the experimenter confirm their hypothesis (motivation account). Alternatively, it may occur because students are more likely to *believe* the communicated hypothesis (expectancy account). In other words, although we have preliminary evidence of contextual modifiers of demand effects, we still lack an explanation of why these contexts matter and how demand effects work more broadly. In Study 2, we begin investigating this outstanding issue through an extension of the meta-analysis.

# General Discussion

Demand characteristics are a puzzling methodological artifact in experimental psychology. Using meta-analysis and a replication of a recent study, we sought to consolidate knowledge about the magnitude, reliability, and mechanisms underlying these effects. The meta-analysis revealed that demand characteristics typically led participants to slightly shift their responses in the direction of the communicated hypothesis. However, publication bias analyses were inconclusive, and the estimated effects were heterogeneous. This heterogeneity led us to estimate the true *distribution* of demand effects. This distribution suggests that approximately 62% of demand characteristic manipulations produce hypothesis-consistent shifts in participants’ responses ( > 0.10), 18% produce hypothesis-*in*consistent shifts in participants’ responses ( < -0.10), and 20% produce negligible shifts in either direction (-0.10 < > 0.10). Most worrisome, the estimated distribution of demand effects ranges from approximately = -1.33 to = 1.82 – remarkable similar to the distribution of theory-relevant effects in experimental psychology (Lovakov & Agadullina, 2021). Thus, to distinguish theory-relevant effects from artifactual demand effects, it is essential that experimental psychologists better understand the latter (Boot, Simons, Stothart, & Stutts, 2013; Corneille & Lush, 2022; Sharpe & Whelton, 2016).

Participants themselves appeared to have little ability to predict or explain the impact of demand characteristics in the studies they reviewed, although it is possible that their performance would improve if they were provided with more information, given better measures, and/or better incentivized to provide accurate predictions. Fortunately, our meta-analysis allows us to make some predictions. Moderator analyses provided preliminary evidence that some methodological factors – such as student samples – are associated with increases in hypothesis-consistent responding. We also found that demand characteristics tended to be more impactful when a nil (as opposed to negative or positive) hypothesis was communicated. Nonetheless, most of the variability we observed in the meta-analysis was not explained.

Although motivation accounts are popular in in the demand characteristics literature (Cook et al., 1970; Orne, 1962; Riecken, 1962; Rosenberg, 1969; Rosnow & Rosenthal, 1997; Sigall et al., 1970), we found little support for this view. In the Study 2 meta-analysis, we did not find that external ratings of two factors oft-discussed in motivation accounts – motivation and opportunity to adjust responses – moderated demand effects. We found some evidence in Study 3 that motivation (but not opportunity) ratings moderated demand effects, but the evidence was weak.

Consistent with expectancy accounts, our results consistently indicated that participants’ expectations about hypothesized effects partially drive demand effects. This may occur because demand characteristics activate pre-existing beliefs about a phenomenon being investigated. It is also possible that demand characteristics cause participants to update or form new expectations. If true, research on how beliefs are formed, updated, and impact participant responses may help explain the heterogeneous effects of demand characteristic manipulations. For example, if expectations are governed by Bayesian principles (for a review, see Kube & Rozenkrantz, 2021), demand characteristics should exert larger effects in contexts where participants have relatively uncertain pre-existing expectations.

Expectancy effects can certainly be reduced – but it is not clear if they can be fully avoided (Boot et al., 2013). Indeed, participants possess a rich array of pre-existing expectations *before* they enter our studies (Dweck, 2012). For example, Coles et al. (2022) estimated that 44% of sampled undergraduates and 34% of sampled online workers believed – before entering the study – that facial poses impact emotion. Even with deception about the purpose of the study, these pre-existing beliefs appear to shape the extent to which participants exhibit facial feedback effects. Similar observations have been made elsewhere, for example in literature examining the effects of videogame training on cognitive tasks (Boot et al., 2013). In other words, deception about the purpose of the study does not guarantee an unbiased estimate of a mechanism-of-interest. In the real world, the mechanisms that psychologists theorize about may be naturalistically confounded with participants’ beliefs.

We end on a note of concern. We estimated that experimentally manipulated demand characteristics have a similar distribution of effects as the theory-relevant phenomena that many psychologists study (Lovakov & Agadullina, 2021). These demand effects appear to be most reliably predicted by participant expectations about the hypothesized effects (i.e., expectancy effects). Even when specific demand characteristics are eliminated, participants have expectations about the phenomena we study – and these expectations may be naturalistically confounded with the theory-relevant mechanisms we wish to study. Thus, if (a) demand characteristics are present or (b) participants are likely to have pre-existing expectations about the phenomenon being studied, researchers should be wary of concluding that an observed effect is not compromised by methodological artifacts (Boot et al., 2013; Corneille & Lush, 2022).

# References

References marked with an asterisk indicate studies included in the meta-analysis.

\* Allen, A. P., & Smith, A. P. (2012). Demand characteristics, pre-test attitudes and time-on-task trends in the effects of chewing gum on attention and reported mood in healthy volunteers. *Appetite*, *59*(2), 349–356.

Aust, F., & Barth, M. (2022). *papaja: Prepare reproducible APA journal articles with R Markdown*. Retrieved from <https://github.com/crsh/papaja>

\* Barabasz, M., Barabasz, A., & O’Neill, M. (1991). Effects of experimental context, demand characteristics, and situational cues: New data. *Perceptual and Motor Skills*, *73*(1), 83–92.

Boot, W. R., Simons, D. J., Stothart, C., & Stutts, C. (2013). The pervasive problem with placebos in psychology: Why active control groups are not sufficient to rule out placebo effects. *Perspectives on Psychological Science*, *8*(4), 445–454.

Borenstein, M. (2009). Effect sizes for continuous data. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of synthesis and meta-analysis* (pp. 221–235). New York, NY: Russell Sage Foundation.

Borenstein, M., Hedges, L. V., Higgins, J. P., & Rothstein, H. R. (2011). *Introduction to meta-analysis*. John Wiley & Sons.

\* Busch, A. M., Kanter, J. W., Sedivy, S. K., & Leonard, J. L. (2007). A follow-up analogue study on the effectiveness of the cognitive rationale. *Cognitive Therapy and Research*, *31*, 805–815.

Coburn, K. M., & Vevea, J. L. (2019). *Weightr: Estimating weight-function models for publication bias*. Retrieved from <https://CRAN.R-project.org/package=weightr>

Cohen, J. (2013). *Statistical power analysis for the behavioral sciences* (Vol. 2). New York, NY: Lawrence Erlbaum Associates.

\* Coles, N. A., Gaertner, L., Frohlich, B., Larsen, J. T., & Basnight-Brown, D. M. (2022). Fact or artifact? Demand characteristics and participants’ beliefs can moderate, but do not fully account for, the effects of facial feedback on emotional experience. *Journal of Personality and Social Psychology*.

Cook, T. D., Bean, J. R., Calder, B. J., Frey, R., Krovetz, M. L., & Reisman, S. R. (1970). Demand characteristics and three conceptions of the frequently deceived subject. *Journal of Personality and Social Psychology*, *14*(3), 185–194.

Corneille, O., & Lush, P. (2022). Sixty years after orne’s american psychologist article: A conceptual framework for subjective experiences elicited by demand characteristics. *Personality and Social Psychology Review*, 81–101.

\* Cramer, D. (2004). Effect of the destructive disagreement belief on relationship satisfaction with a romantic partner or closest friend. *Psychology and Psychotherapy: Theory, Research and Practice*, *77*(1), 121–133.

\* Cramer, D. (2005). Effect of the destructive disagreement belief on satisfaction with one’s closest friend. *The Journal of Psychology*, *139*(1), 57–66.

\* Cramer, D., & Buckland, N. (1995). Effect of rational and irrational statements and demand characteristics on task anxiety. *The Journal of Psychology*, *129*(3), 269–275.

Drevon, D., Fursa, S. R., & Malcolm, A. L. (2017). Intercoder reliability and validity of WebPlotDigitizer in extracting graphed data. *Behavior Modification*, *41*(2), 323–339.

\* Durgin, F. H., Klein, B., Spiegel, A., Strawser, C. J., & Williams, M. (2012). The social psychology of perception experiments: Hills, backpacks, glucose, and the problem of generalizability. *Journal of Experimental Psychology: Human Perception and Performance*, *38*(6), 1582.

Dweck, C. S. (2012). Implicit theories. In P. A. M. V. Lange, A. W. Kruglanski, & T. Higgins (Eds.), *Handbook of theories of social psychology* (Vol. 2, pp. 43–61). London: SAGE Publications Ltd.

\* Earn, B. M. (1979). *Experimental compensation, task interest and the cooperation with demand characteristics of volunteer and sign-up subjects.* (PhD thesis).

Fillenbaun, S., & Frey, R. (1970). More on the" faithful" behavior of suspicious subjects. *Journal of Personality*, *38*(1), 43–51.

Franco, A., Malhotra, N., & Simonovits, G. (2014). Publication bias in the social sciences: Unlocking the file drawer. *Science*, *345*(6203), 1502–1505.

Hayes, C., & King, W. (1967). Two types of phenomenal instructions for size and distance judgments of objects presented on a two-dimensional plane. *Perception & Psychophysics*, *2*(11), 556–558.

\* Hoogeveen, S., Schjoedt, U., & Elk, M. van. (2018). Did i do that? Expectancy effects of brain stimulation on error-related negativity and sense of agency. *Journal of Cognitive Neuroscience*, *30*(11), 1720–1733.

\* Isager, P. (2022). *Student replication of coles et al. 2022*.

\* Kanter, J. W., Kohlenberg, R. J., & Loftus, E. F. (2004). Experimental and psychotherapeutic demand characteristics and the cognitive therapy rationale: An analogue study. *Cognitive Therapy and Research*, *28*, 229–239.

\* Kenealy, P. (1988). Validation of a music mood induction procedure: Some preliminary findings. *Cognition & Emotion*, *2*(1), 41–48.

\* Kersbergen, I., Whitelock, V., Haynes, A., Schroor, M., & Robinson, E. (2019). Hypothesis awareness as a demand characteristic in laboratory-based eating behaviour research: An experimental study. *Appetite*, *141*, 104318.

Kube, T., & Rozenkrantz, L. (2021). When beliefs face reality: An integrative review of belief updating in mental health and illness. *Perspectives on Psychological Science*, *16*(2), 247–274.

\* Lamberth, J., & Byrne, D. (1971). Similarity-attraction or demand characteristics. *Personality*, *2*(2), 77–91.

\* Lana, R. E., & Menapace, R. H. (1971). Subject commitment and demand characteristics in attitude change. *Journal of Personality and Social Psychology*, *20*(2), 136.

Lovakov, A., & Agadullina, E. R. (2021). Empirically derived guidelines for effect size interpretation in social psychology. *European Journal of Social Psychology*, *51*(3), 485–504.

\* Martin, J.-R., Sackur, J., & Dienes, Z. (2018). Attention or instruction: Do sustained attentional abilities really differ between high and low hypnotisable persons? *Psychological Research*, *82*(4), 700–707.

Masling, J. (1966). Role-related behavior of the subject and psychologist and its effects upon psychological data. *Nebraska Symposium on Motivation*, *14*, 67–103.

\* McGinley, H., Kaplan, M., & Kinsey, T. (1975). Subject effects and demand characteristics. *Psychological Reports*, *36*(1), 267–278.

\* McGlynn, F. D., Gaynor, R., & Puhr, J. (1972). Experimental desensitization of snake-avoidance after an instructional manipulation. *Journal of Clinical Psychology*.

\* Mummolo, J., & Peterson, E. (2019). Demand effects in survey experiments: An empirical assessment. *American Political Science Review*, *113*(2), 517–529.

Orne, M. T. (1959). The nature of hypnosis: Artifact and essence. *The Journal of Abnormal and Social Psychology*, *58*(3), 277–299.

Orne, M. T. (1962). On the social psychology of the psychological experiment: With particular reference to demand characteristics and their implications. *American Psychologist*, *17*(11), 776–783.

Orne, M. T. (1969). Demand characteristics and the concept of quasi-controls. In R. Rosenthal & R. L. Rosnow (Eds.), *Artifacts in behavioral research* (pp. 143–179). New York, NY: Academic Press.

\* Orne, M. T., & Scheibe, K. E. (1964). The contribution of nondeprivation factors in the production of sensory deprivation effects: The psychology of the" panic button.". *The Journal of Abnormal and Social Psychology*, *68*(1), 3.

\* Perry, D. G., Roots, R. D., & Perry, L. C. (1978). Demand awareness and participant willingness as determinants of aggressive response to film violence. *The Journal of Social Psychology*, *105*(2), 265–275.

\* Polivy, J., & Doyle, C. (1980). Laboratory induction of mood states through the reading of self-referent mood statements: Affective changes or demand characteristics? *Journal of Abnormal Psychology*, *89*(2), 286.

Pustejovsky, J. E., & Tipton, E. (2018). Small-sample methods for cluster-robust variance estimation and hypothesis testing in fixed effects models. *Journal of Business & Economic Statistics*, *36*(4), 672–683.

R Core Team. (2021). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from <https://www.R-project.org/>

Riecken, H. W. (1962). A program for research on experiments in social psychology. In N. W. Washburne (Ed.), *Decisions, values and groups* (Vol. 2, pp. 25–41). New York, NY: Pergamon Press.

Rodgers, M. A., & Pustejovsky, J. E. (2021). Evaluating meta-analytic methods to detect selective reporting in the presence of dependent effect sizes. *Psychological Methods*, *26*(2), 141.

\* Rose, J. P., Geers, A. L., Fowler, S. L., & Rasinski, H. M. (2014). Choice-making, expectations, and treatment positivity: How and when choosing shapes aversive experiences. *Journal of Behavioral Decision Making*, *27*(1), 1–10.

Rosenberg, M. J. (1969). The conditions and consequences of evaluation apprehension. In R. Rosenthal & R. L. Rosnow (Eds.), *Artifacts in behavioral research* (pp. 280–350). New York, NY: Academic Press.

Rosnow, R. L., & Aiken, L. S. (1973). Mediation of artifacts in behavioral research. *Journal of Experimental Social Psychology*, *9*(3), 181–201.

Rosnow, R. L., & Rosenthal, R. (1997). *People studying people: Artifacts and ethics in behavioral research*. New York, NY: Freeman.

Schardt, C., Adams, M. B., Owens, T., Keitz, S., & Fontelo, P. (2007). Utilization of the PICO framework to improve searching PubMed for clinical questions. *BMC Medical Informatics and Decision Making*, *7*(1), 1–6.

\* Schauer, E. (1969). *Demand characteristics in a quasi-psychophysical experiment.* (PhD thesis).

Sharpe, D., & Whelton, W. J. (2016). Frightened by an old scarecrow: The remarkable resilience of demand characteristics. *Review of General Psychology*, *20*(4), 349–368.

\* Siegel, W. E., & Hagen, R. L. (1982). The influence of demand characteristics and expectancies in the measurement of salivary response. *Journal of Behavioral Assessment*, *4*, 179–185.

Sigall, H., Aronson, E., & Van Hoose, T. (1970). The cooperative subject: Myth or reality? *Journal of Experimental Social Psychology*, *6*(1), 1–10.

\* Smith, J. M., Bell, P. A., & Fusco, M. E. (1986). The influence of color and demand characteristics on muscle strength and affective ratings of the environment. *The Journal of General Psychology*, *113*(3), 289–297.

\* Standing, L. G., Verpaelst, C. C., & Ulmer, B. K. (2008). A demonstration of nonlinear demand characteristics in the’mozart effect’experimental paradigm. *North American Journal of Psychology*, *10*(3), 553–566.

Stanley, T. D., & Doucouliagos, H. (2014). Meta-regression approximations to reduce publication selection bias. *Research Synthesis Methods*, *5*(1), 60–78.

Strohmetz, D. B. (2008). Research artifacts and the social psychology of psychological experiments. *Social and Personality Psychology Compass*, *2*(2), 861–877.

\* Terhune, D. B., & Smith, M. D. (2006). The induction of anomalous experiences in a mirror-gazing facility: Suggestion, cognitive perceptual personality traits and phenomenological state effects. *The Journal of Nervous and Mental Disease*, *194*(6), 415–421.

\* Veitch, J. A., Gifford, R., & Hine, D. W. (1991). Demand characteristics and full spectrum lighting effects on performance and mood. *Journal of Environmental Psychology*, *11*(1), 87–95.

\* Verpaelst, C. C., & Standing, L. G. (2007). Demand characteristics of music affect performance on the wonderlic personnel test of intelligence. *Perceptual and Motor Skills*, *104*(1), 153–154.

Vevea, J. L., & Hedges, L. V. (1995). A general linear model for estimating effect size in the presence of publication bias. *Psychometrika*, *60*(3), 419–435.

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, *36*(3), 1–48.

1. We excluded conditions where the researcher communicated a *non-directional* effect. We did so because participants in these scenarios could not unambiguously infer how their responses were expected to change. For example, if participants were told that an independent variable would “impact mood”, it is not clear if participants should infer that the mood will be boosted or dampened. [↑](#footnote-ref-24)
2. After completing the meta-analysis, we learned that one included record by Lana and Menapace (1971) did include a clinical population: patients who suffered pulmonary or cerebrovascular accidents. Excluding their results does not change the general pattern of results. [↑](#footnote-ref-25)
3. As a reminder, we performed two literature searches. The second literature search was inspired by reviewer feedback, which we received after we started collecting data using the vignette methodology. [↑](#footnote-ref-29)
4. For effect size aggregation, we assumed a default dependent effect size correlation of = .50 but performed sensitivity analysis with = .10, .30, .50, .70, and .90. These sensitivity analyses did not change our overall conclusion about publication bias, so we do not discuss them. [↑](#footnote-ref-34)
5. When assessing publication bias using 3LMA, we also fit an exploratory model that included cluster-robust estimates of the variance covariance matrix. Cluster-robust estimation procedures did not change our inferences. [↑](#footnote-ref-35)