A meta-analysis on demand characteristics

Nicholas A. Coles1 & Michael C. Frank1

1 Center for the Study of Language and Information, Stanford University

Author note

All materials, data, and code are available at .

Correspondence concerning this article should be addressed to Nicholas A. Coles, Cordura Hall, 210 Panama St, Stanford, CA 94305. E-mail: [ncoles@stanford.edu](mailto:ncoles@stanford.edu)

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. In the first quantitative synthesis on the topic, we conducted a three-level meta-analysis of 195 effect sizes from 40 studies that provided strict experimental tests of demand effects by manipulating the hypothesis communicated to participants. Results indicated that these demand characteristics tend to produce small increases in hypothesis-consistent responding. However, these effects were extremely heterogeneous. The estimated distribution of effects ranged from d = 1.91 (a massive increase in hypothesis-consistent responding) to d = -1.67 (a massive increases in hypothesis-*in*consistent responding). This range covers the span of almost every conceivable effect in experimental psychology. Contrary to conventional views, we found little evidence that demand effects were driven by participants’ motivation and opportunity to adjust their responses—i.e., a response bias. Instead, such effects seemed to be more so driven by participants’ beliefs—i.e., placebo effects. Similar findings emerged in a direct replication of a study included in the meta-analysis. Taken together, results challenge conventional distinctions between demand characteristics and placebo effects. More importantly, they highlight a pressing need to understand the mysterious but potentially massive impact of demand characteristics.

*Keywords:* demand characteristics, hypothesis awareness, placebo effect, research methods, meta-analysis

*Word count:* TBD

A meta-analysis on demand characteristics

Imagine that one day a mysterious person approaches you and begins telling you about a new method for understanding humans: Crankology[[1]](#footnote-1). The person explains that Crankology is useful for estimating causal relationships—but adds that it can sometimes be thrown off by a *methodological artifact.* When you ask the Crankologist about this artifact, they explain that it sometimes causes researchers to detect an effect that’s not real, and other times causes them to miss an effect that is real. They add that it sometimes causes relationships to be biased upward and other times causes them to be biased downward. And then they offer a confession: they don’t understand how the artifact works. Sometimes the artifact seems to matter, other times it doesn’t—and its underlying mechanism is poorly understood.

If this scenario were real, you would reasonably question whether Crankology is a valid method of scientific inquiry. However, perhaps we should not be so quick to judge. Because, like Crankology, experimental psychologists deal with a difficult-to-understand methodological artifact: *demand characteristics*.

## Demand characteristics as a methodological artifact

In 1962, Martin Orne published a seminal paper highlighting a view that challenged deeply-ingrained beliefs about the role of human subjects in experiments. Contrary to popular views at the time, Orne argued that research participants are not passive responders to the experimental context. Instead, he suggested that participants are perceptive to demand characteristics—“cues which convey an experimental hypothesis to the subject”—and are motivated to use these cues to help the experimenter confirm their hypothesis (Orne, 1962, p. 779). This idea was controversial at first, with some researchers suggesting that the concern was vague and/or overblown (e.g., Berkowitz, 1971; Kruglanski, 1975; Milgram, 1972). Nonetheless, over the next 60 years, demand characteristics would become recognized as a literal textbook methodological concern in experimental psychology (Sharpe & Whelton, 2016).

Orne initially focused on evidence that demand characteristics can lead to false positives—such as patients exhibiting sham symptoms of hypnosis (Orne, 1959). Follow-up research, though, indicated that demand characteristics can also lead to false negatives. For example, Hayes and King (1967) demonstrated that participants will ignore visual cues of depth when they believe that doing so is the purpose of the experiment. Of course, in addition to creating inferential errors, demand characteristics can bias estimates of causal relationships. For example, Coles, Gaertner, Frohlich, Larsen, and Basnight-Brown (2022) found that the estimated effect of facial poses on self-reported emotion could be amplified *or* attenuated depending on whether the experimenter communicates expectations of positive or nil effects. However, not all researchers have found that demand characteristics matter. For example, in large replications of classic studies in behavioral economics, Mummolo and Peterson (2019) consistently failed to find that manipulations of the communicated hypothesis impacted participants’ responses.

After over 60 years, experimental psychologists are left with an uncomfortable state of affairs. Demand characteristics are a literal textbook methodological concern. However, like Crankology, the magnitude, direction, consistency, and mechanisms underlying this artifact remain mysterious.

## How do demand characteristics bias participant responses?

Historically, theorists have conceptualized the effects of demand characteristics as *response biases* mediated by relatively deliberate changes that participants make to their responses (Orne, 1962; Rosnow & Aiken, 1973; Strohmetz, 2008). In doing so, these theorists distinguished their ideas from conceptually similar work on *placebo effects*: changes in participants’ responses that are mediated by the relatively automatic activation of beliefs and/or conditioned responses (Zion & Crum, 2018). As an example of this distinction, imagine that a participant knows that a researcher expects an intervention to boost mood. Response bias—the historical focus of the demand characteristics literature—would involve a change in participants’ self-reported mood without a concomitant change in actual mood. Placebo effects, on the other hand, would entail an actual change in mood.

As we review below, the most comprehensive demand characteristics framework follows this historic trend, conceptualizing the artifact as a response bias (Rosnow & Rosenthal, 1997). However, this conceptualization was recently challenged by Coles, Gaertner, et al. (2022) and Corneille and Lush (2022), who argued that demand characteristics can lead to both response biases *and* placebo effects (Figure 1). For example, after inferring that a researcher expects an intervention to boost mood, a participant may both (a) deliberately adjust their mood ratings (a response bias), and (b) unintentionally experience a placebo-induced change in mood. We discuss these two mechanisms in greater detail below.

### Response bias.

To date, the most influential framework for conceptualizing the effects of demand characteristics has been developed by Rosnow and Rosenthal (1997). Like most researchers, Rosnow and Rosenthal (1997) suggested that demand characteristics produce response biases. As such, they proposed three key moderators: (1) receptivity to cues, (2) motivation to provide hypothesis-consistent responses, and (3) opportunity to alter responses.



*Figure* *1.*  Rosnow and Rosenthal’s (1997) and Coles, Gaertner, et al.’s (2022) frameworks for conceptualizing how demand characteristics can lead to increases (green), decreases (red), or no shift (light grey) in hypothesis-consistent responding. Rosnow and Rosenthal conceptualized demand effects as response biases moderated by receptivity to cues (not pictured), motivation, and opportunity to adjust responses. Coles, Gaertner, et al. proposed that demand characteristics can also produce placebo biases (dotted boxes) that are driven by the activation of or change in participants’ beliefs.

#### Receptivity to cues.

To start, Rosnow and Rosenthal (1997) reasoned that participants must be perceptive to demand characteristics in order for there to be a response bias (see also Rosnow & Aiken, 1973; Strohmetz, 2008). As an extreme example, imagine that a researcher hands an infant participant 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 (i.e., cannot read). In the present work, we will focus on scenarios where participants are likely to be highly receptive to cues. Thus, we will pay less attention to receptivity as a moderator (but see the General Discussion for more information).

#### Motivation to provide hypothesis-consistent responses.

Early in the history of research on demand characteristics, researchers debated which motivational forces typically underlie its subsequent response bias (for a review, see Weber & Cook, 1972). Orne (1962) originally characterized participants as “good subjects” who change their responses because they are motivated to help the researcher confirm their hypothesis. Others characterized participants as “apprehensive subjects” who are motivated to respond in a manner that will lead them to be evaluated positively (Riecken, 1962; Rosenberg, 1969; Sigall, Aronson, & Van Hoose, 1970). Masling (1966) argued that participants sometimes interfere with the purpose of the study (“negativistic subjects,” see also Cook et al., 1970), whereas Fillenbaun and Frey (1970) argued that participants attempt to follow directions as closely as possible (“faithful subjects”). Although seemingly divided, these early theorists agreed on one overarching principle: response bias is driven by participants’ motivation (or lack thereof) to provide hypothesis-consistent responses.

In the most prolific era of demand characteristics research, investigators sought to understand which subject goal *predominately* mediated response bias. For example, Sigall et al. (1970) found that participants increased performance on a simple task when the experimenter indicated that this was their expectation. However, participants did *not* do so when told that increased performance would be indicative of an obsessive-compulsive personality. Based on these results, Sigall et al. (1970) concluded that participants were predominately motivated to secure a positive evaluation—not help the experimenter confirm their hypothesis.

By focusing on testing hypotheses about a single predominate participant goal, less attention was initially paid to the notion that participants might have multiple, sometimes competing motivations (Barbuto Jr & Scholl, 1998; Boudreaux & Ozer, 2013). Later, though, Rosnow and Rosenthal (1997) demonstrated that people have *multiple* goals in mind when they conceptualize their role as research participants. Participants describe their role as being similar to situations where one is being altruistic (e.g., giving to charity), being evaluated (e.g., being interviewed for a job), *and* obeying authority (e.g., obeying a no-smoking sign). All these goals may impact the extent to which participants are overall motivated to provide hypothesis-consistent responses. Furthermore, these goals can sometimes conflict. For example, in the Sigall et al. (1970) experiment, participants may have been motivated to both (a) secure a positive evaluation, and (b) [presumably to a smaller degree] help the experimenter confirm their hypothesis. The brilliance of Rosnow and Rosenthal’s proposal is that it acknowledged that all previous researchers were [at least somewhat] correct; participants are altruistic, apprehensive, negativistic, *and* faithful. This effectively ended debates about which motivational force predominately drove response biases.

Synthesizing the above observations and reasoning, Rosnow and Rosenthal (1997) suggested that participants 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). Of course, as we later discuss, motivation can also be conceptualized on a continuum ranging from highly motivated to counter-acquiesce to highly motivated to acquiesce.

#### Opportunity to alter responses.

No matter how motivated they are to confirm the hypothesis, Rosnow and Rosenthal (1997) reasoned that there is variability in the extent to which participants have the opportunity/ability to alter the outcome-of-interest. Taking this third moderator into account, Rosnow and Rosenthal concluded that demand characteristics only produce response biases 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).

### Response bias and placebo effects.

Over the past half century, demand characteristics have generally been conceptually divorced from placebo effects (e.g., Orne, 1969). Indeed, in the classic book describing artifacts and behavioral research (Rosnow & Rosenthal, 1997), placebo effects are acknowledged as a historical precursor to research on methodological artifacts but not discussed in the context of demand characteristics. This conceptual separation, however, has recently been challenged by Coles, Gaertner, et al. (2022) and Corneille and Lush (2022), who argued that demand characteristics not only have the potential to lead to response biases, but also placebo effects (Figure 1). Consistent with this reasoning, Coles, Gaertner, et al. (2022) found that participants’ beliefs did not always match the hypothesis communicated to participants; furthermore, both the communicated hypothesis and measures of participants’ beliefs moderated the effects of posed expressions on emotion. Contrary to Rosnow and Rosenthal (1997), this work provides preliminary evidence that demand characteristics can produce both response biases and placebo effects. This means that demand characteristics can still bias responses when participants have neither the motivation nor the opportunity to adjust their responses—challenging the conventional playbook for avoiding the impact of this methodological artifact.

## Goals

The goal of the current paper is to take stock of what we know—and what we don’t know—about demand characteristics as a methodological artifact. In Study 1a, we report a meta-analysis of strict experimental tests of the effects of demand characteristics, with a focus on the the direction, magnitude, and consistency of the effects. We then examine several study features (e.g., whether participants are paid) that researchers have specified as potential moderators.

In Study 1b, we review an extension of the meta-analysis that examines whether observed effect size variability can be explained by factors theorized to underlie response biases (i.e., motivation and opportunity to adjust responses) and placebo effects (i.e., belief in the experimenter’s hypothesis). To do so, we derived estimates of these factors from a new set of participants. These participants read descriptions of each study in the meta-analysis and then reported the extent to which they hypothetically would have (a) been motivated to confirm the experimenter’s hypothesis, (b) had the opportunity to adjust their responses, and (c) believed the experimenter’s hypothesis. We also examined how well this new set of participants could predict the effects of the studies’ demand characteristic manipulations.

In Study 2, we review a small replication study that re-examines the extent to which demand effects are driven by response biases and placebo effects. In this replication study, we manipulated demand characteristics in an experimental investigation of the proposed effects of facial poses on emotional experience (Coles, Larsen, & Lench, 2019; Coles, March, et al., 2022). We then examined the extent to which the effect of facial poses was moderated by factors believed to underlie response biases (i.e., self-reported motivation and opportunity to adjust responses) and placebo effects (i.e., self-reported belief in facial feedback effects).

# Study 1a

Study 1a was designed to provide the first quantitative synthesis of strict experimental tests of demand effects, with a focus on their direction, magnitude, and consistency.

## Methodology

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 research studies. We excluded clinical research studies so that we could focus on research that better isolated the discipline (experimental psychology) and mechanism (response bias) most often discussed in the demand characteristics literature. Given that there is a sizable literature on placebo effects, excluding clinical research studies also helped us improve the feasibility of the meta-analysis.

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. Orne (1962) more broadly defined demand characteristics as *any* cue that may impact participants’ beliefs about the purpose of the study, including instructions, rumors, and experimenter behavior. However, such a definition creates a blurry and potentially boundless conceptual space where any systematic change in a research design might be considered a test of demand characteristics. Thus, 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” or “mood-dampening”, the outcome-of-interest would be any measure of mood.

### Literature search.

Our literature search strategy was developed in consultation with a librarian at Stanford University. Given the broad nature of the demand characteristics construct, we determined that a truly comprehensive strategy was not feasible. Thus, we sought to design a strategy that best balanced comprehensiveness and feasibility.

We searched APA PsycInfo using broad search terms: “demand characteristics” OR “hypothesis awareness”. This yielded 850 records. 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. This yielded 3 additional records. In total, 97 of the records were unpublished.

### 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 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, this was compared to a *control* condition, where participants were not told about an effect of an independent variable on a dependent variable.
* We excluded conditions where the researcher communicated a *non-directional* effect. We did so because participants in these scenarios could not unambiguoursly 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.
* The demand characteristics manipulation was not strongly confounded. 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).
* Information necessary for computing at least one effect size was included.

N. C. and a research assistant screened records independently, reviewed potentially relevant records together, and worked together to code the information for moderator analyses and effect size computations. Disagreements and discrepancies were resolved through discussion. It total, 42 studies from 31 records were eligible for inclusion. However, one record (Allen & Smith, 2012) was removed because the information reported led to implausibly large effect size estimates (e.g., = -212.57).

### Effect size index.

We used standardized mean difference scores (Cohen’s and ) as our effect size index (Borenstein, 2009; Cohen, 2013).

In most scenarios, we estimated the main effect of demand characteristics. For example, Coles, Gaertner, et al. (2022) manipulated whether participants were told that posing smiles would increase happiness. Here, the main effect of demand characteristics can be computed by comparing happiness ratings from smiling participants who were either informed or not informed of its mood-boosting effect.

In some scenarios, we estimated the *interactive* effect of demand characteristics. For example, in the same Coles, Gaertner, 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 demand characteristics. In this scenario, the interactive effect of demand characteristics was computed by calculating a standardized difference-in-differences score. These scores were computed similar to Cohen’s and , but with mean-difference scores (as opposed to means).

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 participants were told that the intervention should be mood dampening, an increase in mood would be coded as a negative effect.

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.

85% of studies 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 in our models (described later).

### Potential moderators.

We coded several moderators that may help explain variability in demand effects. The first of these moderators allowed us to assess whether demand effects are additive. As a reminder, Cohen’s represents a standardized difference between two groups. Often, this involved a single demand characteristic condition (positive, negative, or nil demand) compared to a control group. Sometimes, however, this comparison involved *two* demand characteristic conditions (e.g., positive demand vs. negative demand). If demand characteristics can be additive, their effects should be larger when two demand characteristic conditions are compared (as opposed to one condition being compared to a control group).

Instances where a demand characteristic condition was compared to a control group allowed us to additional test whether participants respond more strongly to positive, nil, or negative demand characteristics. We thus coded whether the comparison was positive demand vs. control, nil demand vs. control, or negative demand vs. control.

We also coded several study feature moderators that researchers have speculated may moderate demand effects. This included: (1) whether the sample was student, non-student (e.g., MTurk), or mixed, (2) whether the study was conducted online or in-person, (3) whether demand characteristics were manipulated within- vs. between-subjects, and (4) whether participants were paid or unpaid.

### Meta-analytic approach.

85% of studies in our meta-analysis contained multiple effect sizes of interest. To model this nested structure, we used three-level meta-analysis (3LMA; also referred to as “multilevel” meta-analysis). 3LMA accommodates nested effect sizes by modeling three sources of variability: the sampling error of individual studies (level 1), variability within studies (level 2), and variability between studies (level 3; often referred to as “random effects”). To estimate the overall effect size, we fit an intercept-only 3LMA model. For moderator analyses, the dummy-coded categorical moderators were separately entered into the model, which were used to estimate the moderating relationship and the effect size within each subgroup of the moderator.

#### 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 produced multiple effect sizes. Thus, we examined two funnel plots: one with all effect sizes and one with dependent effect sizes aggregated. 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 further.

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 generally interpreted as an estimate of publication bias, and the intercept is interpreted as the bias-corrected overall effect. 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.

Third, we used weight-function modeling (Vevea & Hedges, 1995). In weight-function modeling, weighted distribution theory is used to model biased selection based on the significance of observed effects. 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 predictor to our overall-effect 3LMA. This allowed us to estimate the difference in the magnitude of published vs. unpublished effects.

## Results

Results indicated that, overall, explicit manipulations of demand characteristics cause participants’ responses to shift in a manner consistent with the communicated hypothesis, = 0.22, 95% CI [0.11, 0.33], = 3.93, < .001. For example, if participants were hypothetically told that an intervention should improve mood (positive demand), they would generally report slightly improved moods; if told that an intervention should worsen mood (negative demand), they would generally report slightly worsened moods.



*Figure* *2.*  Forest plot of estimated effect sizes (grey diamonds), their 95% confidence intervals (grey error bars), and their citations (left). The estimated effect size distribution is also shown and colored based on whether demand characteristics produce more hypothesis-consistent responding (green; d > 0.10), more hypothesis-inconsistent responding (red; d < -0.10), or negligible shifts in responding (grey; |d| < 0.10).

Although demand characteristics produce more hypothesis-consistent responding *on average*, demand effects are not consistent (between-study = 0.30; within-study = 0.20; Figure 2). For the sake of example, we arbitrarily classified any effect size less than 0.10 standard deviation in either direction as “negligible”. Based on the meta-analytic mean and standard deviation (between-study + within-study ), the estimated distribution of effects suggests that demand characteristics most often produce hypothesis-consistent shifts (63%), but sometimes produce negligible shifts (18%) or shifts in the opposite direction of the communicated hypothesis (19%)

the estimated distribution of effects suggests that demand characteristics (a) sometimes produce negligible shifts in responding and (b) other times cause participants’ responses to shift in the *opposite* direction of the communicated hypothesis (Figure 2). For the sake of example, we arbitrarily classified any effect size less than 0.10 standard deviation in either direction as “negligible”. Based on this classification, 63 % of demand characteristics manipulations produce hypothesis-consistent shifts in responding, 19 % produce hypothesis-inconsistent shifts, and 18% produce negligible shifts in either direction.

### Moderator analyses.

The observed variability in demand effects drastically exceeded what would be expected from sampling error alone, (194) = 901.77 , < .001. This suggests the existence of moderators.

Results indicated that the effects of demand characteristics tended to differ by participant pool, (2, 182) = 4.12, = .018. The effects were generally positive and medium-to-large in studies with students ( = 0.33, 95% CI [0.20, 0.46], < .001), positive and extremely small in studies with a mix of students and non-students ( = 0.04, 95% CI [-0.38, 0.47], = .838), and near-zero in studies with non-students ( = 0.00, 95% CI [-0.20, 0.20], = .993). The effects of demand characteristics also tended to be slightly more positive for in-person ( = 0.31, 95% CI [0.18, 0.45], < .001) vs. online ( = 0.09, 95% CI [-0.10, 0.28], = .373) studies, although this did not meet conventional thresholds of statistical significance, (1, 189) = 3.61, = .059.

The effects of demand characteristics appeared to be additive. Compared to instances where a demand characteristic condition was compared to a control group ( = 0.16, 95% CI [0.04, 0.28], = .009), effect sizes were approximately twice as large when two demand characteristic conditions were compared ( = 0.37, 95% CI [0.24, 0.51], < .001), (1, 193) = 19.26, < .001. Instances where a demand characteristic condition was compared to a control group allowed us to additional test whether participants respond more strongly to positive, nil, or negative demand characteristics. Results indicated that they might, (2, 131) = 5.41, = .006. The effect of demand characteristics tended to be nearly twice as large in the nil ( = 0.42, 95% CI [0.24, 0.60], < .001) vs. positive ( = 0.20, 95% CI [0.08, 0.32], = .002), and negative demand conditions ( = 0.16, 95% CI [0.01, 0.30], = .034).

We did not find that the effects of demand characteristics tended to differ depending on whether they were manipulated within- ( = 0.24, 95% CI [0.12, 0.36], < .001) vs. between-subjects ( = 0.09, 95% CI [-0.14, 0.32], = .427), (1, 193) = 1.66, = .199. We also did not find that the effects of demand characteristics differed by the year the record was completed or published, = 0.00, 95% CI [-0.01, 0.00], (194) = -0.51, = .607.

The effects of demand characteristics tended to be *numerically* larger in unpaid ( = 0.23, 95% CI [0.11, 0.35], < .001) vs. paid ( = 0.13, 95% CI [-0.05, 0.31], = .157) studies—but this difference was not statistically significant, (1, 192) = 0.87, = .352.

#### Exploratory attempt to reduce confounding (tmp).

The above moderator analyses indicated that demand effects tend to be positive in studies with student populations, in-person procedures, and no compensation. However, an exploratory inspection of the data revealed that these variables may be confounded. For example, students were more likely to be participants in in-person (61% student samples) vs. online (14% student samples) studies. Students were also more likely to be participants in unpaid (65% student samples) vs. paid (11% student samples) studies. In hindsight, this confounding seems obvious—but it was not anticipated when we pre-registered our analysis plan.

As an exploratory analysis, we fit a three-level meta-analysis with student status, data collection medium, and payment status entered as effect-coded factors. The results should be interpreted with some skepticism because the model may be overfit. Nonetheless, this exploratory analysis indicated that student status—but not data collection medium or payment status—was a significant moderator of demand effects. In other words, only student status is robustly associated with differences in demand effects.

### Publication bias analyses.

Overall, publication bias analyses were inconclusive. For instance, a funnel plot containing all effect sizes appeared to indicate 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 appeared to indicate 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 fit (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 provided a non-significant estimate of publication bias that favored hypothesis-consistent shifts in participants’ responses, = 0.68, 95% CI [-0.07, 1.44], = .076. The bias-corrected overall effect size estimate did not significantly differ from zero, = 0.06, 95% CI [-0.16, 0.27], = .606. On the other hand, two-level precision-effect tests with aggregated dependent effect size estimates yielded an opposite pattern: that there was a slightly (but not statistically significant) preference for non-significant or hypothesis-inconsistent shifts in participants’ responses, = -0.34, 95% CI [-1.39, 0.70], = .519. The bias-corrected overall effect size estimate was thus slightly adjusted upward, = 0.23, 95% CI [0.01, 0.45], = .038. In other words, depending on how dependencies were handled, precision-effect tests yielded inconsistent conclusions about the direction of publication bias and the significance of the bias-corrected overall effect of demand characteristics.

A weight-function model suggested that better fit was achieved with a model indicating that 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 ( = 0.46, 95% CI [0.00, 0.91], = .050) and published ( = 0.21, 95% CI [0.09, 0.32], < .001) studies yielded a similar pattern, although the difference was not statistically significant, (1, 193) = 1.08, = .301.



*Figure* *3.*  Raw (A) or aggregated (B) effect sizes ploted against their corresponding standard errors.

## Discussion

Study 1a provides the first quantitative synthesis of strict experimental tests of demand characteristics. Overall, explicit manipulations of demand characteristics cause participants’ responses to shift in a manner consistent with the communicated hypothesis. However, significant heterogeneity was observed. Using arbitrary thresholds, we estimated that 63% of demand characteristics manipulations produce hypothesis-consistent shifts ( > 0.10), 19% produce hypothesis-*in*consistent shifts ( < -0.10), and 18% produce negligible shifts in either direction (-0.10 < > 0.10). Moderator analyses revealed three study features that are associated with more hypothesis-consistent shifts in responses: (1) sampling student populations, (2) running studies in-person, and (3) communicating that the researchers hypothesizes there will be *no* shift in responses (i.e., using nil demand manipulations). We also found non-significant evidence of increases in hypothesis-consistent responding when participants were paid. However, attempts to unconfound these moderator analyses failed to provide robust evidence of moderation by in-person and payment status. Furthermore, publication bias analyses yielded opposite conclusions about the direction and impact of publication bias. Consequently, these results are non-definitive.

Study 1a provides preliminary insights on the magnitude, consistency, and contextual moderators of demand effects. However, it was not designed to evaluate outstanding questions regarding the extent to which these effects are driven by response bias vs. placebo effects. For example, consider our finding that demand characteristics tend to produce more hypothesis-consistent shifts in responses when students are sampled. If this is true, it may occur because students are more motivated to help the experimenter confirm their hypothesis (a response bias). Alternatively, it may occur because students are more likely to *believe* the communicated hypothesis (a placebo effect). 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 1b, we begin investigating this outstanding issue through an extension of the meta-analysis.

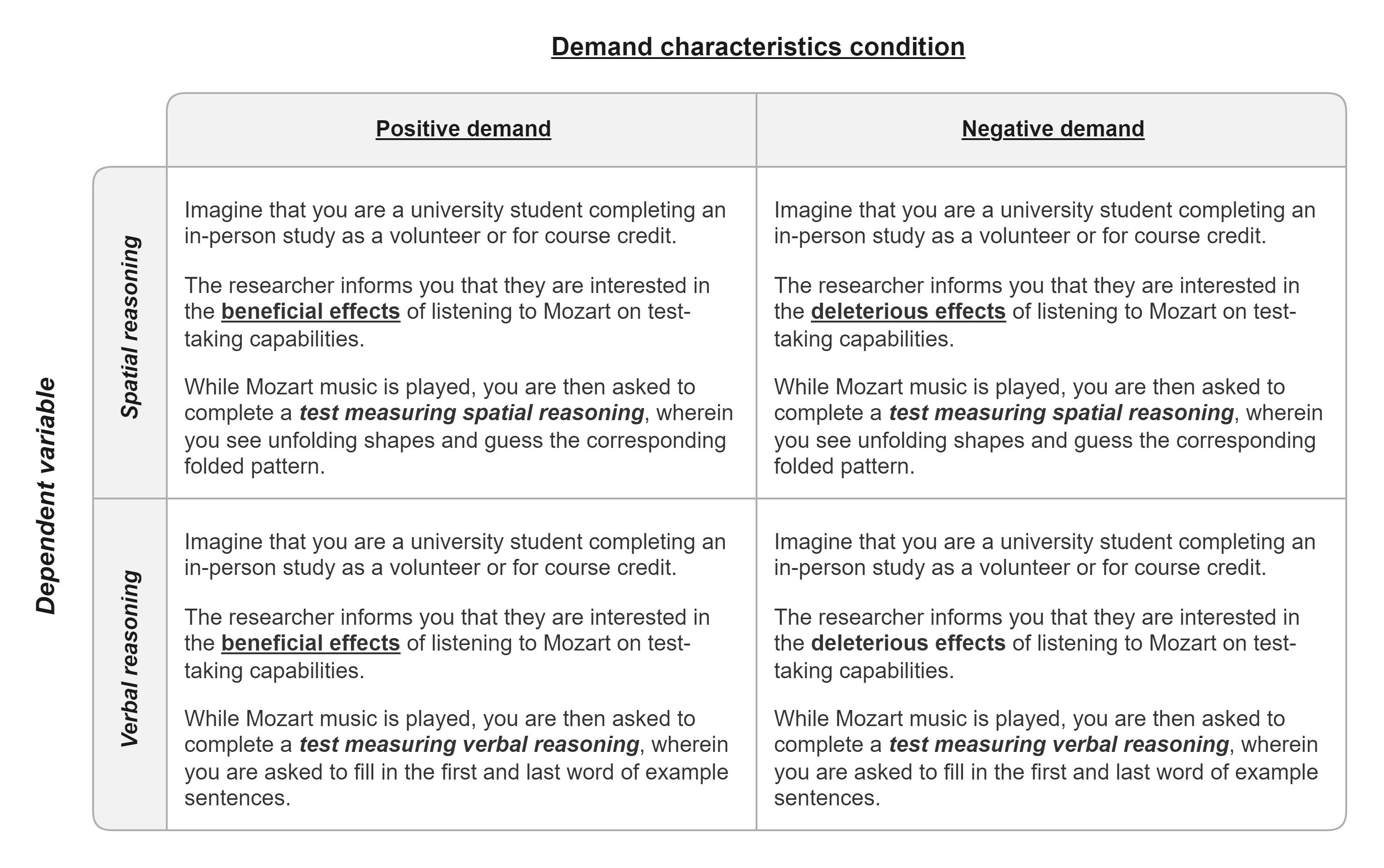
# Study 1b

Study 1b was designed to examine whether the meta-analytic effect size variability can be explained by factors theorized to underlie response biases (i.e., motivation and opportunity to adjust responses) and placebo effects (i.e., belief in the experimenter’s hypothesis; Figure 1). Unfortunately, these factors were rarely measured in the studies included in the meta-analysis. (See General Discussion for our call for more direct tests of underlying mechanisms.) Thus, we (a) estimated their values through a new set of participants and then (b) tested their moderating role by entering the values into meta-regressions. Also through meta-regression, we examined whether a new set of participants could retroactively predict the effects of the demand characteristic manipulations in the Study 1a meta-analysis.

## Methodology

For each study in the meta-analysis, 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 (see Figure 4).

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.



*Figure* *4.*  Vignettes for Standing et al. (2008), which described the key details for each demand characteristic condition (bolded and underlined) and dependent variable (bolded and italicized) combination.

Using a web-based survey, 224 undergraduates from Stanford University reviewed 10 randomly-selected vignettes in exchange for course credit. 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 hypothesis (-3 = “extremely motivated to adjust responses to be inconsistent” to 3 = “extremely motivated to adjust responses to be consistent”), (2) be able to adjust their responses on the outcome-of-interest (0 = “extremely incapable” to 4 = “extremely capable), and (3) believe the experimenter’s hypothesis (-3 =”strong disbelief” to 3 = “strong belief”). Raters also indicated whether they believed participants would change their responses to confirm the hypothesis, which we discuss later. These questions were presented in random order.

Ratings were removed in instances where the rater did not correctly identify the hypothesis communicated in the vignette. The remaining ratings were averaged across raters to provide mean estimates of motivation, opportunity, and belief.



*Figure* *5.*  Hypothetical data from a study where a procedure is either described as mood-boosting (positive demand), described as mood-dampening (negative demand), or not described at all (control). Data provides examples of how the effects of demand characteristics (d) on self-reported mood are moderating by participants’ reports of their motivation to confirm the stated hypothesis (m, Panel A), belief in the stated hypothesis (b, Panel B), and opportunity to adjust responses (o, Panel C). In each panel, separate examples are provided for scenarios where motivation is invariant (Column 1) and variant (Column 2) across demand characteristic manipulations

### Accounting for different demand comparisons.

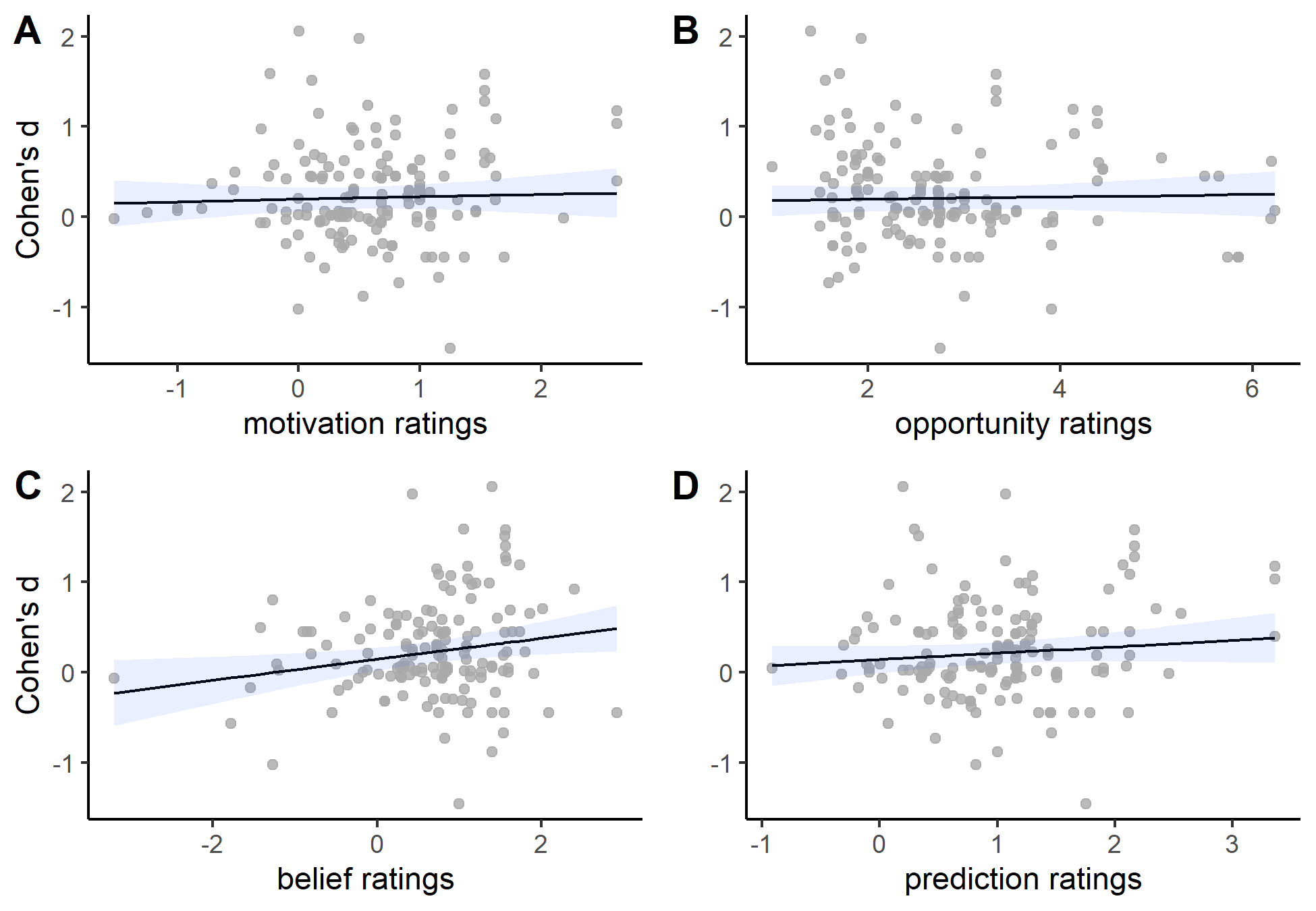
As mentioned before, Cohen’s represents the standardized difference between *two* groups. Thus, for each observation, we summed the motivation, opportunity, and belief ratings for the two groups being compared. Doing so allowed us to accommodate the fact that some comparisons involved two demand characteristics 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 (see Figure 5, Panel A, Column 1). When comparing the two demand conditions, the size of the demand effect should be doubled because the motivational forces in the two conditions produce an additive effect. In a different hypothetical context, these motivational forces could cancel each other out. This might happen if participants were (a) motivated to confirm the hypothesis in the positive demand condition, and (b) motivated to *dis*confirm the hypothesis in the negative demand condition (see Figure 5, Panel A, Column 2). Summing motivation scores allowed us to accommodate this possibility, and we used the same approach for belief (Figure 5, Panel B) and opportunity ratings (Figure 5, Panel C).

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 extremely 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.

### Rater forecasts of demand effects.

Even if researchers cannot explain how demand characteristics work, it might be valuable to be able to predict their effects (Yarkoni & Westfall, 2017). Orne suggested that one group that may be particularly good at predicting these effects is participants themselves (1969). To examine this, raters also predicted whether other participants would confirm vs. disconfirm the researcher’s hypothesis (-3 = “extremely likely to adjust responses to be inconsistent” to 3 = “extremely likely to adjust responses to be consistent”). We processed these data using the same approach as the motivation, opportunity, and belief scores (e.g., summed ratings when comparing two demand conditions).

## Results



*Figure* *6.*  The effects of demand characteristics on participants’ responses were not significantly moderated by motivation (Panel A) or opportunity (Panel B) ratings. They were, however, significantly moderated by belief (Panel C) and prediction (Panel D) ratings.

If demand effects are driven by response biases, their effects are expected to be moderated by participants’ motivation and opportunity to adjust responses (Figure 1). Inconsistent with this view, we did not find that demand effects were moderated by ratings of motivation, = 0.03, 95% CI [-0.09, 0.14], (150) = 0.47, = .640) or opportunity to adjust responses, = 0.01, 95% CI [-0.05, 0.08], (150) = 0.40, = .689 (Figure @(fig:modfig), Panels A and B).

If demand effects are driven by placebo, their effects should be moderated by participants’ belief in the communicated hypothesis. Consistent with this view, demand characteristic effects were positively associated with ratings of belief in the experimenter’s hypothesis, = 0.12, 95% CI [0.02, 0.21], (150) = 2.48, = .014 (Figure 6, Panel C). Last, we did find that raters’ predictions were significantly associated with observed demand effects, = 0.07, 95% CI [-0.03, 0.17], (150) = 1.37, = .172 (Figure 6, Panel D).

## Discussion

Contrary to classic conceptualizations of the impact of demand characteristics and comprehensive frameworks proposed by Rosnow and Rosenthal (1997) and Coles, Gaertner, et al. (2022), we did not find evidence of two moderators that have been theorized to underlie a response bias mechanism: motivation and opportunity to adjust responses. We did, however, find evidence that such effects are moderated by a measure of participants’ belief in the communicated effect. This provides preliminary evidence of a placebo-based mechanism that was initially proposed as an *extension* of classic demand characteristic frameworks.

To test the moderating role of participants’ motivation to adjust responses, opportunity to adjust responses, and belief in the experimenter’s hypothesis, we had to rely on ratings from an new set of participants. This was necessary because researchers have rarely measured these proposed moderators. However, it is not without limitations. First, it is possible that raters did not have enough information to make an accurate prediction about participants’ motivation, opportunity to adjust responses, and belief in the communicated hypothesis. For the sake of feasibility, we gave participants a short summary of the study. However, it is not clear if participants could accurately imagine the reality of being in these studies based on these relatively short descriptions. Indeed, when trying to gauge the impact of demand characteristics, Orne (1969) often would provide participants with extensive information about the study—perhaps even by running them through the procedures. It is thus possible that participants would have provided more valid ratings if we would have provided them with more information about the study (e.g., video recreations of the procedures).

Second, it is possible that our specific sample of raters—or maybe even modern-day participants in general—are not representative of the people sampled in previous research. In other words, maybe our 21th century Stanford University undergraduates have different study-related motivations, judgments, and beliefs than the participants who completed previous studies on demand characteristics. This seems likely to be true—but we did not find that it explains our pattern of results. To test the idea, we re-ran our motivation, opportunity, and belief moderator analyses focusing only on studies completed in the *past decade*. Doing so helped minimize differences between the participants who completed the original studies and our rating task. The pattern of results in this sensitive analysis, however, was largely the same as those from the full dataset.

To address these two major limitations, we re-examined the moderators in a small exploratory replication of an experiment in the demand characteristics literature.

# Study 2

After Study 1b participants completed the vignette ratings task, we had them complete an exploratory close replication of Coles, Gaertner, et al. (2022).

## Methodology

After completing the Study 1b ratings, we told 222 participants that we hypothesized that posed smiles will either increase (positive demand, n = 111) or not impact (nil demand, n = 111) feelings of happiness. Participants then posed happy and neutral expressions across two blocks. For happy poses, participants were instructed to move the corner of their lips toward their ears, elevating their cheeks. For neutral poses, participants were instructed to maintain a blank expression. Participants held each pose for 5 seconds with the assistance of an on-screen timer. After each pose, participants self-reported the extent to which they experienced happiness, satisfaction, and enjoyment (0 = “not at all” to 6 = “maximally”), which were averaged to form a happiness composite score. As filler items, participants also self-reported the extent to which they experienced fear (alarmed, scared, and fear) and anger (irritation, aggravation, and annoyance). Participants also completed several attention check items that instructed them to a choose a specific response option (e.g., “please choose ‘slightly’ as your response to this question”). Following Coles, Gaertner, et al. (2022), we excluded participants who failed an attention check (final n = 146). We report the two instances where this exclusion rule impacted our results.

Using similar items as Study 1b, participants at the end of the study reported the extent to which they were motivated to confirm the hypothesis, had the opportunity to adjust their responses, and believed in facial feedback effects. Altogether, the study used a 2 (facial pose: happy or neutral) × 2 (block: first or second) × 2 (demand characteristics: positive demand or nil demand) mixed design, with demand characteristics manipulated between subjects.

## Results

Following Coles, Gaertner, et al. (2022), we fit a mixed-effect regression with (a) facial pose, demand characteristics, and block number entered as effect-coded factors and (b) random-intercepts for participants. Results indicated that participants reported higher levels of happiness after posing happy vs. neutral expressions, *Mdiff* = 0.36, *F*(1, 428.20) = 158.77, *p* < .001. Furthermore, this effect was more pronounced in the positive (*Mdiff* = 0.90) vs. nil (*Mdiff* = 0.56) demand conditions, *F*(1, 428.20) = 9.06, *p =* .003.

Next, we examined the role of motivation, opportunity and belief. For each of these potential moderators, we fit mixed-effect regressions containing (a) facial pose and block number as effect-coded factors, (b) the moderator entered mean-centered as a continuous variable, (c) a higher-order facial pose by moderator interaction term, and (d) random intercepts for participants. The effect of facial poses on happiness tended to be larger among participants who reported being more motivated to confirm the hypothesis, = 0.10, *t*(431.38) = 3.95, *p =* < .001. However, this finding was not robust when including participants who failed an attention check, = 0.10, *t*(431.38) = 3.95, *p =* < .001. Further, we did not find evidence that the effect of facial poses on happiness was moderated by ratings of the perceived opportunity to adjust responses ( = 0.03, *t*(431.16) = 1.32, *p =* .187. However, consistent with previous evidence of placebo effects in facial feedback research (Coles, Gaertner, et al., 2022; Coles, March, et al., 2022), the effect of facial poses tended to be larger among participants who reported believing in the effect, = 0.05, *t*(431.20) = 2.94, *p =* .003

The previous analyses indicate that participants’ beliefs—and potentially also their motivation to provide hypothesis consistent responses—may moderate facial feedback effects. They do not, however, indicate whether the these factors moderate the effects of *demand characteristics*—i.e., whether there are three way interactions between (1) facial poses, (2) demand characteristics, and (3) ratings of motivation, opportunity, and/or belief. For each of those ratings, we fit separate mixed-effect regressions containing (a) facial pose and demand characteristics as effect-coded factors, (b) the rating entered mean-centered as a continuous variable, (c) all higher-order interactions, and (d) random intercepts for participants. Results indicated that that there was a three-way interaction between facial poses, demand characteristics, and participants’ self-reported motivation to provide hypothesis-consistent responses. Specifically, the interaction between facial poses and demand characteristics ( = 0.05, *t*(430.06) = 1.50, *p =* .133) tended to be larger among participants who provided higher motivation ratings, = 0.07, *t*(430.21) = 2.57, *p =* .010. This three-way interaction, however, was not robust when including participants who failed attention checks, = 0.04, *t*(584.31) = 1.81, *p =* .071.

We did not find evidence of a three-way interaction between facial poses, demand characteristics, and participants’ self-reported opportunity to adjust responses, = 0.00, *t*(430.13) = -0.09, *p =* .926.We did, however, find robust evidence of a three-way interaction involving self-reported belief in facial feedback effects. Specifically, the interaction between facial poses and demand characteristics ( = 0.08, *t*(430.15) = 2.69, *p =* .007) tended to be larger among participants with higher belief ratings, = 0.07, *t*(430.16) = 3.59, *p =* < .001.

To summarize, Study 2 provided mixed evidence of a response bias. We found some evidence that facial feedback and demand effects are moderated by self-reported motivation to provide hypothesis-consistent responses—but this finding was not robust across different analysis strategies. Furthermore, we consistently failed to find evidence that these effects were moderated by self-reported opportunity to adjust responses. We did, however, find consistent evidence that facial feedback and demand effects are moderated by self-reported belief in the communicated hypothesis.

# General Discussion

A comprehensive examination of strict experimental manipulations of demand characteristics reveal that they typically lead participants to slightly shift their responses in the direction of the communicated hypothesis. However, publication bias analyses are inconclusive, and the estimated effects are heterogeneous. Using admittedly arbitrary thresholds, we estimated that 63% of demand characteristics manipulations produce hypothesis-consistent shifts in participants’ responses ( > 0.10), 19% produce hypothesis-*in*consistent shifts in participants’ responses ( < -0.10), and 18% produce negligible shifts in either direction (-0.10 < > 0.10). Most worrisome, the current estimated distribution of demand effects suggests that they can range from approximately = -1.63 to = 2.02. *This range covers the magnitude of almost every conceivable effect in experimental psychology.* Thus, in order to distinguish theory-relevant effects from artifactual demand effects, it is essential that experimental psychologists understand when the latter bias their participant responses.

Participants themselves appeared to have little-to-no ability to predict the impact of demand characteristics, although it is possible that their performance would improve if they were provided with more information and better incentivized to provide accurate predictions. Unfortunately, it is also clear that our meta-analysis does not allow us to make predictions that are much better. Moderator analyses provided preliminary evidence that some methodological decisions—such as sampling students, running studies in-person, and not offering payment—are associated with increases in hypothesis-consistent responding. However, only student status yielded robust results across sensitivity analyses designed to reduce confounding. We also found that demand characteristics tended to be more impactful when a nil (as opposed to negative or positive) hypothesis is communicated. Nonetheless, most of the variability we observed in the meta-analysis is currently unaccounted for.

Demand effects would become easier to predict if we understood how they operate. Fortunately, an examination of the evidence we were able to synthesize provides *some* clues. Specifically, we found robust evidence that such effects are at least partly driven by participants’ beliefs. This challenges historical distinctions made between placebo effects and demand characteristics—the later which have been conventionally conceptualized as a relatively deliberate response bias driven by participants’ motivation and ability to adjust their responses. Nonetheless, our results do not rule out the possibility that demand characteristics produce *both* placebo effects and response biases. In the Study 1b meta-analysis, we did not find that external ratings of two factors theorized to underlie response biases—motivation and opportunity to adjust responses—moderated demand effects. However, we did find some evidence in Study 2 that motivation (but not opportunity) ratings moderated the effects of demand characteristics in a specific experimental context.

## Implications for conceptualizations of participant roles

In his pioneering work on demand characteristics, Orne (1962) characterized participants as “good subjects” who are motivated to help the researcher confirm their hypothesis. Our results—although not without their limitations—do not suggest that this is a prominent participant goal. Across all experimental contexts reviewed by Study1b participants, the estimated mean motivation to help confirm the study hypothesis was near zero (*M* = 0.42, *SD* = 0.53). To be sure, there were some study contexts in which participants tended to report slight motivation to confirm the experimenter’s hypothesis, such as when they were told the researcher expected them to (a) prefer a news article that favors their political party (*M* = 2.18), (b) avoid perceiving a shift in an optical illusion (*M* = 1.62), and (c) feel moved by happy and sad music (*M* = 1.36). However, this can alternatively be interpreted as participants’ motivation to respond in a manner that is consistent with their beliefs—not motivation to help the experimenter. Unfortunately, our results provide mixed evidence for this explanation. One one hand, the Study 1b meta-analysis indicated that there was a modest-sized correlation between study-level estimates of participants’ (a) motivation to confirm the hypothesis, and (b) belief in the hypothesis, = 0.52, 95% CI [0.38, 0.64], < .001. This suggest that, across all study contexts, participants were generally more motivated to confirm the hypothesis when it conformed with their own beliefs. However, this was not replicated in the specific experimental context examined in Study 2, where there was little-to-no correlation between motivation and belief ratings, = 0.04, 95% CI [-0.12, 0.20], = .613.

A placebo account—if true—could accommodate findings from many classic studies that seemingly demonstrated participants’ motivation to (a) help the experimenter, or (b) secure positive evaluations. For instance, when participants exhibited sham symptoms of hypnosis, Orne (1962) concluded that the participants did so to please the experimenter. However, an alternative explanation is that these participants were merely acting in accordance with their beliefs about the [sham] symptoms of hypnosis. Similarly, when participants reduced performance on a simple task after being told that high performance was indicative of an obsessive-compulsive personality, Sigall et al. (1970) concluded that participants did so to secure a positive evaluation. Once again, though, an alternative explanation is that these participants simply believed they did not possess a personality disorder and behaved accordingly.

Although our results most strongly support a placebo account of demand characteristics, it would be premature to dismiss conventional response bias accounts. Indeed, it seems likely that there are at least *some* contexts where participants are highly motivated to secure positive evaluations and/or help researchers confirm their hypotheses (even when controlling for beliefs). For instance, to avoid negative evaluations, participants may be unwilling to explicitly report racial biases—even if they (a) believe that the experimenter expects them to possess the bias, and (b) are consciously aware of the bias. Indeed, other reviews have provided evidence of this “socially desirable’ responding (Sedikides & Gebauer, 2010; Vesely & Klöckner, 2020; but see Lanz, Thielmann, & Gerpott, 2022). Conversely, there may be some contexts where participants are motivated to help the experimenter confirm their hypothesis—even if they don’t believe the hypothesis or think their response will impact how they’re evaluated.

## Future directions

In work originally published in 1969, McGuire (2009) suggested that there are three stages to a methodological artifact: ignorance, coping, and understanding/exploitation. At that same time, McGuire suggested that research on demand characteristics was entering the third stage. Unfortunately, over 50 years since McGuire’s initial publication, it would seem that only limited progress through this third stage has been made. We found very few direct tests of the mechanisms believed to underlie demand effects. Furthermore, our attempts to test these mechanisms through external ratings and a small replication study did not yield consistent for support pre-existing demand characteristic frameworks (e.g., Figure 1).

If researchers hope to progress through this third and final stage, we suggest that (1) theories will have to be refined, (2) mechanisms will have to be directly probed, and (3) reform for increasing the trustworthiness of study results will have to be implemented.

Theoretically, it is no longer tenable to keep demand characteristics conceptually divorced from related work on placebo effects. Consistent with recently proposed extensions of demand characteristic frameworks, our meta-analysis and replication study consistently indicated that *participants’* *beliefs* partially drive demand effects (Coles, Gaertner, et al., 2022; Corneille & Lush, 2022). This may occur because demand characteristics activate pre-existing beliefs about a phenomenon being investigated—but it is also possible that they cause participants to update pre-existing beliefs or form new beliefs. If true, research on how beliefs are formed, updated, and impact participant responses may help explain the unreliable effects of demand characteristic manipulations. For example, if beliefs 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 beliefs.

Methodologically, the mechanisms believed to underlie demand effects will have to be more directly probed through measurement and manipulation. For instance, similar to Study 2, researchers investigating demand characteristics could measure the extent to which participants believe the hypothesized effect and are motivated to (a) help the experimenter, (b) secure a positive evaluation, and/or (c) adjust their responses. These potential mechanisms could also be manipulated. For example, researchers could manipulate participants’ motivation to help the experimenter by providing financial incentives for doing so (Mummolo & Peterson, 2019).

Implementation-wise, we urge future demand characteristic researchers to engage in open science practices. Records of unpublished or “file-drawered” studies would help address our conflicting evidence regarding the existence and impact of publication bias. Access to open materials would better enable researchers to resolve discrepancies between previous studies through replication efforts (Coles, Tiokhin, Scheel, Isager, & Lakens, 2018; Zwaan, Etz, Lucas, & Donnellan, 2018). Last, open data and code would better allow researchers to verify published results, reproduce analytic workflows, and explore new questions through secondary analyses.

**Practical recommendations (or lack thereof)**

In his writings on demand characteristics, McGuire (2009) quipped that “one man’s artifact may be another man’s main effect” (p. 16). Although we argue that more attention to demand characteristics as a main effect is warranted, we acknowledge that, for most, it remains a pesty artifact. For those researchers, we suggest a major amendment to the playbook for avoiding the artifact. Following an influential framework developed by Rosnow and Rosenthal (1999), many researchers intuitively believe that demand effects can be avoided by using deception (reducing receptivity), incentivizing honest reporting (reducing motivation), and/or deploying difficult-to-control outcome measures (reducing opportunity to adjust responses). Evidence that demand characteristics can create placebo effects, however, suggest that these last two strategies will not be fully effective. Regardless of whether a participant is motivated or able to adjust their responses, awareness of the purpose of the study may unintentionally create placebo-induced changes in their responses. These placebo effects can no longer be a concern relegated to investigations of clinical outcomes. Just as placebo effects may bias estimates of the efficacy of an analgesic drug, they may bias our estimates of peoples’ attitudes, feelings, and behaviors.

Placebo effects can certainly be reduced—although it is not clear if they can be fully avoided. Existing demand characteristic frameworks suggest that they can at least be reduced by reducing receptivity, e.g., by using deceptive cover stories. However, it is important to note that participants’ possess a rich array of pre-existing beliefs before they step foot into our labs. For example, Coles, Gaertner, et al. (2022) found that approximately 44% of sampled undergraduates and 34% of sampled online workers believed—a priori—that facial poses impact emotion. Even with extensive deception about the purpose of the study, these pre-existing likely shape the extent to which participants exhibit facial feedback effects. In other words, extensive deception does not guarantee an unbiased estimate of a mechanism-of-interest. In the real world, the mechanisms that scientists theorize about may be naturally confounded with participants’ beliefs.

# Conclusion

We began our paper by mocking Crankology: a fictitious discipline plagued by a methodological artifact that could bias results in any direction, had unreliable effects, and had poorly understood mechanisms of action. However, our quantitative examination of a textbook methodological concern in experimental psychology—demand characteristics—raises humbling questions about the superiority of our own scientific endeavors. After all, the evidence we were able to synthesize indicates that demand characteristics also (a) can bias participant responses in any direction, (b) have heterogeneous effects, and (c) have somewhat unclear mechanisms of action. Contrary to conventional beliefs in the demand characteristics literature, our results provide evidence that placebo-based mechanisms may play a much larger role than previously believed. However, such conclusions are ultimately preliminary given the high heterogeneity, inconclusive publication bias analyses, and primitive measures of potential underlying mechanisms (Flake & Fried, 2020).

The estimated range of demand effects covers the span of almost every conceivable effect in experimental psychology. After taking stock of what we know about demand characteristics, we think it is *reasonable* to question whether the potentially valid methods of experimental psychology are distinguishable from the clearly invalid methods of Crankology. Despite our own crankiness, we remain somewhat optimistic—if, that is, researchers will continue to work

# To-do

[] Double check prediction scores.

[] Rosnow and Rosenthal were right that there can be counter and regular acquiesence.

[] can placebo account for superiority of nil demand manips?

[] Clean up folder structure

[] Codebooks

# References

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.

Barbuto Jr, J. E., & Scholl, R. W. (1998). Motivation sources inventory: Development and validation of new scales to measure an integrative taxonomy of motivation. *Psychological Reports*, *82*(3), 1011–1022.

Berkowitz, L. (1971). The" weapons effect," demand characteristics, and the myth of the compliant subject. *Journal of Personality and Social Psychology*, *20*, 332–338.

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.

Boudreaux, M. J., & Ozer, D. J. (2013). Goal conflict, goal striving, and psychological well-being. *Motivation and Emotion*, *37*(3), 433–443.

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*.

Coles, N. A., Larsen, J. T., & Lench, H. C. (2019). A meta-analysis of the facial feedback literature: Effects of facial feedback on emotional experience are small and variable. *Psychological Bulletin*, *145*(6), 610–651.

Coles, N. A., March, D. S., Marmolejo-Ramos, F., Larsen, J. T., Arinze, N. C., Ndukaihe, I. L., et al.others. (2022). A multi-lab test of the facial feedback hypothesis by the many smiles collaboration. *Nature Human Behaviour*, 1–12.

Coles, N. A., Tiokhin, L., Scheel, A. M., Isager, P. M., & Lakens, D. (2018). The costs and benefits of replication studies. *The Behavioral and Brain Sciences*.

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.

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.

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

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

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.

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

Kruglanski, A. W. (1975). The human subject in the psychology experiment: Fact and artifact. *Advances in Experimental Social Psychology*, *8*, 101–147.

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.

Lanz, L., Thielmann, I., & Gerpott, F. H. (2022). Are social desirability scales desirable? A meta-analytic test of the validity of social desirability scales in the context of prosocial behavior. *Journal of Personality*, *90*(2), 203–221.

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

McGuire, W. J. (2009). Suspiciousness of experimenter’s intent. *Artifacts in Behavioral Research. New York: Oxford*, 15–47.

Milgram, S. (1972). Interpreting obedience: Error and evidence. A reply to orne and holland. In A. G. Miller (Ed.), *The social psychology of psychological research* (pp. 138–154). New York, NY: Free Press.

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.

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.

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.

Sedikides, C., & Gebauer, J. E. (2010). Religiosity as self-enhancement: A meta-analysis of the relation between socially desirable responding and religiosity. *Personality and Social Psychology Review*, *14*(1), 17–36.

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.

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

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.

Vesely, S., & Klöckner, C. A. (2020). Social desirability in environmental psychology research: Three meta-analyses. *Frontiers in Psychology*, *11*, 1–9.

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.

Weber, S. J., & Cook, T. D. (1972). Subject effects in laboratory research: An examination of subject roles, demand characteristics, and valid inference. *Psychological Bulletin*, *77*(4), 273–295.

Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science*, *12*(6), 1100–1122.

Zion, S. R., & Crum, A. J. (2018). Mindsets matter: A new framework for harnessing the placebo effect in modern medicine. *International Review of Neurobiology*, *138*, 137–160.

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

1. “Crankology” is a portmanteau. It is a combination of the authors’ last names and the name of the discipline they study: psychology. [↑](#footnote-ref-1)