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

MARS guidelines: The problem of relation(s) under investigation; Study eligibility criteria; Type(s) of participants included in primary studies; Meta‐analysis methods (indicating whether a fixed‐effects or random‐effects model was used); Main results (including the more important effect sizes and any important moderators of these effect sizes); Conclusion (including limitations); Implications for theory, policy, and/or practice

*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: Colesology. The person explains that Colesology is useful for estimating causal relationships—but adds that it can sometimes be thrown off by a *methodological artifact.* When you ask the Colesologist 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. Because sometimes the artifact seems to matter and other times it doesn’t.

Like Colesology, experimental psychologists also deal with a mysterious methodological artifact: *demand characteristics*, or “cues which convey an experimental hypothesis to the subject” (Orne, 1962, p. 779). This artifact can be invoked in almost any study context. However, like the Colesologist, it is difficult for experimental psychologist to predict the direction, magnitude, and impact of this artifact in advance.

## Demand characteristics as a methodological artifact

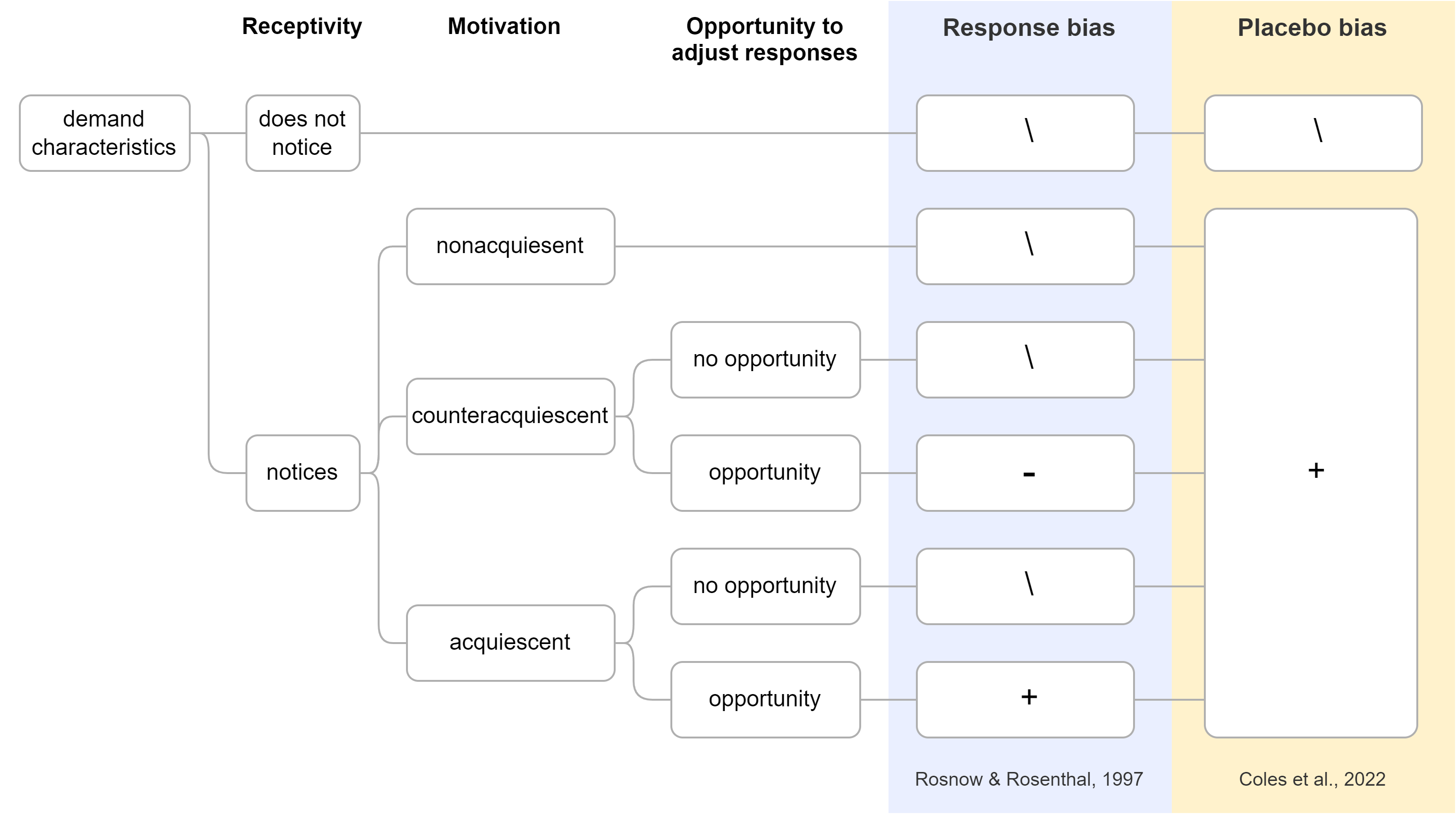
In 1962, Martin Orne published a seminal paper highlighting a view that challenged deeply-ingrained beliefs about experimental psychology. In this paper, Orne argued that research participants are not passive responders to the experimental context. Instead, he suggested that participants actively try to make sense of the situation based on their assumptions, beliefs, and motivations. One factor that Orne believed played a particularly powerful role were *demand characteristics*. This idea was controversial at first, with some researchers arguing that Orne’s claims about demand characteristics were 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. 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, but it is not clear when and how their effects emerge. The goal of the current paper is to use meta-analysis to take stock of what we know—and what we don’t know—about this methodological artifact. In Study 1, we provide the first quantitative synthesis of strict experimental tests of the effects of demand characteristics, with a focus on the the direction, magnitude, and consistency of the effect. We then examine several study features (e.g., whether participants are paid) that researchers have specified as potential moderators. In Study 2, we collect ratings from a new set of participants that provide tests of three mechanisms that may drive demand effects: motivation to confirm the experimenter’s hypothesis, perceived opportunity to adjust responses, and belief in the stated hypothesis. We also examine how well this new set of participants can predict the effects of demand characteristics. In Study 3, we present the results of a small replication study that re-examines motivation, opportunity, and belief as potential moderators of the effects of demand characteristics. We end with a discussion of the steps we believe are required to transform vague frameworks about demand characteristics into formal theories—theories we believe might help distinguish a valid experimental psychology from the invalid methods of Colesology.

## Rosnow and Rosenthal’s (1997) demand characteristics framework

Rosnow and Rosenthal (1997) proposed that there are three key moderators of the effects of demand characteristic: (1) receptivity to cues, (2) motivation to provide hypothesis-consistent responses, and (3) opportunity to alter their responses (Figure 1).



*Figure* *1.*  Rosnow and Rosenthal’s (1997) and Coles et al.’s (2022) frameworks for conceptualizing the impact of demand characteristics on participants’ responses.

### Receptivity to cues.

Rosnow and Rosenthal (1997) argued that participants must be perceptive to demand characteristics in order for them to impact downstream responses (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 researcher’s 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).

### Motivation to provide hypothesis-consistent responses.

Early in the history of research on demand characteristics, researchers debated which motivational forces typically underlie the effects of demand characteristics (for a review, see Weber & Cook, 1972). Orne (1962) originally characterized participants as “good subjects” who change their responses because they are altruistically 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 respond as naturally as possible (“faithful subjects”). Although seemingly divided, these early theorists agreed on one overarching principle: that participants’ motivation to provide hypothesis-consistent responses drives the effects of demand characteristics.

Because early demand characteristic theorists often focused on a single predominant subject goal–such as the goal to help the experimenter, be evaluated positively, or respond faithfully–less attention was paid to the notion that participants may have multiple, sometimes competing motivations (Barbuto Jr & Scholl, 1998; Boudreaux & Ozer, 2013). Indeed, when the idea of multiple motivations was examined, it was often done so to highlight the more prominent role of a specific goal (e.g., evaluation apprehension vs. motivation to help the experimenter, Sigall et al., 1970). However, Rosnow and Rosenthal (1997) found 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, imagine that an experimenter is friendly towards the participant and that the participant is thus motivated to help the experimenter. Now imagine that the participant learns that the experimenter hypothesizes that they will show a race-based preference for job applicants. In this scenario, the motivation to help the experimenter may conflict with the participant’s desire to respond in a socially desirable manner.

Based on the above observations and reasoning, Rosnow and Rosenthal (1997) suggested that participants can be characterized as being motivated to either (a) non-acquiesce (i.e., not change their responses), (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 might 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 to alter the outcome-of-interest. Thus, they posited that demand characteristics can only impact outcomes that participants can readily alter.

Taken together, Rosnow and Rosenthal (1997) posited that demand characteristics only bias participants responses when they (1) notice the cues, (2) are motivated to adjust their responses, and (3) are capable of adjusting their responses.

## Coles et al.’s (2022) framework

Demand characteristic theorists have traditionally conceptualized the effects of the methodological artifact as a *response bias* (Orne, 1962; Rosnow & Aiken, 1973; Strohmetz, 2008). For example, demand characteristics that indicate the researcher expects an intervention to boost mood is posited to impact participants’ mood *reports*–not necessarily their actual mood.

Coles et al. (2022) argued that demand characteristics not only have the potential to lead to response biases, but also placebo biases (Figure 1). They defined (a) response biases as changes mediated by relatively deliberate changes that participants make to their responses, and (b) placebo effects as changes mediated by relatively automatic activation of beliefs or pre-existing conditioned responses (Zion & Crum, 2018). Thus, unlike Rosnow and Rosenthal (1997), Coles et al. (2022) argued that demand characteristics can impact responses when participants have neither the motivation nor the opportunity to adjust their responses. Coles et al. (2022) also provided preliminary evidence that demand characteristics and placebo effects have distinct mechanisms. They found that participants’ beliefs did not always match the demand characteristics manipulation–and that both the demand characteristics manipulations and measures of participants’ beliefs moderated the effects of posed expressions on emotion.

# Methodology

The present meta-analysis was designed to (1) provide the first quantitative synthesis of strict experimental tests of demand effects, and (2) test predictions made by Rosnow and Rosenthal (1997) and Coles et al. (2022).

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 mechanism most often discussed in the demand characteristics literature: response biases (as opposed to placebo effects). Given that there is a sizable literature on placebo effects, excluding clinical tests of demand characteristics also helped us improve the feasibility of the project.

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 boundless and blurry 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 (see Limitations section). 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 additionally 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, but not always, this is compared to a *control* condition where participants were not told about an effect of an independent variable on a dependent variable.
* We excluded scenarios where the researcher described an effect that was *non-directional*. We did so because participants in these scenarios could not readily infer how their responses should 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 (positive demand) or dampened (negative demand).
* The demand characteristics manipulation was not strongly confounded. For example, a study by Sigall et al. (1970) was excluded because the manipulation of the stated hypothesis was confounded with a disclosure about the meaning of the behavior. Specifically, participants were either informed or not informed that the researcher expected them to copy a large quantity of numbers. When participants were informed about this hypothesis, they were also told that such behavior would be indicative of an undesirable personality trait.
* 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 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 about its expected effect.

In some scenarios, we estimated the *interactive* effect of 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 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 demand characteristics manipulation (i.e., acquiescence occurred). For example, if participants were told that an intervention should increase mood, an increase in mood would be coded as a positive effect. If participants were told that an intervention should decrease mood, 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 the 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 only 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 account for dependencies in our models (described later).

## Types of demand characteristic comparisons

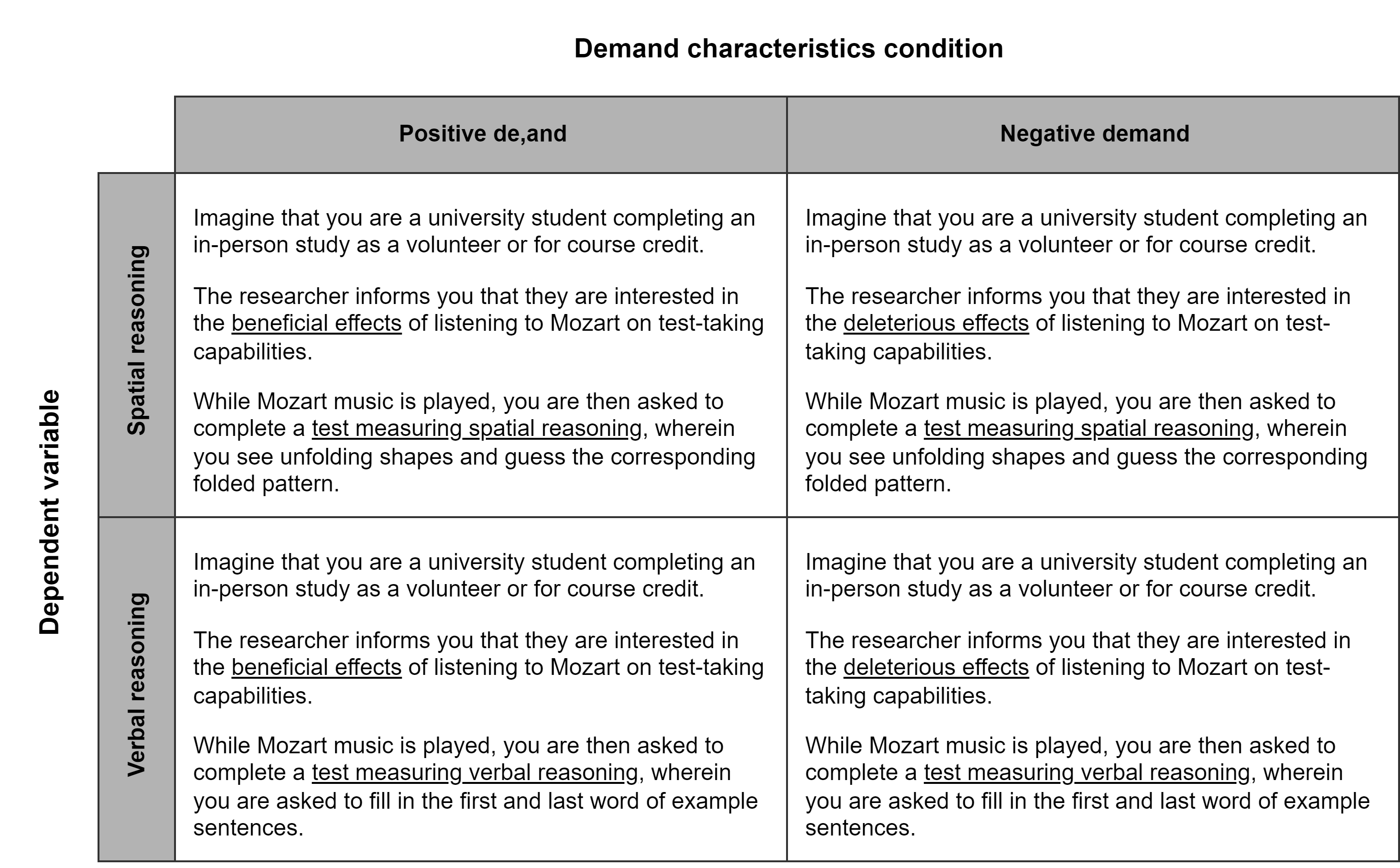
Cohen’s represents a standardized difference between two groups. Often, this involved a single demand characteristic condition (positive, negative, or nil demand) compared to a control group. Sometimes, however, this comparison involved two demand characteristic conditions (e.g., positive demand vs. negative demand). We thus coded whether the comparison involved one vs. two demand characteristic conditions. In addition, we coded each type of comparison: positive demand vs. control, nil demand vs. control, negative demand vs. control, positive vs. nil demand , positive vs. negative demand, and nil vs. negative demand.

## Post-hoc measures of motivation, opportunity, and belief

Both Rosnow and Rosenthal (1997) and Coles et al. (2022) posited that the effects of demand characteristics are moderated by participants’ (1) motivation to provide hypothesis-consistent responses, and (2) opportunity to adjust their responses (Figure 1). Coles et al. (2022) additionally predicted a third moderator: (3) participants’ belief in the hypothesized effect. Unfortunately, these variables were rarely measured in the studies included in the meta-analysis.

As an indirect measure of these moderators-of-interest, we estimated their values through a new set of participants. For each demand characteristic condition and dependent variable combination, we created vignettes that described key study details. For example, Standing, Verpaelst, and Ulmer (2008) had two demand characteristics 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 2).

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* *2.*  Vignettes for Standing et al., 2008.

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 a different 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 (1) be motivated to provide hypothesis-consistent responses (-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 the actual participants would confirm the hypothesis, which we discuss later. These questions were presented in random order.

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

### Accounting for different demand comparisons.



*Figure* *3.*  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 (c, Panel C). In each panel, separate examples are provided for scenarios where motivation is invariant (Column 1) and variant (Column 2) across levels of demand characteristics

As we’ve mentioned above, Cohen’s represents the standardized difference between *two* groups. Thus, for each effect size, 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 will have upward-biased responses in the positive demand condition and downward-biased responses in the negative demand condition (see Figure 3, 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. Alternatively, these motivational forces could hypothetically 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 disconfirm the hypothesis in the negative demand condition (see Figure 3, Panel A, Column 2). Summing motivation scores allowed us to accommodate this possibility, and we used the same approach for belief (Figure 3, Panel B) and opportunity ratings (Figure 3, 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 this condition. 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 (Orne, 1969). To examine this, raters who reviewed the vignettes 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).

## Other moderators

We also coded several moderators that researchers have speculated may moderate demand effects but for which we personally had no a-priori hypotheses. 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 the comparison involved one (e.g., positive demand vs. control) or two demand conditions (e.g., positive vs. negative-demand), (4) in the former, whether a positive, nil, or negative demand manipulation was used, (5) whether demand characteristics were manipulated within- vs. between-subjects, (6) the year the record was completed or published, and (7) 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 random-effect three-level meta-analysis (3LMA; also referred to as “multivariate” or “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 effect”). To estimate the overall effect size, we fit an intercept-only 3LMA model. For moderator analyses, continuous and dummy-coded categorical moderators were separately entered into the model. For categorical moderators, we used the models to estimate overall effect sizes 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. Consequently, we used three main approaches for assessing and correcting for potential publication bias in our overall effect size model.

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 omitted from the scientific record (i.e., there is publication bias), the distribution is often asymmetric or sloped. Funnel plots traditionally contain one effect size per study. However, given that many of our studies produced multiple effect sizes, we also examined funnel plots where non-independent effect sizes were aggregated.

Second, we conducted precision-effect tests (Stanley & Doucouliagos, 2014). In precision-effect tests, the relationship between observed effect sizes and their standard errors—which would be absent when there is no publication bias—is estimated and controlled for in a meta-regression model. This subsequently produces estimates of publication bias and 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, which is then compared to an unadjusted model. 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 meta-analyses with independent effect sizes, but it has fairly good statistical properties when non-independent effect sizes are aggregated (Rodgers & Pustejovsky, 2021). We thus aggregated non-independent effect sizes when using this procedure.

As a sensitivity analysis, we included publication status 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.

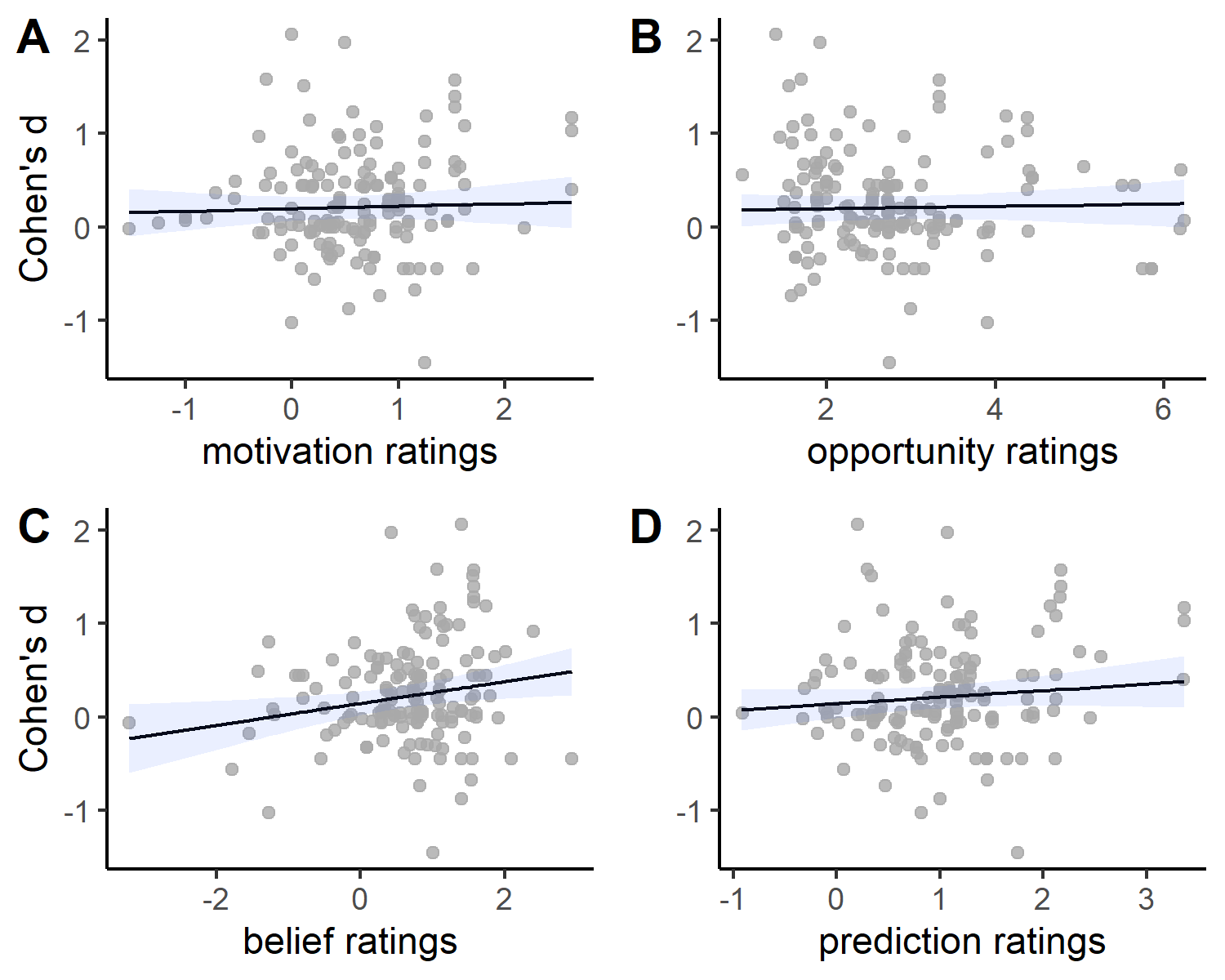
# Results

Results indicated that, overall, explicit manipulations of demand characteristics create a small acquiescence effect, = 0.22, 95% CI [0.11, 0.33], = 3.93, < .001. In other words, when explicitly told a hypothesis, participants responses tend to shift in a manner consistent with that hypothesis. 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.

### Moderator analyses.

The observed effects of demand characteristics were highly heterogeneous. Indeed, the between-study ( = 0.30) and within-study ( = 0.20) effect standard deviation was approximately the same size of the overall effect. This suggests that moderators (as opposed to mere sampling error) are producing meaningful shifts in the effects of demand characteristics.

#### Motivation, opportunity, and belief.



*(#fig:mod.fig)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.

Contrary to predictions made by Rosnow and Rosenthal (1997) and Coles et al. (2022), we did not find that the effects of demand characteristics were moderated by ratings of the extent to which participants would be motivated to confirm the experimenter’s hypothesis, = 0.03, 95% CI [-0.09, 0.14], (150) = 0.47, = .640 (Figure @ref(fig:mod.fig). We also did not find that demand characteristic effects were moderated by ratings of the extent to which participants could control the outcome of interest, = 0.01, 95% CI [-0.05, 0.08], (150) = 0.40, = .689 (Figure @ref(fig:mod.fig). Consistent with Coles et al. (2022), however, demand characteristic effects were larger when raters indicated that they believed the experimenter’s hypothesis, = 0.12, 95% CI [0.02, 0.21], (150) = 2.48, = .014 (Figure @ref(fig:mod.fig).

#### Other moderators.

Results indicated that the effects of demand characteristics were moderated by participant pool, (2, 182) = 4.12, = .018. The effects were medium-to-large in studies with students ( = 0.33, 95% CI [0.20, 0.46], < .001), 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 larger for in-person ( = 0.31, 95% CI [0.18, 0.45], < .001) vs. online ( = 0.09, 95% CI [-0.10, 0.28], = .373) studies, (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 do, (2, 131) = 5.41, .006. The effect of demand characteristics was 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 were moderated by whether researchers manipulated demand characteristics 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 were moderated 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 also not statistically significant, (1, 192) = 0.87, = .352.

### Forecasts.

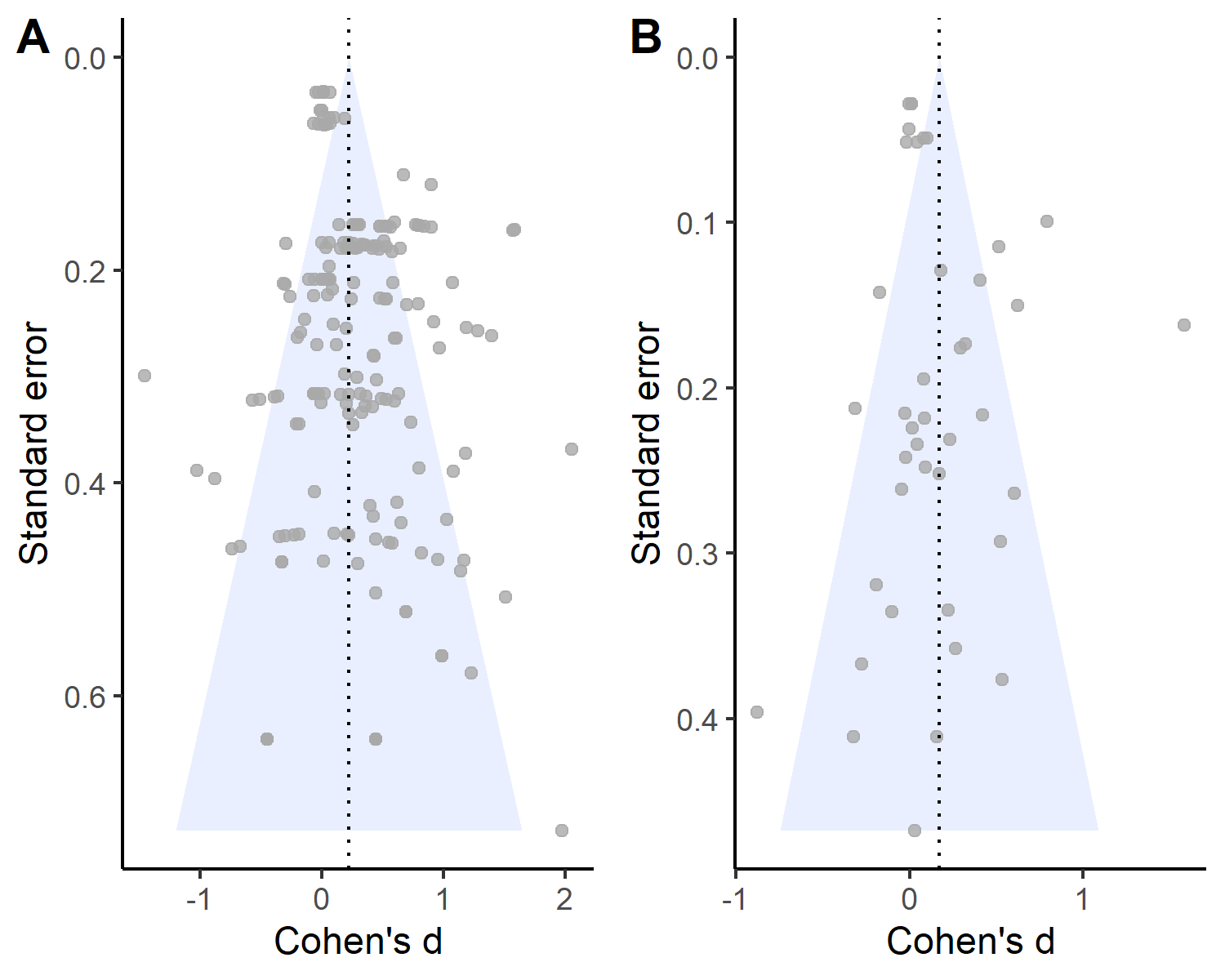
Raters predictions about the effects of demand characteristics were associated with the observed effects, but this association was weak, = 0.07, 95% CI [-0.03, 0.17], (150) = 1.37, = .172 (Figure @ref(fig:mod.fig).

### Publication bias analyses.

Overall, publication bias analyses were inconclusive. For example, a funnel plot containing all effect sizes appeared to indicate that publication bias favored acquiescence effects (i.e., effects where participants shifted responses to be consistent with the demand characteristics). However, a funnel plot where non-independent effect sizes were aggregated appeared to indicate the opposite: that publication bias favored non-significant or counter-acquiescence effects.

Precision-effect tests with 3LMA provided a non-significant estimate of publication bias that favored acquiescence effects, = 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. Precision-effect tests with aggregated non-independent effect sizes, however, estimated the opposite: that there was a slightly (but not statistically significant) preference for non-significant or counter-acquiescence effects, = -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.

The weight-function model suggested that better fit was achieved with a model indicating that publication bias favored non-significant or counter-acquiescence effects, (1) = 10.80, = .001. The bias-corrected overall effect size was thus upward-adjusted, = 0.41, 95% CI [0.19, 0.62], < .001. We did not, however, find evidence that effect sizes differed among unpublished ( = 0.46, 95% CI [0.00, 0.91], = .050) and published ( = 0.21, 95% CI [0.09, 0.32], < .001) studies (1, 193) = 1.08, = .301.



*Figure* *4.*  Funnel plot of raw (A) or aggregated (B) effect sizes ploted against their corresponding standard errors.

# Discussion

Contrary to classic conceptualizations of the impact of demand characteristics and frameworks proposed by Rosnow and Rosenthal (1997) and Coles et al. (2022), we did not find evidence of two moderators that have been theorized to underlie response biases: 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 unexpected pattern of results is consistent with a more parsimonious view: that phenomena typically described by experimental psychologists as “demand effects” are actually *placebo effects*. Applied to classic work by Orne, maybe participants did not exhibit sham symptoms of hypnosis because they were motivated to please the experimenter; instead maybe they intentionally or unintentionally responded in a manner that was consistent with their personal beliefs (a self-fulfilling prophecy or placebo effect).

## Moderator analysis limitations and converging evidence

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 raters. This was necessary because researchers have rarely measured these proposed moderators—but it is not without limitations.

First, it is possible that raters did not have enough information to make an accurate prediction about participants’ motivation, ability to adjust responses, and belief in the experimenter’s communicated hypothesis. For the sake of feasibility, we gave participants a short summary of the hypothesis communicated by the researcher and the study set-up. However, it is not clear if participants could accurately imagine the reality of being in these studies based on this relatively short description. 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 some of the procedures. It is thus possible that raters would have provided more valid ratings if more information was provided (e.g., video recordings of recreations of the study).

Second, it is possible that our specific sample of raters—or maybe even modern-day participants in general—are not representative of the participants 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 have completed previous studies on demand characteristics. This seems likely to be true. However, we do not believe it explain our pattern of results. To test it, we re-ran our motivation, opportunity, and belief moderator analyses focusing only studies completed in the *past decade*. Focusing on the past decade, we believe helped minimize differences between participants who completed the original studies and participants who completed our rating task. The pattern of results, though, were largely the same.

Last, like in most meta-analysis, it is possible that our moderator analyses were underpowered, contained unidentified confounds, and/or were impacted by publication bias. Unfortunately, though, our publication bias analyses were too inconclusive to weigh in on this issue.

(Also, maybe the ratings weren’t that precise)

### Converging evidence from a replication study.

Despite the limitations of our approach for studying the moderating role of motivation, opportunity, and belief, we observed similar patterns in a large replication of Coles et al. (2022). After our participants complete the vignette ratings task, we had them complete a close replication of Coles et al. (2022). At the beginning of this study, we told 222 participants that we hypothesized that posed smiles will either (a) increase (positive demand, n = 111) or (b) not impact (nil demand, n = 111) feelings of happiness. Participants than posed happy and neutral expressions across two blocks. After each pose, participants self-reporting 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). Using similar items as the vignette study, participants at the end of the study reported the extent to which they were motivated to confirm the hypothesis, able to adjust responses, and believed the hypothesized effect.

Replicating Coles et al. (2022), a mixed-effect regression with (a) facial pose, demand characteristics, and block entered as effect-coded factors and (b) random-intercepts for participants, indicated that participants reported higher levels of happiness after posing happy vs. neutral expressions, *Mdiff* = 0.66, *F*(1, 581.62) = 180.15, *p* < .001. Further, this effect was more pronounced in the positive (*Mdiff* = 0.83) vs. nil (*Mdiff* = 0.55) demand conditions, *F*(1, 581.62) = 7.75, *p =* .006.

Most importantly, though, we replicated the pattern of results from the meta-analysis. For each moderator (motivation, opportunity, and belief), we fit a mixed-effect regression containing (a) trial and block number as effect-coded factors, (b) the moderator entered as a continuous variable, (c) a higher-order trial by moderator interaction term, and (d) random intercepts for participants. We did not find that the effect of facial poses was moderated by ratings of motivation to confirm the hypothesis (*F*(1, 585.46) = 2.46, *p =* .117 or opportunity to adjust responses (*F*(1, 585.25) = 1.48, *p* = .224. We did, however, find that the effect of facial poses was moderated by participants’ beliefs about the hypothesized effect (*F*(1, 583.41) = 6.42, *p =* .012. Specifically, the facial feedback effect was larger among participants who more strongly believed in the effect, $\beta = 0.63.

### Can a “demand as placebo” account explain discrepancies in the literature?.

Despite the allure of the parsimonious “demand characteristics as placebo effects” account, belief ratings did not fully explain observed discrepancies in the demand characteristics literature. For example, the mean belief ratings in the largest demonstration (Coles et al., 2022) and largest failure-to-demonstrate demand effects were quite similar (Mummolo & Peterson, 2019): M = X, and M = Y respectively. That being said, there are several limitations to our approach to testing these potential moderators (see Limitation sections). Thus, we believe future research should examine this question through replications of the Coles et al. (2022) and Mummolo and Peterson (2019) procedures. Like our own replication, of Coles et al. (2022), this may involve asking participants questions about their motivation to confirm the experimenter’s hypothesis, perceived ability to adjust their responses, and belief in the hypothesized effect at some point in the experiment. Alternatively, future researchers may find it useful to *manipulate* these proposed moderators. This may involve, for instance, providing incentives to increase participants motivation, relying on indirect and difficult-to-control measures of the dependent variables, and providing participants with information that make the communicated hypothesis more or less believable.

## Can participants help us predict the effects of demand characteristics?

Even in the continued absence of a clear explanation of how demand characteristics work, it would certainly be valuable to be able to predict their effects (Yarkoni & Westfall, 2017). Orne suggested that participants themselves may be particularly adept at making these predictions—but we did not find much evidence in support of this assertion.

It is certainly possible that participants are unaware of how and why demand characteristics impact their responses in experiments. Alternatives:

* They didn’t have enough information to make a good judgment. We only gave them a short summary of the hypothesis communicated and experimental set-up. Maybe if they had more information—such as a video recording of the actual experiment—they could make good predictions. This is more similar to what Orne did.
* There’s a history effect. Modern-day participants’ motivations and belief are different than the people that Orne worked with a half-century ago. Modern day participants may be able to predict the results of modern-day experiments–but not historical ones. Intriguingly, when we [arbitrarily] limited analyses to observations made in the past 10 years, the patterns of results were similar.
* Could be a statistical power issue. These data are really messy. But one should then still expect to see it when working with the cleaner replication data, which we didn’t

##   
## Multivariate Meta-Analysis Model (k = 64; method: REML)  
##   
## Variance Components:  
##   
## estim sqrt nlvls fixed factor   
## sigma^2.1 0.0085 0.0919 15 no id.study   
## sigma^2.2 0.0018 0.0424 64 no id.study/id.es   
##   
## Test for Residual Heterogeneity:  
## QE(df = 62) = 162.0633, p-val < .0001  
##   
## Test of Moderators (coefficient 2):  
## F(df1 = 1, df2 = 62) = 2.2624, p-val = 0.1376  
##   
## Model Results:  
##   
## estimate se tval df pval ci.lb ci.ub   
## intrcpt 0.0508 0.0371 1.3712 62 0.1753 -0.0233 0.1249   
## pre 0.0331 0.0220 1.5041 62 0.1376 -0.0109 0.0772   
##   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

##   
## Multivariate Meta-Analysis Model (k = 64; method: REML)  
##   
## Variance Components:  
##   
## estim sqrt nlvls fixed factor   
## sigma^2.1 0.0078 0.0884 15 no id.study   
## sigma^2.2 0.0027 0.0518 64 no id.study/id.es   
##   
## Test for Residual Heterogeneity:  
## QE(df = 62) = 163.8966, p-val < .0001  
##   
## Test of Moderators (coefficient 2):  
## F(df1 = 1, df2 = 62) = 0.9579, p-val = 0.3315  
##   
## Model Results:  
##   
## estimate se tval df pval ci.lb ci.ub   
## intrcpt 0.0751 0.0307 2.4445 62 0.0174 0.0137 0.1366 \*   
## mot 0.0236 0.0241 0.9787 62 0.3315 -0.0246 0.0717   
##   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

##   
## Multivariate Meta-Analysis Model (k = 64; method: REML)  
##   
## Variance Components:  
##   
## estim sqrt nlvls fixed factor   
## sigma^2.1 0.0102 0.1010 15 no id.study   
## sigma^2.2 0.0013 0.0360 64 no id.study/id.es   
##   
## Test for Residual Heterogeneity:  
## QE(df = 62) = 164.5483, p-val < .0001  
##   
## Test of Moderators (coefficient 2):  
## F(df1 = 1, df2 = 62) = 0.2365, p-val = 0.6285  
##   
## Model Results:  
##   
## estimate se tval df pval ci.lb ci.ub   
## intrcpt 0.0613 0.0611 1.0032 62 0.3196 -0.0608 0.1833   
## opp 0.0083 0.0171 0.4863 62 0.6285 -0.0258 0.0424   
##   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

##   
## Multivariate Meta-Analysis Model (k = 64; method: REML)  
##   
## Variance Components:  
##   
## estim sqrt nlvls fixed factor   
## sigma^2.1 0.0104 0.1019 15 no id.study   
## sigma^2.2 0.0013 0.0362 64 no id.study/id.es   
##   
## Test for Residual Heterogeneity:  
## QE(df = 62) = 162.9602, p-val < .0001  
##   
## Test of Moderators (coefficient 2):  
## F(df1 = 1, df2 = 62) = 3.3539, p-val = 0.0719  
##   
## Model Results:  
##   
## estimate se tval df pval ci.lb ci.ub   
## intrcpt 0.0541 0.0363 1.4890 62 0.1416 -0.0185 0.1267   
## bel 0.0466 0.0254 1.8314 62 0.0719 -0.0043 0.0975 .   
##   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

## $mod  
##   
## Multivariate Meta-Analysis Model (k = 64; method: REML)  
##   
## Variance Components:  
##   
## estim sqrt nlvls fixed factor   
## sigma^2.1 0.0085 0.0919 15 no id.study   
## sigma^2.2 0.0018 0.0424 64 no id.study/id.es   
##   
## Test for Residual Heterogeneity:  
## QE(df = 62) = 162.0633, p-val < .0001  
##   
## Test of Moderators (coefficient 2):  
## F(df1 = 1, df2 = 62) = 2.2624, p-val = 0.1376  
##   
## Model Results:  
##   
## estimate se tval df pval ci.lb ci.ub   
## intrcpt 0.0508 0.0371 1.3712 62 0.1753 -0.0233 0.1249   
## pre 0.0331 0.0220 1.5041 62 0.1376 -0.0109 0.0772   
##   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
##   
## $sub  
##   
## Multivariate Meta-Analysis Model (k = 64; method: REML)  
##   
## Variance Components:  
##   
## estim sqrt nlvls fixed factor   
## sigma^2.1 0.0069 0.0830 15 no id.study   
## sigma^2.2 0.0024 0.0490 64 no id.study/id.es   
##   
## Test for Residual Heterogeneity:  
## QE(df = 63) = 162.8975, p-val < .0001  
##   
## Model Results:  
##   
## estimate se tval df pval ci.lb ci.ub   
## pre 0.0544 0.0175 3.1139 63 0.0028 0.0195 0.0894 \*\*   
##   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

not good at predicting

Although participants’ failure to predict responses to demand characteristics manipulations may be due to issues with our approach to testing this moderator (see Limitations section), it is possible that participants do not correctly guage

suggested that one group that may be particularly good at predicting these effects is participants themselves (Orne, 1969).

of belief

This still doesn’t make sense when explaining why Mummolo didn’t find the effect. Had above average belief ratings

## # A tibble: 1 x 1  
## m.bel  
## <dbl>  
## 1 0.636

## # A tibble: 1 x 1  
## m.bel  
## <dbl>  
## 1 0.527

## Implications

Based on the belief that demand characteristics is a response bias moderated by motivation and opportunity to adjust responses, two recommendations have been proposed to combat demand characteristics: lower participants’ (1) motivation to adjust responses (e.g., by emphasizing the importance of natural responding), and/or (2) ability to adjust responses (e.g., by using difficult-to-control measures, like implicit or physiological measures). The “demand-as-placebo” account, however, suggests that these approaches are likely to be ineffective. Regardless of whether participants are motivated or able to adjust responses, knowledge of the experimenter’s hypothesis can create placebo-induced changes in their responses.

# Limitations

Our literature search was non-comprehensive.

* For our literature review, we focused on identifying a workable conceptual space, focusing on explicit experimental manipulations of the communicated hypothesis.
* But demand characteristics is a broad–and some might even argue vague–construct. Anything that potentially influences participants beliefs about the experimenter’s hypothesis can be considered demand characteristics, including the text in informed consents and study materials, the experimenter’s delivery, etc.
* Furthermore, Orne’s writings on demand characteristics often included discussions of even broader (but difficult to measure) aspects of the social situations, including social norms about participant behavior, pacts of ignorance about admitting awareness of the true purpose of the study, etc.

Publication bias analyses are problematic. Evidence of moderation is correlational. There may be confounding.

# thoughts

Compare average effect of demand to average effect in psychology. It suggest that it’s plausible that these are driven by demand. When looking at the distribution, it’s clear that no effect is too big to rule out concerns about demand.

Might want to plot es distribution and identify proportion of (a) non-negligible acquiescence, (b) non-negligible counter-acquiescence, and (c) non-acquiescence.

### Exploratory stuff.

##   
## Test of Moderators (coefficients 2:4):  
## F(df1 = 3, df2 = 148) = 2.2156, p-val = 0.0887

# To-do

[] Look into bug that’s causing studies to only have one effect size

[] Add forest plot

[] F-value looks off here “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.”

[] Clean up folder structure

[] Have Mike review again?

[] Have M.W. work on outcome type

[] Have M.W. build 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*.

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.

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.

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.

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

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.

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.

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.