A meta-analysis of the effects of demand characteristics

Nicholas A. Coles1, Morgan H. Wyatt1, & 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

TBD

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

*Word count:* TBD

A meta-analysis of the effects of 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 in a new approach for estimating causal relationships. However, they adds that the method can sometimes be thrown off by a *methodological artifact.* You ask the Colesologist about this artifact. They explain that, sometimes, it causes researchers to detect an effect that’s not real; other times, it causes researchers to miss an effect that *is* real. They add that sometimes the artifact causes estimated relationships to be biased upward; other times, it causes the estimates to be biased downward. And, in general, they explain, it means that the things researchers observe using Colesology don’t necessarily capture real-world human behavior.

At this point, you might become skeptical and ask the Colesologist, “How does this methodological artifact work?” Their response? They don’t know, because sometimes the artifact seems to matter and other times it doesn’t.

If this scenario were real, you would reasonably question whether Colesology is a valid method of scientific inquiry. But here’s the thing: we’re not describing Colesology, we’re describing experimental psychology.

## 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. 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 was *demand characteristics*: “cues which convey an experimental hypothesis to the subject” (p. 779). Not everybody agreed about the importance of these demand characteristics (e.g., Berkowitz, 1971; Milgram, 1972). Nonetheless, over the next 60 years, demand characteristics become recognized as a literal textbook methodological concern in experimental psychology (Sharpe and Whelton, 2016).

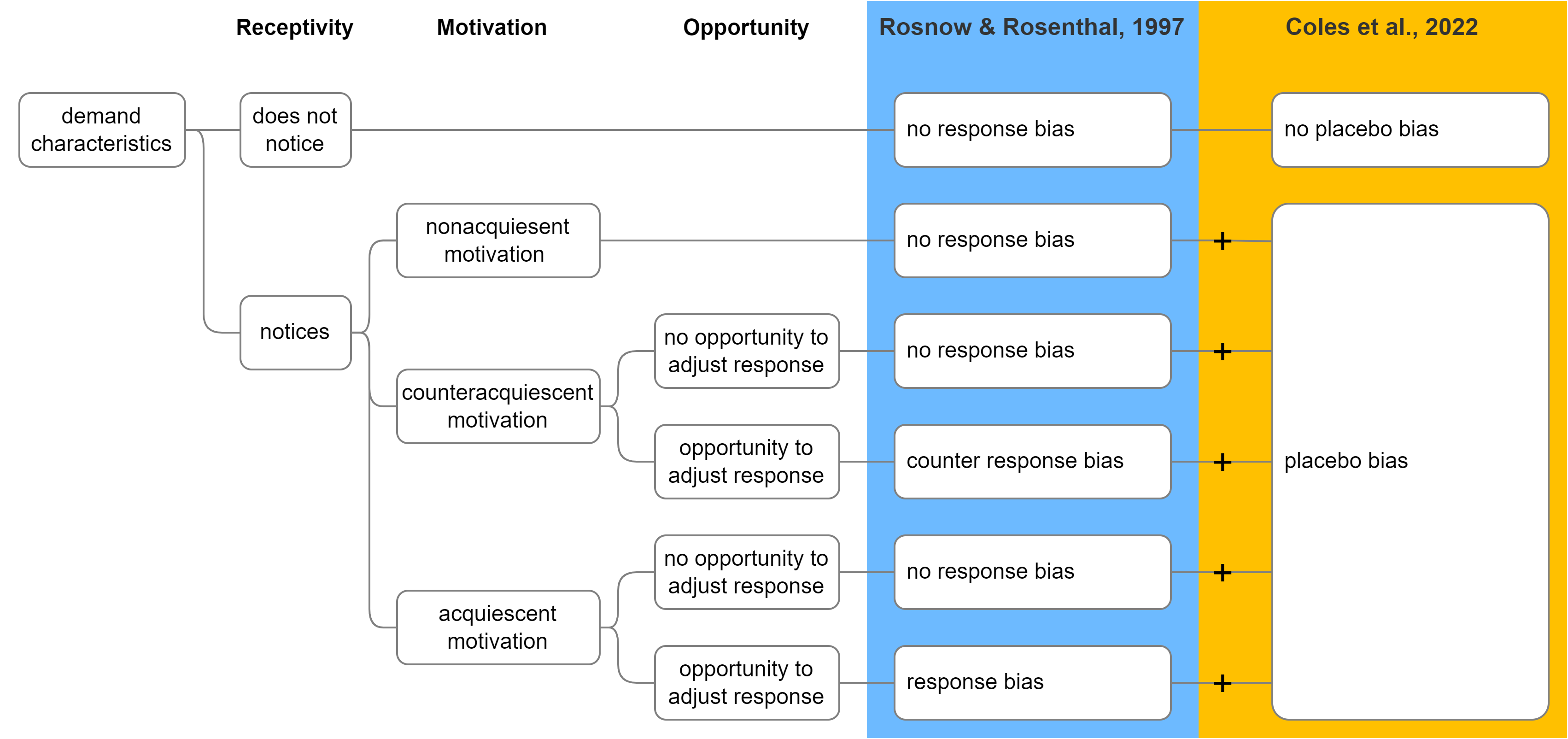
Orne initially focused on evidence that demand characteristics can lead to false positives—such as patients exhibiting sham symptoms of hypnosis (Orne, 1959). Follow-up research, though, indicated that demand characteristics can also lead to false negatives. For example, Hayes and King (1967) demosntrated 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 also bias estimates of causal relationships. For example, Coles et al. (2022) found that the estimated effect of facial poses on self-reported emotion could be amplified *or* attenuated based on the communicated purpose of the study. Yet, it’s still not clear when demand characteristics matter. For example, in large replications of classic studies in behavioral economics, Mummolo and Peterson (2021) consistently failed to find that manipulations of their communicated hypothesis impacted participants’ responses.

After over 60 years, experimental psychologists are left with an uncomfortable truth: 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 don’t know) about demand characteristics.

We first provide an overview of a framework designed to accommodate the potentially varied effects of demand characteristics (Rosnow & Rosenthal, 1997). We also review a complementary framework that attempts to bridge the gap between research on demand characteristics and research on placebo effects (Coles et al., 2022). We then use meta-analysis to conduct the first quantitative synthesis of strict experimental tests of the effects of demand characteristics. Through this meta-analysis, we not only estimate the overall impact of demand characteristics, but also use moderator analyses to provide preliminary tests of predictions made by Rosnow and Rosenthal (1997) and Coles et al. (2022). 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 distinguishing 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 the cues.

Rosnow and Rosenthal (1997) argued that participants must be perceptive to demand characteristics in order for them to impact downstream responses (Rosnow & Aiken, 1973; Strohmetz, 2007). 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 response bias (for a review, see Rosnow and Rosenthal, 1997; Weber and 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 cause them to be evaluated positively (Riecken, 1962; Rosenberg, 1969, 2009; Sigall, Aronson, and Van Hoose, 1970). Masling (1966) argued that participants sometimes interfere with the purpose of the study (“negativistic subjects”; see also Cook et al. 1970; Silverman, 1977), whereas Fillenbaum (1966) argued that participants attempt to respond as naturally as possible (“faithful subjects”). Although seemingly divided, these early theorists actually agreed on one overarching principle: that a key driver of the effects of demand characteristics is participants’ motivation to provide hypothesis-consistent responses.

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 and Scholl, 1998; Boudreaux & Ozer, 2013). Indeed, when the idea of multiple motivations was explored, it was often done so to highlight the more prominent role of a single goal (e.g., evaluation apprehension vs. motivation to help the experimenter; Sigall, Aronson, and Hoose, 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 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 can also be conceptualized on a continuum ranging from highly motivated to acquiesce to highly motivated to counter-acquiesce.

### Opportunity to alter responses.

No matter how motivated they are to confirm the hypothesis, Rosnow and Rosenthal (1997) suggested 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 impact outcomes that participants can readily alter.

## Coles et al.’s (2022) framework

Researchers have generally conceptualized the effects of demand characteristics on participants’ responses as a *response bias* (Orne, 1962; Rosnow and Aiken, 1973; Strohmetz, 2007). For example, demand characteristics that indicate the researcher expects an intervention to boost mood is *not* posited to impact participants’ actual mood; Instead, the demand characteristics are posited to merely impact participants’ mood *reports.*

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 that are mediated by the relatively automatic activation of beliefs or pre-existing conditioned responses (Zion and Crum, 2018). Thus, unlike Rosnow and Rosenthal (1997), Coles and colleagues argued that demand characteristics can impact responses even when participants have neither the motivation nor opportunity to adjust their responses. Preliminary evidence for this assertion comes from Coles et al.’s observation that participants’ beliefs did not always match the demand characteristics manipulation. For example, some participants disclosed that they (a) did not personally believe that posed expressions impacted emotion, but (b) recognized that the experimenter did. Both the manipulation of demand characteristics and measures of participants’ beliefs independently moderated facial feedback effects, providing preliminary evidence of distinct psychological mechanisms.

# Methodology

We describe the scope of our meta-analysis using the Population, Intervention, Comparison, Outcome framework (Schardt et al. 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. We focused on this intervention because it provides a relatively overt test of the impact of demand characteristics.

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 infeasible (see Limitations section). Thus, we sought to design a strategy that best balanced comprehensiveness and feasibility.

We searched APA PsycInfo using relatively 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; Facebook Psychological Methods Discussion Group and PsychMAP groups. 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.
* We excluded scenarios where the researcher described an effect that was non-nil and non-directional. We did so because participants in these scenarios could not readily infer how to adjust their responses. 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 (nil demand).
* The demand characteristics manipulation was not strongly confounded. For example, a study by Sigall and Adair (1974) 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 M. W. screened records independently, reviewed potentially relevant records together, and coded the information for moderator analyses and effect size computations. Disagreements were resolved through discussion. It total, 41 studies from 30 records were eligible for inclusion.

## Effect size index

We used standardized mean difference scores as our effect size index (Cohen’s and ; Borenstein, 2009; Cohen, 1988).

In some scenarios, we estimated the main effect of demand characteristics. For example, Coles et al. (2022) manipulated whether participants were told that smiling 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 other scenarios, we estimated the interactive effect of demand characteristics. For example, in the same Coles et al. (2022) study, participants reported happiness 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 difference-in-differences score.

Effect sizes were calculated so that positive values indicated an effect consistent with the demand characteristics manipulation. 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.

For repeated-measure comparisons, the correlation between the repeated measures is needed to calculate Cohen’s $d\_{rm}$. This correlation is rarely reported, so we followed Borenstein’s (2009) recommendation 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.

Often, studies contained multiple effect sizes of interest. For example, Coles et al. (2022) had 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

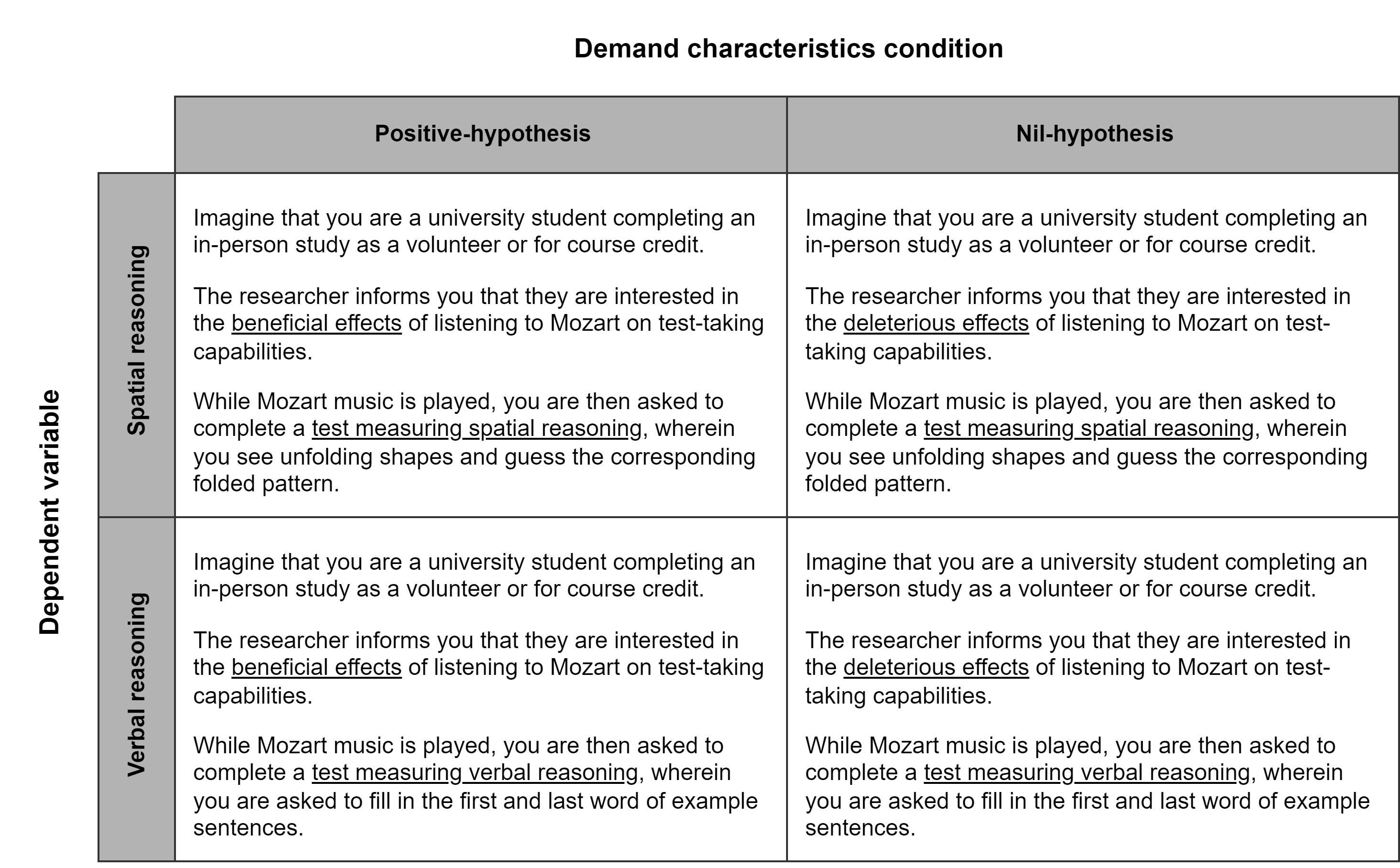
Cohen’s represents a standardized mean difference between two groups. Often, this comparison involved a single demand characteristic condition (positive, negative, or nil demand) compared to a control group. Sometimes, however, this comparison involved two demand characteristic comparisons (e.g., positive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

## 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 three moderators-of-interest, we estimated their values through a new set of participants. (See SI for construct validity analyses.) 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 demand and nil 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. We assumed that not communicating a hypothesis meant that participants (a) could not be motivated to confirm the hypothesis, and (b) would not have any pre-existing beliefs about the effect-of-interest activated. Thus, we assumed that motivation and belief scores in this condition were zero. We also assumed that opportunity ratings in the control condition were equivalent to the average ratings from non-control conditions.

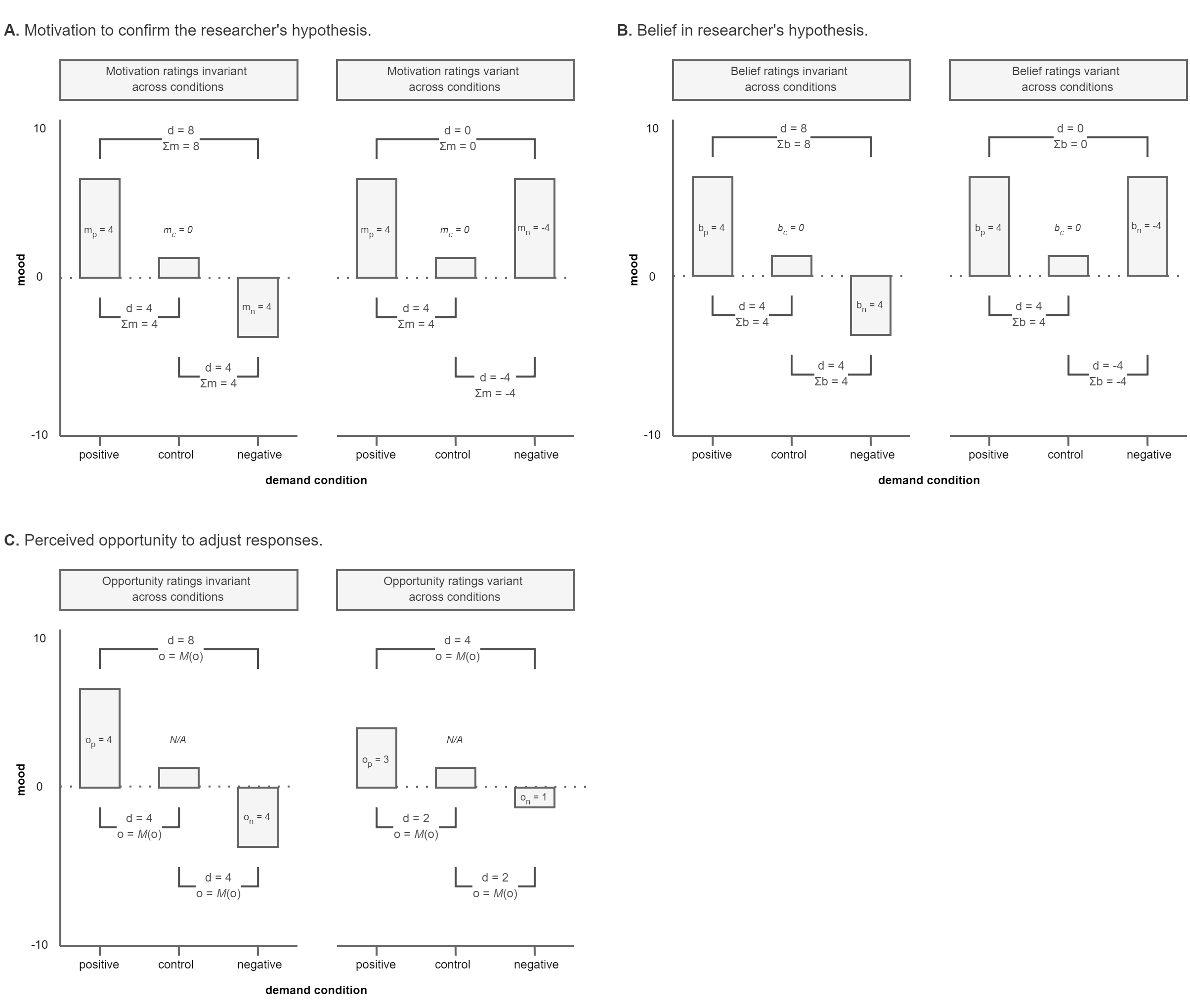


*Figure* *2.*  Vignettes for Coles et al. 2022, Study 1.

N undergraduates from Stanford University reviewed 20 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 7 = “extremely capable), and (3) believe the experimenter’s hypothesis (0 =”strong disbelief” to 7 = “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 scores were averaged across rater 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

Cohen’s represents the standardized difference between *two* groups. Thus, for each effect size, we summed the motivation and belief scores 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 a procedure is either described as mood-boosting (positive demand), described as mood-dampening (negative demand), or not described at all (control). If participants are equally motivated to confirm the hypothesis in both conditions, mood reports will be biased upward in the first condition and downward in the second condition (see Figure 3, Panel A, Column 1). In other words, the size of the demand effect is doubled because the motivational forces in the two conditions produce an additive effect. Similarly, these motivational forces could hypothetically cancel each other out. This would 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). We used a similar approach for belief scores (Figure 3, Panel B).

Theoretically, the opportunity to adjust responses should not differ between conditions. Thus, we averaged the opportunity ratings from the non-control conditions (see 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 (negative effect direction) *or* decreasing (positive effect direction) their mood report. Nonetheless, we discuss potential strategies in the Limitations sections for addressing this question in future primary research.

## Rater forecasts of demand effects

Even if researchers cannot necessarily explain how demand characteristics work, it might be valuable to be able to predict their effects (Yarkoni & Westfall, 2017). Orne suggested that participants themselves may be able make such forecasts (1969, 1970). 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”). These values were computed using the same approach as motivation and belief scores.

## Other moderators

We also coded several moderators for which we had no a-priori hypotheses. This included: (1) whether the study was published vs. unpublished, (2) whether the sample was student or non-student (e.g., MTurk), (3) whether participants were paid, (4) whether the study was conducted online, and (5) whether demand characteristics were manipulated within vs. between subjects.

## **Meta-analytic approach**

For estimating overall effects and moderators, we used random-effects meta-analysis with robust variance estimates (Hedges, Tipton, & Johnson, 2010). We included random-effects because Rosnow and Rosenthal (1997) and Coles et al. (2022) posited that moderators create a distribution of true effects—as opposed to a single fixed-effect (Hedges & Vevea, 1998). We used robust variance estimates because 85% of studies provided multiple effect sizes of interest, which violates the statistical assumption that effect sizes are independent. Like many meta-analysis procedures, meta-analysis with robust variance estimates uses an inverse-variance weighting scheme that adjusts for dependencies among effect sizes. For the present work, we used a heirarchical effects weighting scheme (Tanner-Smith & Tipton, 2014).

To estimate the overall effect size, we fit an intercept-only model. The intercept of this model can be interpreted as the precision-weighted overall effect size, adjusted for correlated-effect dependencies. We used the same approach to estimate subgroup overall effect sizes by subsetting the data by each level of each moderator. For moderator analyses, continuous and effect-coded categorical moderators were separately entered into the meta-regression model.

### **Publication bias analyses**.

The most common way to assess publication bias with dependent structures is to aggregate the dependent effect sizes and perform standard publication bias analyses. Following this approach, we used the MAd R package to aggregate dependent effect sizes (Del Re & Hoyt, 2010). We then used precision effect tests (i.e., PET-PEESE; Stanley & Doucouliagos, 2014) and weight-function modeling (Vevea & Hedges, 1995) to test for the presence of publication bias and estimate the bias-corrected overall effect size. As a further sensitivity analysis, we used an approach by Mathur and VanderWeele (2020) to estimate the maximal publication bias required to shift the observed overall effect size estimate to zero.

# Results

Results indicated that, overall, explicit manipulations of demand characteristics create a small acquiescence effect, = 0.28 , 95% CI [0.17, 0.39], = 5.16, = < .001.

### Moderator analyses.

Heterogeneity

# To-do

[] Start writing preliminary results

[] Have Mike review again?

[] Update pre-registration

[] Separate processing and analysis code

[] Clean up folder structure

[] Have M.W. add refs

[] Have M.W. build moderator table

# References