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

# To-do

[] Update and upload Figure 3

[] Look over Qualtrics to see if any details are missing

[] Start writing preliminary results

[] Have Mike review again?

[] Update pre-registration

[] Separate processing and analysis code

[] Have M.W. add refs

[] Have M.W. build moderator table

Imagine that one day a mysterious person approaches you and begins telling you about a new method for understanding humans: Colesology. The person tells you that Colesology is an extremely effective method for estimating causal relationships. However, they add that there’s a tiny issue: sometimes the method is 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 is not real; other times it causes them to miss an effect that is real. Sometimes it causes estimated relationships to be biased upward; other times it causes the estimates to be biased downward. And, in general, it means that the things researchers observe using this method 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 sometimes it does not.

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

## Debates about demand characteristics as a methodological artifact

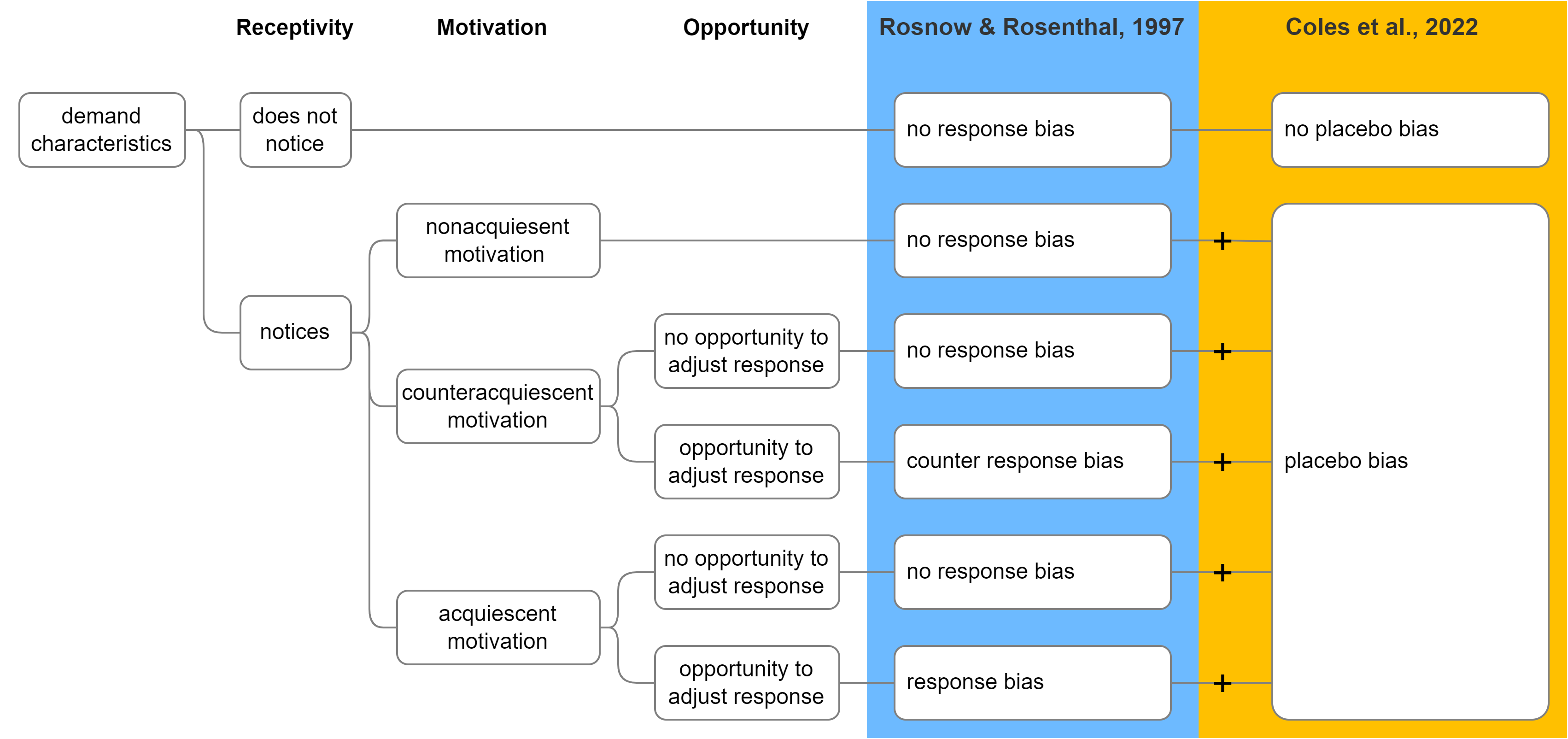
In 1962, Martin Orne published a seminal paper highlighting a view that challenged deeply-ingrained beliefs about the role of experiments in psychology. Rather than conceptualizing research participants as passive responders to the experimental context, Orne argued that participants actively navigate and try to make sense of the study context based on their assumptions, beliefs, and motivations. In particular, Orne emphasized that participants are perceptive to *demand characteristics:* “cues which convey an experimental hypothesis to the subject” (p. 779).

In the 60 years since Orne published his seminal paper, demand characteristics have become a literal textbook methodological concern in human subjects research (Sharpe and Whelton, 2016). But textbook psychology often does not capture the reality of the discipline. In actuality, disagreements about the extent to which demand characteristics impact participants’ response persist; Uncertainty about how such effects work remains. For example, Mummolo and Peterson (2021) consistently failed to find that manipulations of their communicated hypothesis impacted participants’ responses in large replications of classic studies in behavioral economics. Based on these findings, some researchers have argued that the effects of demand characteristics are “rare” (Clayton et al., 2020, p. 1092), supported by “little evidence” (Guess & Coppock, 2020, p. 1512), and/or “limited” (Vial, Brescoll, & Dovidio, 2019, p. 94). Yet, there are some contexts where these effects are more reliable. For example, in three experiments, Coles et al. (2022) found that the effects of posed facial expressions on emotional experience were moderated by demand characteristics. These facial feedback effects were larger when participants were told that the purpose of the study was to demonstrate that the effects are real and smaller when told the purpose was to demonstrate the effects are not real. Nonetheless, Coles et al. were left with the following conclusion: it is still unclear “how and when demand characteristics impact behavior” (p. 61).

In this paper, we provide an overview of a comprehensive framework for conceptualizing the impact of demand characteristics on participants’ responses (Rosnow & Rosenthal, 1997). We also review an alternative account that attempts to bridge the gap between research on demand characteristics and research on placebo effects (Coles et al., 2022). Next, we use meta-analysis to conduct the first quantitative synthesis of strict experimental tests of the effects of demand characteristics on participants’ responses. 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 required to transform frameworks about demand characteristics into formal theories—theories we believe might help distinguishing 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 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).

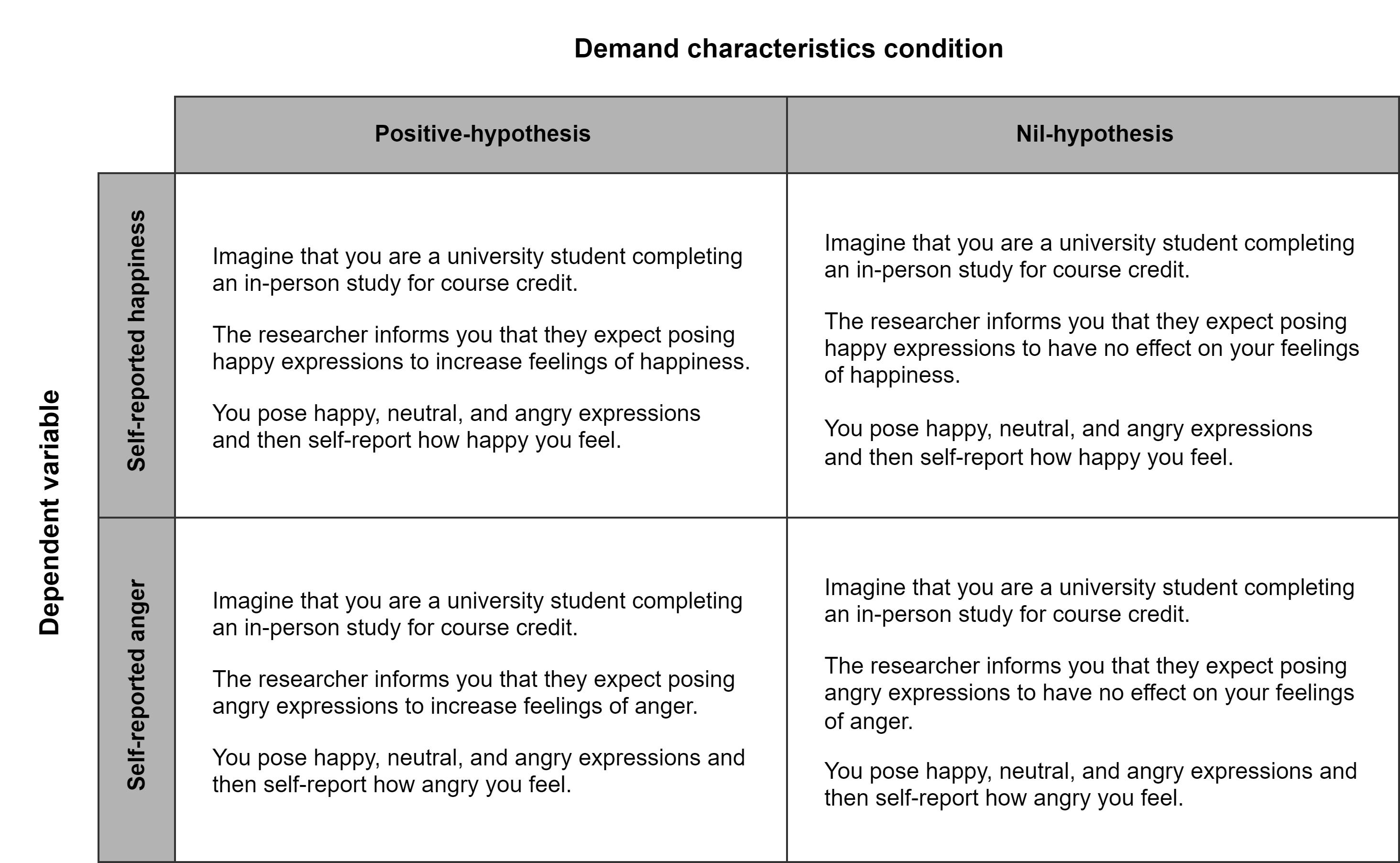
## Types of demand characteristic comparisons

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, Coles et al. (2022, Study 1) had two demand characteristics manipulations (positive demand and nil demand) and two dependent variables (self-reported happiness and anger). Thus, we created four vignettes (see Figure 2).

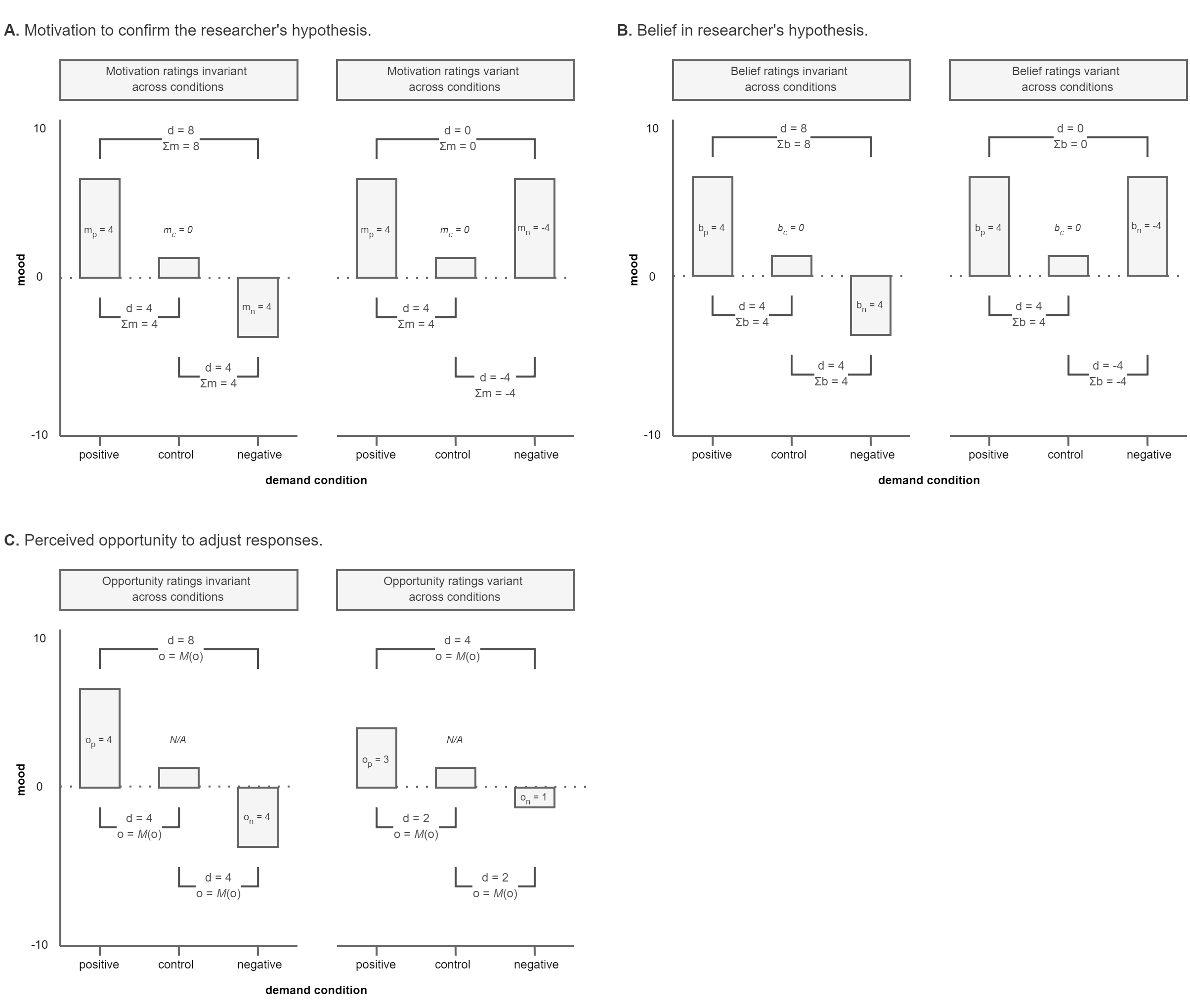


*Figure* *2.*  Vignettes for Coles et al. 2022, Study 1.

In total, there were X vignettes. N participants reviewed 20 randomly-selected vignettes. For each vignette, participants 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, participants rated the extent to which they would (1) be motivated to provide hypothesis-consistent responses, (2) be able to adjust their responses on the outcome-of-interest, and (3) believe the experimenter’s hypothesis. For each vignette, participants’ motivation, opportunity, and belief ratings were removed if they did not correctly identify the communicated hypothesis. The remaining scores were averaged.

For the control conditions, we assumed that motivation and belief ratings were 0 because participants were not given information about the experimenter’s hypothesis. We also assumed that opportunity ratings in the control condition were equivalent to the average ratings from non-control conditions.

### 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

For each effect size, we summed the motivation and belief scores for the two conditions being compared. Doing so allows 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-hypothesis), and (b) is extremely motivated to disconfirm the hypothesis. Relative to a control condition, this participant could act upon their motivation by either increasing (negative effect direction) *or* decreasing (positive effect direction) their mood. Nonetheless, we discuss potential strategies in the Limitations sections for addressing this question in future primary research.

# References