Meta-analysis suggests that the effects of demand characteristics can be consequential, unreliable, and difficult to explain

Nicholas A. Coles[1,2], Morgan Wyatt[3], & Michael C. Frank[3]

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

[2]Department of Psychology, University Florida

[3]Department of Psychology, Stanford University

Abstract

Demand characteristics are a fundamental methodological concern in experimental psychology. Yet, little is known about the direction, magnitude, consistency, and mechanisms underlying their effects. We conducted a three-level meta-analysis of 252 effect sizes from 52 studies that provided experimental tests of demand effects by explicitly manipulating cues about the study hypothesis. These manipulations tended to produce small overall increases in hypothesis-consistent responding (g = 0.21, 95% CI [0.12, 0.31]). However, effects were extremely heterogeneous (between-study τ = 0.28; within-study σ = 0.18), with the prediction interval ranging from g = 0.89 (a large increase in hypothesis-consistent responding) to g = -0.46 (a moderate *decrease* in hypothesis-consistent responding). Both the observed and estimated distribution of these effects suggested that demand characteristics can create false positives, false negatives, upward bias, and downward bias. This heterogeneity is currently difficult to explain. New participants who reviewed key study details were neither able to predict nor provide insights into psychological mechanisms theorized to underlie demand effects. Participants' ratings of three theorized moderators – motivation to adjust responses, opportunity to adjust responses, and belief in the researcher's hypothesis – failed to predict observed demand effects. Coded methodological features (e.g., whether participants were paid) also often failed to predict observed effects – explaining approximately 15% of in-sample variability. Although the meta-analysis did not capture the full depth of the demand characteristics construct, the synthesis of even a narrow subset of the literature suggests that their effects can be inferentially consequential, unreliable, and difficult to explain.

Keywords: demand characteristics, expectancies, meta-analysis, methodology, confound

Meta-analysis suggests that the effects of demand characteristics can be consequential, unreliable, and difficult to explain

"All scientific inquiry is subject to error, and it is far better to be aware of this, to study the sources in an attempt to reduce it, and to estimate the magnitude of such errors in our findings, than to be ignorant of the errors concealed in the data" (Hyman, 1954, p. 4)

Imagine that one day a mysterious person approaches you and begins telling you about a new method they invented for understanding humans. They tell you that their method is useful for estimating causal relationships, but add that there is one issue: it can sometimes be thrown off by a *methodological artifact*. They explain that this artifact sometimes causes researchers to detect an effect that's not real, and other times causes them to miss an effect that is real; that it sometimes biases estimates upward and other times downward. Then, they offer a confession: the artifact doesn't always impact their conclusions, and they don't know why. After over a half century of inquiry, they explain, the artifact's effects ultimately remain difficult to predict and explain.

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

In a seminal paper published over a half century ago, Martin Orne argued that human subjects are perceptive to demand characteristics – "cues which convey an experimental hypothesis" – and generally use these cues to help the experimenter confirm their hypothesis (1962, p. 779). Orne initially presented evidence that demand characteristics can lead to false

positives, such as patients exhibiting sham symptoms of hypnosis (Orne, 1959). However, demand characteristics can also lead to false negatives. For example, participants will ignore visual cues of depth when they believe that disregarding them is the purpose of the experiment (Hayes & King, 1967). In addition to creating inferential errors, demand characteristics can bias estimates of causal relationships. For example, the effects of facial poses on self-reported emotion can be amplified *or* attenuated depending on whether the experimenter communicates expectations of positive or nil effects (Coles, Gaertner, Frohlich, Larsen, & Basnight-Brown, 2022). Puzzlingly, though, demand characteristics do not always seem to matter. For example, in a set of large replications of classic studies in behavioral economics, direct manipulations of demand characteristics consistently failed to significantly impact participants' responses (Mummolo & Peterson, 2019).

In the present work, we advance an unexpected[1] thesis based on the above observations and our own follow-up meta-analytic work: demand characteristics are uncomfortably close to the mysterious methodological artifact described in the opening of the paper. Demand effects are a literal textbook methodological concern in experimental psychology (Sharpe & Whelton, 2016). They can create false positives, false negatives, upward bias, and downward bias. Yet, over 50+ years after Orne influentially described them (1962), demand characteristics remain difficult to predict and explain. To begin, we review an influential theoretical framework that initially guided our investigation.

[1] The thesis advanced in the present work is unexpected in the sense that we initially pre-registered an investigation we thought might reveal insights into the nature of demand effects. Instead, several years later, we are left with more questions than we began with.

How do demand characteristics alter participant responses?

One of the most influential frameworks for conceptualizing demand effects was developed by Rosnow and colleagues (Rosnow & Aiken, 1973; Rosnow & Rosenthal, 1997; Strohmetz, 2008). In this framework, they described three key moderators we discuss in the present work: (1) receptivity to cues, (2) motivation to provide hypothesis-consistent responses, and (3) opportunity to alter responses.

To start, Rosnow and colleagues reasoned that participants must be receptive to demand characteristics for there to be subsequent shifts in participants' responses (see also, Orne, 1958). As an extreme example, imagine that a researcher hands an infant a sheet of paper that precisely explains the study hypothesis. Demand characteristics are certainly present, but they are not predicted to have an impact because the infant is not receptive to the cues. Even if the infant possessed the astonishing ability to read, it's possible they would misunderstand the cues (Corneille & Lush, 2023) – which we will consider another form of non-receptivity in the present work.

If participants correctly interpret demand characteristics, Rosnow and colleagues theorized that subsequent changes in participants' responses would be driven by their motivation (or lack thereof) to provide hypothesis-consistent responses. Early work on demand characteristics was marked by debates about whether participants are motivated to adjust their responses to (a) help the researcher confirm their hypothesis (Orne, 1962), (b) receive positive evaluations (Riecken, 1962; Rosenberg, 1969; Sigall, Aronson, & Van Hoose, 1970), (c) interfere with the purpose of the study (Cook et al., 1970; Masling, 1966), or (d) follow directions as closely as possible (Fillenbaun & Frey, 1970). Rosnow and colleagues advanced this line of thinking by illustrating that participants have *multiple* shifting motivations in mind when they conceptualize their roles as

subjects (Rosnow & Rosenthal, 1997; see also Silverman & Marcantonio, 1965). For example, participants appear to be motivated to increase performance on simple tasks when told that this is the experimenter's expectation – but not when the experimenter adds that the increase in performance will be indicative of a negative personality trait (Sigall et al., 1970). Rosnow and colleagues, thus, suggested that participants in any given context can be characterized as being overall motivated to either: (a) non-acquiesce (i.e., not change their responses based on knowledge about the hypothesis), (b) acquiesce (i.e., provide hypothesis-consistent responses), or (c) counter-acquiesce (i.e., provide hypothesis-inconsistent responses).

If participants are motivated to adjust their response, Rosnow and colleagues theorized that subsequent changes in participants' responses would then be driven by their ability to alter the outcome of interest. As elaborated by Corneille and Lush (2023), this could occur through faking, imagination, or phenomenological control (voluntary changes experienced by the participant as involuntary). Taking this third moderator – opportunity – into account, Rosnow and colleagues concluded that demand characteristics bias responses when participants (1) notice the cues, (2) are motivated to adjust their responses, and (3) can adjust their responses. This framework directly maps onto many psychologists' typical playbook for avoiding the impact of demand characteristics: use deception and/or unobtrusive procedures (reduce receptivity), incentivize honest reporting (reduce motivation), and/or deploy difficult-to-control outcome measures (reduce opportunity to adjust responses).

Of course, other researchers have since expanded upon and/or challenged parts of Rosnow and colleagues' framework. For example, by elaborating upon underlying mechanisms like imagination, Corneille and Lush (2023) more clearly highlight that participants can willingly change many outcomes that may initially seem outside their control. For example, a participant

who wants to help a researcher confirm that a manuscript reviewing research artifacts is physiologically arousing could likely do so by simply imagining a physiologically arousing context. Relatedly, Coles et al. (2022) argued that demand characteristics may sometimes impact participants in cases where they are *not* motivated to adjust responses – e.g., via conditioned responses or other mechanisms discussed in conceptually-related work on placebo effects (Stewart-Williams & Podd, 2004). We focus our review on Rosnow and colleagues' influential framework, but we revisit complementary ideas throughout.

Methodology

The goal of the current paper is to quantitatively take stock of what experimental psychologists have learned – if anything – about demand effects in the 50+ years since Orne influentially described them (Orne, 1962). Although several excellent *narrative reviews* exist (Corneille & Lush, 2023; Rosnow & Rosenthal, 1997; Sharpe & Whelton, 2016; Strohmetz, 2008), meta-analysis allows us to evaluate the magnitude, consistency, and potential moderators of demand effects.

We defined the scope of the meta-analysis using the Population, Intervention, Comparison, Outcome framework (Schardt, Adams, Owens, Keitz, & Fontelo, 2007). Our population-of-interest was human subjects participating in non-clinical psychology experiments. Given that there is a sizable literature and number of reviews on conceptually-related placebo effects, excluding clinical studies improved the feasibility and reduced the redundancy of our work.

The intervention-of-interest was explicit manipulations of the hypothesis communicated to participants – i.e., scenarios where a researcher tells participants about the effect of an

independent variable on a dependent variable. Demand characteristics are sometimes defined as *any* cue that may impact participants' beliefs about the purpose of the study, including instructions, rumors, and experimenter behavior (Orne, 1962). However, such a definition creates a potentially boundless conceptual space where *any* systematic change in a research design might be considered a test of demand characteristics. Furthermore, our own review of the literature revealed that researchers tend to study such effects by manipulating the study hypothesis *explicitly* communicated to participants. For these reasons, the present review focuses on relatively explicit manipulations of the study hypothesis. (See *Limitations* section for further reflection.)

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

Literature search. Figure 1 provides a PRISMA-style flowchart summarizing our literature search and screening process (Page et al., 2021).

The literature search was initially developed in consultation with a librarian at (anonymous for peer review) and later expanded based on reviewer feedback. On January 12, 2022, we searched APA PsycInfo using broad search terms: "demand characteristics" OR "hypothesis awareness" (n = 850 records identified). On April 17, 2024, we repeated the search to identify records published after the initial search (n = 29 records identified). At that time, we also expanded the search to include conceptually similar terms found in the appendix of Rosnow and Rosenthal's (1997) book on experimental artifacts: "participant role" OR "demand effects"

OR “good subject effect” OR “expectancy effect” OR “evaluative apprehension” (n = 572 records identified). We also released a call for unpublished studies on the Society for Personality and Social Psychology Open Forum, Twitter, the Facebook Psychological Methods Discussion group, and the Facebook PsychMAP group (n = 6 records identified).

Our search did not have language restrictions and went as far back as 1840, which yielded 1457 records, 168 of which were unpublished.

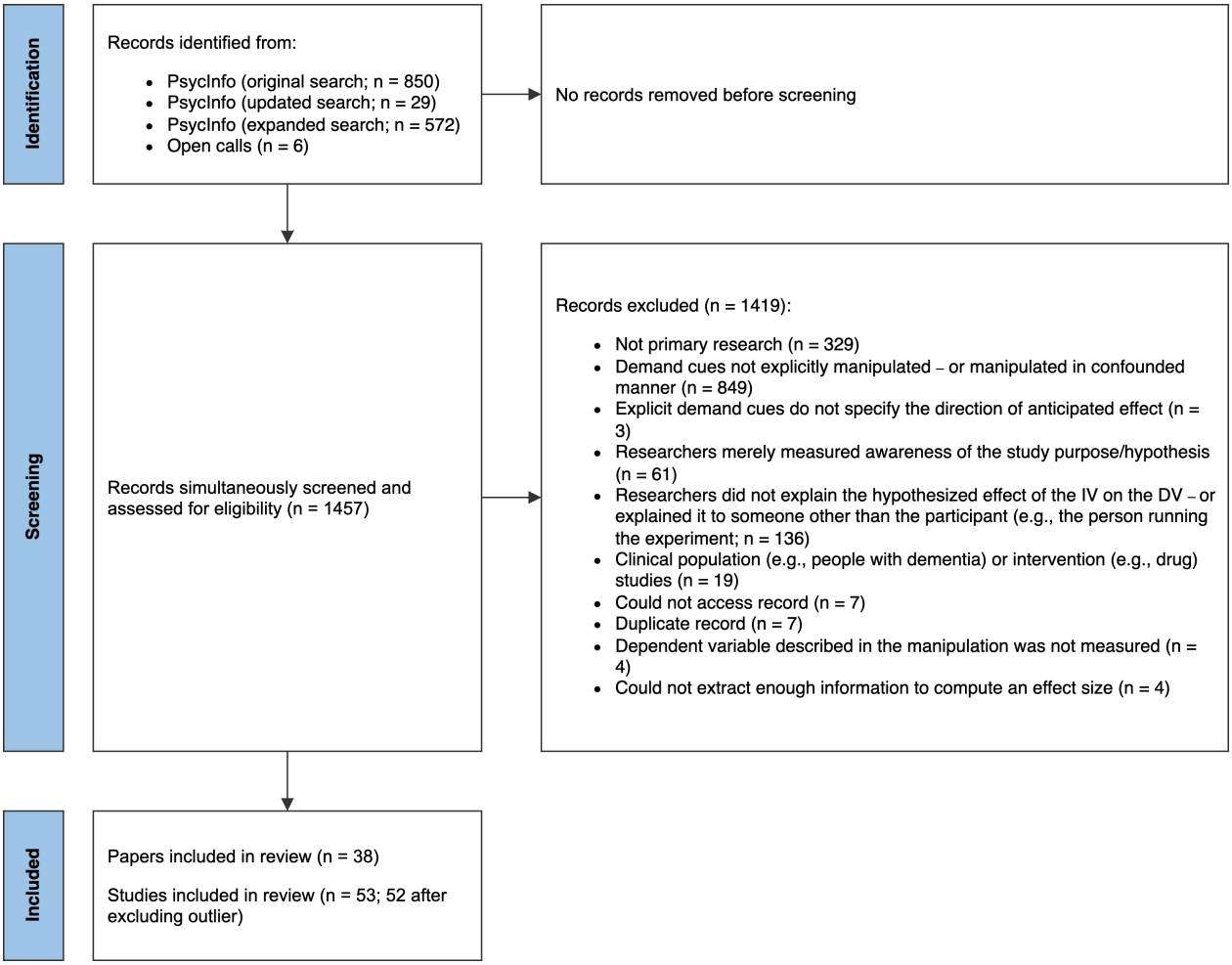

Figure 1. PRISMA-style flowchart illustrating the identification, screening, and selection of studies.

Screening. Put simply, records must have met the following criteria in order to be eligible for inclusion in the meta-analysis:

- The researcher manipulated what participants were told about the effect of an independent variable on a dependent variable.[2] In most cases, the effect of the independent variable was described explicitly, but there were some included studies where it was strongly implied.
- The demand characteristics manipulation was not strongly confounded with another manipulation. For example, we excluded a study by Sigall et al. (1970) because the manipulation of the stated hypothesis was confounded with a disclosure about the meaning of the behavior (i.e., that confirming the hypothesis would be indicative of an obsessive-compulsive personality disorder).
- A non-clinical population was studied.
- Information necessary for computing at least one effect size was included.

Figure 1 more thoroughly summarizes exclusion criteria. In instances where multiple exclusion criteria applied, coders were asked to choose only one option.

[2] We excluded conditions where the researcher communicated a *non-directional* effect. We did so because participants in these scenarios could not unambiguously infer how their responses were expected to change. For example, if participants were told that an independent variable would "impact mood", it is not clear if participants should infer that the mood will be boosted or dampened.

N. C. and M. W. screened records independently, reviewed potentially relevant records together, and worked together to code the information for moderator analyses and effect size computations. Any disagreements were resolved through discussion. Abstracts and (if necessary) full texts were reviewed in a single step so that records did not have to be reviewed twice during screening. In total, 53 studies from 38 records were eligible for inclusion. However, one record (Allen & Smith, 2012) was removed because the information provided led to implausibly large effect size estimates (e.g., d = -209.16).

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

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

In some scenarios, we estimated the *interactive* effect of explicit demand characteristics. For example, in the same Coles et al. (2022) study, participants provided happiness ratings both after smiling and scowling. Participants' mood generally improved when smiling vs. scowling (i.e., there was a main effect of facial pose). However, the difference was more pronounced when participants were told about the mood-boosting effects of smiling. In other words, there was an interaction between facial pose and explicit demand characteristics. In this scenario, the interactive effect of explicit demand characteristics was computed by calculating a standardized difference-in-differences score.

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

We calculated Hedge's g by applying a small sample correction to Cohen's d_s (for between-subject designs) and d_{rm} (for within-subject designs[3]) estimates. Whenever possible, we used the *M*'s and *SD*'s reported in a paper to compute Cohen's *d*. If these values were not reported, we used (in order of preference), (1) *t*-values, (2) descriptive statistics extracted from figures (e.g., bar charts) using the WebPlotDigitizer (Drevon, Fursa, & Malcolm, 2017), (3) *F*-values, or (4) *p*-values. In instances where relevant information was not provided but the significance and direction of the effect was described, we assumed *p*-values of .04 and .50 for significant and non-significant effects respectively (e.g., Kenealy, 1988). In a few instances, an outcome variable in a study was discrete, as opposed to continuous (e.g., Orne & Scheibe, 1964). In these cases, we approximated a Cohen's *d* score based on a transformation of the log odds ratio (Borenstein, Hedges, Higgins, & Rothstein, 2011).

[3] 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 a recommendation by Borenstein (2009) and performed sensitivity analyses on an assumed correlation. We preregistered a default correlation of r = .50 but performed sensitivity analysis with r = .10, .30, .50, .70, and .90. These sensitivity analyses produced virtually no change in overall effect size estimates, so we do not discuss them further.

Nearly all studies (75%) contained multiple effect sizes of interest. For example, the full design in Coles et al. (2022) included a positive demand, nil demand, and control condition. Participants also completed several facial expression poses (happy, angry, and neutral) and self-reported several emotions (happiness and anger). To be comprehensive, we recorded all reported effect sizes and accounted for dependencies using three-level meta-analysis (described later).

Potential study feature moderators. The studies we included in our meta-analysis were methodologically varied (for more information, see *Results* and *Limitations*). Below, we describe study features we coded as potential moderators of demand effects:

- *Group comparison.* Most studies included in our meta-analysis examined the effects of *positive demand*, wherein participants were told that the dependent variable will increase. However, a notable subset of studies examined the impact of *negative demand* (participants told that the dependent variable will decrease) or *nil demand* (participants told the dependent variable will be unaffected). Often these conditions were compared to a *control* condition, wherein participants were not told about an effect of an independent variable on a dependent variable. Sometimes, though, one demand condition was compared to another.

- *Control vs. non-control group comparison.* Demand effects should presumably be additive. For example, imagine a study where the effect of a task is either (a) not described at all (a control condition), (b) described as mood-boosting (positive demand) or (c) described as mood-dampening (negative demand). Further imagine that participants are motivated and able to adjust their responses. Compared to the control condition, participants' moods are predicted to be boosted in the positive demand condition and dampened in the negative demand condition. If this is the case, the mean difference in

mood should be larger when the positive demand condition is compared to the negative demand condition (as opposed to the control condition). To test this, we coded whether comparisons were made to a control group or a different demand condition.

- *Control group comparison.* Instances where a demand characteristic condition was compared to a control group also allowed us to test whether participants' responses shift more when the researcher hypothesizes an increase (positive demand), a decrease (negative demand), or no change in the dependent variable (nil demand).

- *Design of demand characteristics manipulation.* Whether demand characteristics were manipulation within- vs. between-subjects.

- *Participant pool.* Whether students, non-students (e.g., MTurk workers), or a mix of students and non-students were sampled.

- *Setting.* Whether the study was conducted online or in-person.

- *Payment.* Whether participants were paid or unpaid.

- *Publication status.* Whether the study was published or unpublished.

Can research participants help us understand demand effects? During our literature review, we found very few papers that tested mechanisms that may help predict demand effects. We thus turned to a population that Orne (1969) believed may help researchers understand demand effects: research participants themselves. As recently reviewed by Corneille and Béna (2023), participants can successfully predict a variety of effects in experimental psychology, including the approach-avoidance effect, mere exposure effect, and the rubber hand illusion. When this occurs, it raises concerns that the original effect may have been driven by

demand characteristics (Bartels, 2019). Here, we attempt to extend this methodology not to raise concerns about participants' potential responses to demand characteristics – but instead to evaluate whether they can explain *when* and *how* such effects operate.

As we describe below, we asked a new set of participants to review vignettes describing key details of studies included in the meta-analysis. We then solicited judgments of not only whether they believed demand effects would emerge, but also the extent to which they (a) correctly identified the communicated hypothesis, (b) would be motivated to adjust responses, (c) would be able to adjust responses, and (d) would believe the experimenter's hypothesis.

Vignette rating methodology. For each study included in the meta-analysis after our original literature search[4], we created vignettes that described the key details for each demand characteristic condition and dependent variable combination. For example, Standing, Verpaelst, and Ulmer (2008) had two demand characteristic manipulations (positive and negative demand) and two dependent variables (measures of verbal and spatial reasoning). Thus, we created four vignettes for this study (Figure 2). In an effort to help participants understand the study context, vignettes also contained information about (a) whether students vs. non-students were sampled, (b) whether subjects received compensation, and (c) whether the study was conducted online or in-person.

[4] As a reminder, we performed two literature searches. The second literature search was inspired by reviewer feedback, which we received after we started collecting data using the vignette methodology.

Dependent variable	Demand characteristics condition: Positive demand	Demand characteristics condition: Negative demand
Spatial reasoning	Imagine that you are a university student completing an in-person study as a volunteer or for course credit. The researcher informs you that they are interested in the **beneficial effects** of listening to Mozart on test-taking capabilities. While Mozart music is played, you are then asked to complete a ***test measuring spatial reasoning***, wherein you see unfolding shapes and guess the corresponding folded pattern.	Imagine that you are a university student completing an in-person study as a volunteer or for course credit. The researcher informs you that they are interested in the **deleterious effects** of listening to Mozart on test-taking capabilities. While Mozart music is played, you are then asked to complete a ***test measuring spatial reasoning***, wherein you see unfolding shapes and guess the corresponding folded pattern.
Verbal reasoning	Imagine that you are a university student completing an in-person study as a volunteer or for course credit. The researcher informs you that they are interested in the **beneficial effects** of listening to Mozart on test-taking capabilities. While Mozart music is played, you are then asked to complete a ***test measuring verbal reasoning***, wherein you are asked to fill in the first and last word of example sentences.	Imagine that you are a university student completing an in-person study as a volunteer or for course credit. The researcher informs you that they are interested in the **deleterious effects** of listening to Mozart on test-taking capabilities. While Mozart music is played, you are then asked to complete a ***test measuring verbal reasoning***, wherein you are asked to fill in the first and last word of example sentences.

Figure 2. Vignettes for Standing et al. (2008), which described the key details for each demand characteristic condition (bolded and underlined) and dependent variable (bolded and italicized) combination.

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 (i.e., there were no explicit demand characteristics to act upon).

Using a web-based Qualtrics survey, participants reviewed 10 randomly selected vignettes. Much like the sample in the studies they reviewed, these participants were a convenience sample. For each study, participants were asked to first identify the researcher's hypothesis. Here, participants chose between four options that described a filler effect (usually involving an irrelevant dependent variable) or a positive, negative, or nil effect of the independent variable on the dependent variable. Although not originally pre-registered, the proportion of participants who correctly identified the hypothesis in each vignette (0 - 1) will later be used to evaluate Rosnow and colleagues' proposed receptivity moderator.

Afterwards, participants rated the extent to which they would hypothetically (a) be motivated to adjust responses based on the researcher's stated hypothesis (-3 = "extremely motivated to adjust responses to be inconsistent" to 3 = "extremely motivated to adjust responses to be consistent"), (b) be able to adjust their responses on the outcome-of-interest (0 = "extremely incapable" to 4 = "extremely capable"), and (c) believe the hypothesized effect would occur (-3 = "strong disbelief" to 3 = "strong belief"). Participants also indicated the extent to which they expected other participants to adjust their responses to confirm the hypothesized effect (-3 = "extremely likely to adjust responses to be *inconsistent*" to 3 "extremely likely to adjust responses to be consistent"). These rating scales were presented in random order.

Sample size was initially based on availability of resources.[5] We originally collected as much data as possible (n = 192) in a single quarter from undergraduates from (anonymous for peer review). Following a reviewer recommendation, we performed post-hoc examinations of the reliability of their ratings. More specifically, we calculated intraclass correlations using mixed effects models. For ratings of predicted demand effects, motivation to adjust responses, opportunity to adjust responses, and belief in the hypothesized effect, we used the lme4 package (Bates, Mächler, Bolker, & Walker, 2015) in R (R Core Team, 2021) to fit an intercept-only mixed effect model with random intercepts at the level of participant and vignette. We then used the performance package (Lüdecke, Ben-Shachar, Patil, Waggoner, & Makowski, 2021) to calculate the intraclass correlation for the participant random intercept. The intraclass coefficient for predicted demand effects (ICC = 0.21), motivation to adjust responses (ICC = 0.23), opportunity to adjust responses (ICC = 0.21), and belief in the researcher's stated hypothesis (ICC = 0.14) was low.

The low intraclass correlations from our original sample indicates that participants strongly disagree about how they will respond to explicit demand cues. Nonetheless, the Law of Large Numbers stipulates that these relatively imprecise ratings should converge into relatively precise estimates of the true mean at larger samples. We attempted to exploit this statistical tendency by collecting additional ratings from Prolific workers. This left us with a total of 412 participants (55% female; 41% male; all other participants indicated they were transgender,

[5] For transparency, we would like to note that earlier analyses with our initial sample of participants suggested that observed demand effects were moderated by ratings of the extent to which they believed the researcher's stated hypothesis. This finding, however, did not replicate in our full sample.

gender non-conforming, some other gender, or unwilling to disclose gender). 54% of participants reported they were White/Caucasian, 20% Asian, 11% Black/African American. All other participants declined to respond or indicated their ethnicity could not be described by a single (or any) provided category. The average participant age was 30.10 (SD = 13.82).

Accounting for different demand comparisons. As mentioned before, Hedge's g represents the standardized difference between *two* groups. Thus, for each observation in the meta-analysis, we summed participants' average motivation, opportunity, and belief ratings (after removing cases where they identified the wrong hypothesis). We also summed the estimates of how likely participants were to correctly identify the communicated hypothesis. Doing so allowed us to accommodate the fact that some comparisons involved two demand characteristics conditions. For example, imagine a study where participants are told a procedure will boost mood (positive demand), told a procedure will dampen mood (negative demand), or not told about an expected effect (control). Compared to a control condition, participants who are motivated to confirm the hypothesis are theorized to have upward-biased responses in the positive demand condition and downward-biased responses in the negative demand condition. If those demand conditions are compared to each other – instead of a control condition – their effects should be additive. Summing participants ratings allowed us to accommodate this possibility.

We did not include nil-hypothesis comparisons in our analyses because our coding strategy could not accommodate the potential moderating role of motivation and belief in these conditions. For example, imagine that a participant is (a) told that an intervention will not impact mood (nil demand), and (b) is motivated to disconfirm the hypothesis. Relative to a control condition, this participant could disconfirm the hypothesis by either increasing *or* decreasing their

mood report. Thus, even if motivation does moderate the effects of demand characteristics, we would not expect a systematic pattern to emerge with our coding scheme.

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

We fit all models using the metafor package (Viechtbauer, 2010) in R (R Core Team, 2021). We weighed effect sizes based on their inverse-variance and used cluster-robust methods for estimating variance-covariance matrices (Pustejovsky & Tipton, 2018). To estimate the overall effect size, we fit an intercept-only 3LMA model. We conducted moderator analyses by separately entering variables into a new model. In doing so, we hoped to avoid issues with collinearity and overfitting. Categorical moderators were dummy coded. To test the significance of each moderator, we used model comparison *F*-tests. To estimate effect sizes within each subgroup of the moderator, we used model-derived estimates.

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

First, we visually examined *funnel plots,* wherein observed effect sizes are plotted against a measure of their precision (e.g., standard error). In the absence of publication bias, the distribution typically resembles a funnel; relatively large studies estimate the effect with high precision, and effect sizes fan out in *both* directions as the studies become smaller. If, however, non-significant findings are disproportionately omitted from the scientific record (i.e., there is publication bias), the distribution is often asymmetric/sloped. Funnel plots traditionally contain one effect size per study, but many of our studies included multiple relevant effect sizes. Thus, we examined two funnel plots: one with all effect sizes and one with the dependent effect sizes aggregated[6].

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

[6] For effect size aggregation, we assumed a default dependent effect size correlation of r = .50 but performed sensitivity analysis with r = .10, .30, .50, .70, and .90. These sensitivity analyses did not change our overall conclusion about publication bias, so we do not discuss them further.

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

As a sensitivity analysis, we used the PublicationBias package in R (Mathur & VanderWeele, 2020a) to estimate the ratio in which publication bias would have to favor affirmative studies in order make the overall effect size in a robust random effects model non-significant (Mathur & VanderWeele, 2020b). We also estimated the difference in the magnitude of published vs. unpublished effects in a moderator analysis.

Transparency and openness. The project pre-registration, materials, data, and code are openly available at https://osf.io/3hkre/?view_only=2dc92af53f194e5eab0d7aecafaf01c2. This link also contains a list of amendments/deviations we made to our pre-registration as the project evolved and reviewer feedback was received. These amendments were largely concerned with (a) whether and how many vignette ratings to collect, (b) whether to make vignette-related tests confirmatory vs. exploratory, (c) whether new primary data collected for a Coles et al. (2024) record is described in the main text vs. summarized in the meta-analysis, (d) whether to code the quality of included records[7], (e) expansions to the literature search, (f) whether to use Cohen's *d* or Hedge's

[7] N.C. coded the quality of each record included in the meta-analysis using a modified version of the Downs and Black (1998) checklist. On a 0-1 scale, ratings of reporting quality were modest

g as our effect size index, and (g) decisions regarding whether to use robust variance estimation, mixed-effects models, or both.

For the meta-analysis, sample size was determined by the availability of relevant records. For the vignette ratings, sample size was initially determined by the availability of resources (i.e., we collected as much data as possible in a single quarter). However, our second wave of participant recruitment considered precision – and was designed to (a) yield an equal number of participant ratings per vignette, and (b) decrease the length of the 95% confidence intervals of the predicted demand effect, motivation, opportunity, and belief ratings to 1.

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

Results

In total, we extracted 252 effect sizes from 52 studies from between the years 1964 and 2024 (M = 2003, SD = 18.63). 11 of these studies were unpublished.

In order of frequency, effect sizes represented a positive demand compared to a control group (k = 114), positive demand to negative demand (k = 44), negative demand to a control

---

(M = 0.71, SD = 0.30); ratings of internal validity were high (M = 0.91, SD = 0.17); and ratings of external validity were consistently 0. The low external validity scores were driven by the reliance on non-representative sampling methods, an unfortunately common limitation in experimental psychology (Frank et al., 2023). Quality ratings were not significantly associated with observed effect sizes and are thus not discussed.

group (k = 43), positive demand to a nil demand group (k = 34), or nil demand to a control group (k = 17).

More broadly, effect sizes tended to compare one demand condition to a control group (k = 174) – as opposed to a group exposed to a different type of explicit demand cues (k = 78). Regardless of what type of demand manipulation was used, it was more common to manipulate the cues between (k = 208) vs. within subjects (k = 44).

Most effect sizes came from student samples (k = 160), although some samples were non-students (k = 25), a mix of students and non-students (k = 19), or not described thoroughly enough to make a determination (k = 48). Most effect sizes came from unpaid samples (k = 162), although some were paid (k = 50) and some were not described thoroughly enough to make a determination (k = 40). The majority of effect sizes came from in-person studies (k = 187), but some were from online studies (k = 52) or not described thoroughly enough to make a determination (k = 13).

Overall results. Overall, results indicated that explicit manipulations of demand characteristics cause participants' responses to shift in a manner consistent with the communicated hypothesis, g = 0.21, 95% CI [0.12, 0.31], $t(47.30) = 4.53$, $p < .001$. As a hypothetical example, if participants were told that the researcher hypothesizes that an intervention will improve mood (positive demand), they would generally report slightly improved moods; if told that the researcher hypothesizes that an intervention will worsen mood (negative demand), they would generally report slightly worsened moods.

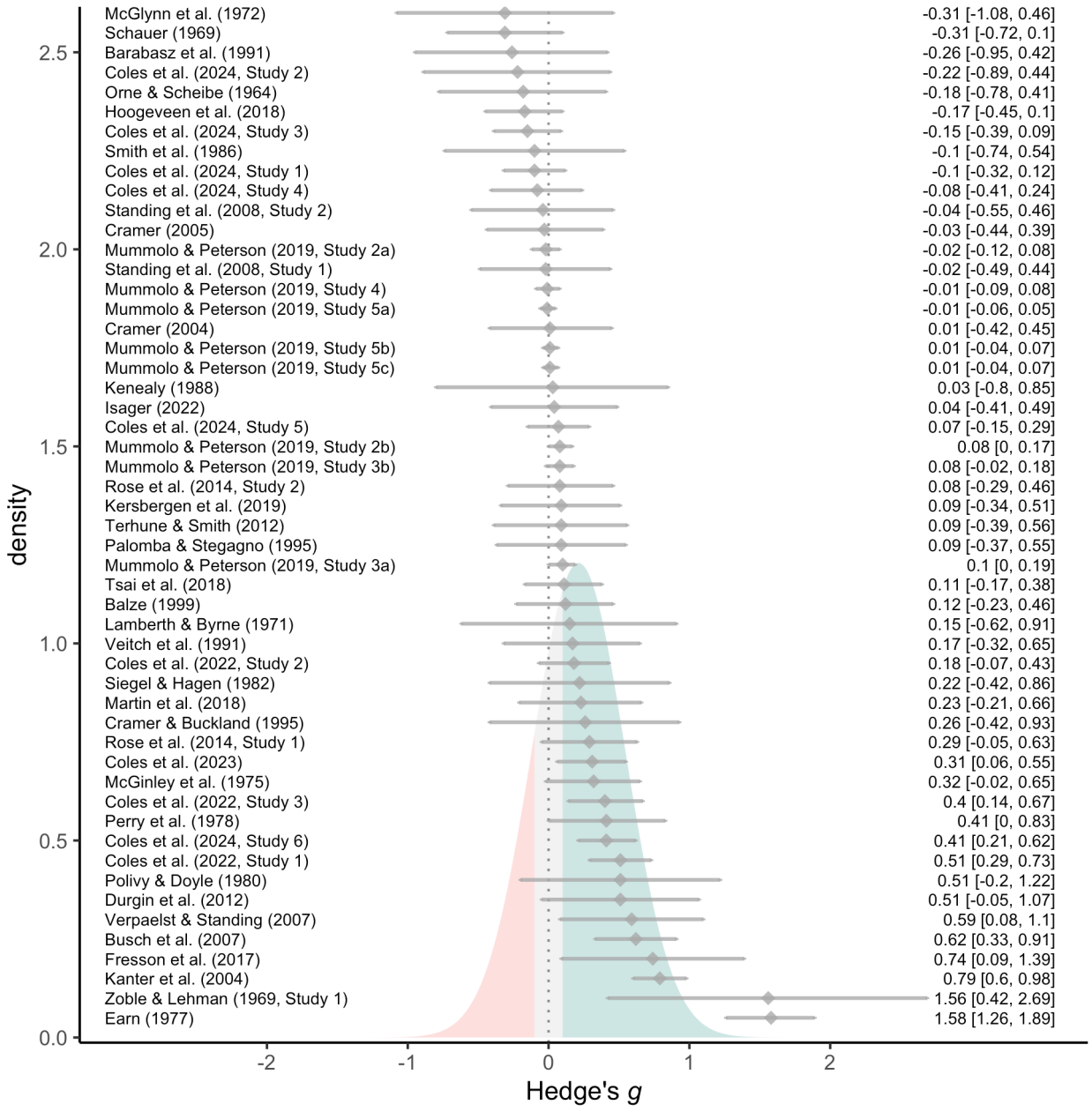

Figure 3. Forest plot of effect sizes (grey diamonds), their 95% confidence intervals (grey error bars), and their citations (left). For visualization purposes, effect sizes are aggregated within-studies (see openly-available data for non-aggregated effect sizes). The estimated effect size distribution is also shown and colored based on whether demand characteristics produce more hypothesis-consistent responding (green; $g > 0.10$), more hypothesis-inconsistent responding (red; $g < -0.10$), or negligible shifts in responding (grey; $|g| < 0.10$).

As a reminder, rather than assuming that there is a *single true effect* of demand characteristics, 3LMA assumes a distribution of *multiple true effects.* Consistent with this assumption, observed variability in demand effects drastically exceeded what would be expected from sampling error alone (between-study $\tau = 0.28$; within-study $\sigma = 0.18$; $Q(251) = 972.42$, $p < .001$). 3LMA often assumes that the multiple true effects form a normal distribution, which we recreated based on estimates of the average effect size and variability attributed to sources other than sampling error (between-study τ + within-study σ). As shown in Figure 3, this estimated distribution illustrates that demand characteristics can have a wide range of effects. Indeed, the 95% prediction interval for a single future study ranges from $g = -0.46$ to $g = -0.46$.

As a heuristic, we arbitrarily classified any effect size less than 0.10 standard deviation in either direction (i.e., $|g| < .10$) as "negligible". Based on this classification, the estimated distribution of effects suggested that demand characteristics most often produce hypothesis-consistent shifts (63%), but sometimes produce negligible shifts (20%) or shifts in the *opposite* direction of the communicated hypothesis (17%). Such results are consistent with Rosnow and colleagues' prediction that demand characteristics can lead to both hypothesis-consistent and hypothesis-*inconsistent* shifts in participants' responses.

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

Study features. In general, we did not find much evidence that demand effects are moderated by study features (see Table 1). The two exceptions were (1) whether the demand characteristics condition was compared to a control group (vs. another condition with demand characteristics), and (2) whether the study was conducted in-person (vs. online).

The average effect sizes was estimated to be twice as large when two demand characteristic conditions were compared (d = 0.34, 95% CI [0.19, 0.49], p < .001), as opposed to one demand characteristic condition being compared to a control group (d = 0.16, 95% CI [0.08, 0.25], p = < .001), $F(1, 10.41) = 10.55$, p = .008. This provides preliminary evidence that the effects of demand characteristics are additive. However, these results should be interpreted with some caution, as a broader test of whether *all* specific types of comparisons varied was not statistically significant, $F(4, 3.55) = 1.89$, p = .292.

Instances where a demand characteristic condition was compared to a control group allowed us to test whether participants responses shift more when the researcher hypothesizes an increase (i.e., positive demand; d = 0.18, 95% CI [0.07, 0.29], p = .002), a decrease (i.e., negative demand; d = 0.20, 95% CI [0.07, 0.33], p = .005), or no change in the dependent variable (i.e., nil demand; d = 0.27, 95% CI [-0.20, 0.75], p = .169). We did not find this to be the case, $F(2, 4.16) = 0.18$, p = .842. We also did not find that demand effects significantly varied depending on whether they were manipulated within- (d = 0.23, 95% CI [0.12, 0.35], p < .001) vs. between-subjects (d = 0.14, 95% CI [0.03, 0.25], p = .016), $F(1, 10.61) = 1.76$, p = .213

Demand effects tended to be slightly more positive for in-person (g = 0.31, 95% CI [0.16, 0.46], p < .001) vs. online (g = 0.10, 95% CI [0.01, 0.19], p = .029) studies, $F(1, 30.58) = 5.92$, p = .021. However, we did not find that demand effects significantly varied depending on whether students (d = 0.27, 95% CI [0.13, 0.40], p < .001), non-students (d = 0.08, 95% CI [-0.01, 0.17], p = .076), or a mix of students and non-students (d = 0.05, 95% CI [-1.00, 1.09], p = .680) were sampled, $F(2, 2.11) = 2.20$, p = .304. We also did not find that demand effects significantly varied depending on whether those participants were paid (d = 0.13, 95% CI [0.00, 0.26], p = .048) vs. unpaid (d = 0.21, 95% CI [0.09, 0.32], p = < .001), $F(1, 20.94) = 0.84$, p = .371.

Table 1. Study feature moderator and subgroup analyses.

Moderator (bolded) and level	*s*	*k*	*g*	95% CI	*F*	*p*
Group comparison	52	252	–	–	1.89	.292
positive vs. control	41	114	0.16	[0.05, 0.27]	8.35	.006
nil vs. control	4	17	0.23	[-0.13, 0.58]	3.15	.152
negative vs. control	17	43	0.16	[0.03, 0.29]	7.13	.016
positive vs. nil	8	34	0.37	[0.02, 0.72]	6.39	.040
positive vs. negative	16	44	0.33	[0.15, 0.51]	15.74	.001
Control vs. non-control group comparison	52	252	–	–	10.55	.008
control	44	174	0.16	[0.08, 0.25]	14.15	< .001
non-control	24	78	0.34	[0.19, 0.49]	22.02	< .001
Control group comparison (see levels above)	44	174	–	–	0.18	.842
Design of demand characteristics manipulation	52	252	–	–	1.76	.213
within-subjects	14	44	0.14	[0.03, 0.25]	7.84	.016
between-subjects	44	208	0.23	[0.12, 0.35]	16.48	< .001
Participant pool	48	204	–	–	2.2	.304
students	36	160	0.27	[0.13, 0.4]	16.16	< .001

Moderator (bolded) and level	*s*	*k*	*g*	95% CI	*F*	*p*
non-students	11	25	0.08	[-0.01, 0.17]	3.96	.076
mix	2	19	0.05	[-1, 1.09]	0.3	.680
Setting	49	239	–	–	5.92	.021
online	18	52	0.1	[0.01, 0.19]	5.75	.029
in-person	32	187	0.31	[0.16, 0.46]	18.03	< .001
Payment	48	212	–	–	0.84	.371
yes	13	50	0.13	[0, 0.26]	4.91	.048
no	36	162	0.21	[0.09, 0.32]	13.75	< .001
Publication status	52	252	–	–	0.11	.748
published	41	239	0.22	[0.13, 0.32]	21.9	< .001
unpublished	11	13	0.17	[-0.17, 0.51]	1.27	.287

Note. s = number of studies; *k* = number of effect size estimates; *g* = Hedge's *g*; 95% CI corresponds to the estimated value of Hedge's *g*; *F*-values represent the test of moderation in bolded rows – and tests of the model-derived overall effect size in non-bolded rows; The number of studies listed for a moderator analysis is not necessarily the sum of the number of studies listed for the individual levels of the moderators because many studies yielded effect sizes for multiple levels of the moderator.

Residual variability. To evaluate how much in-sample variability in demand effects is currently accounted for by study feature moderators, we calculated a pseudo-R^2 statistic. We did so by comparing the sum of the variance components (between-study τ^2 + within-study σ^2) in a model containing only an intercept and a model containing the two study feature moderators that achieved statistical significance: (1) whether the demand characteristics condition was compared to a control group (vs. another condition with demand characteristics), and (2) whether the study was conducted in-person (vs. online). Results indicated that the significant moderators accounted for approximately 15.52% of in-sample variability in demand effects.

Can participants help us understand demand effects? Participants correctly identified the described hypothesis 83% of the time Participants did not generally report having strong beliefs about whether such hypothesized effects would occur (M = 0.50, SD = 0.72). Participants reported that they would be highly capable of adjusting their responses (M = 2.24, SD = 0.44), but not very motivated to do so (M = 0.33, SD = 0.37). Participants also predicted that other subjects would be generally unlikely to adjust their responses to fit the experimenter's hypothesis (M = 0.74, SD = 0.41).

The above results suggest that participants generally report being receptive to demand characteristics, agnostic about hypothesized effects, capable of adjusting their responses, but not motivated to do so. That being said, we remind the reader that these ratings exhibited low reliability (motivation ICC = 0.23; opportunity to adjust responses ICC = 0.23; belief ICC = 0.16). This may be indicative of strong individual differences, but we also later describe the possibility of measurement difficulties (see *Limitations*).

As shown in Table 2, we did not uncover a significant association between observed demand effects and (a) the extent to which participants correctly identified the hypothesis

described in the vignettes, (β = 0.14, 95% CI [-0.02, 0.31], $t(8.23) = 1.99$, $p = .081$), (b) ratings of motivation to adjust responses (β = 0.01, 95% CI [-0.21, 0.22], $t(11.18) = 0.09$, $p = .932$), (c) ratings of opportunity to adjust responses (β = 0.04, 95% CI [-0.02, 0.10], $t(8.66) = 1.56$, $p = .155$), and (d) rated belief in the hypothesized effect (β = 0.06, 95% CI [-0.05, 0.18], $t(11.11) = 1.21$, $p = .252$). Of course, Rosnow and colleagues posited that receptivity, motivation, and opportunity *interact* to shape demand effects. However, when we explored this question, we did not find robust evidence that including all possible higher order interactions significantly improved model fit, $F(7, 2.43) = 1.36$, $p = .463$.

Even after averaging across a large number of noisy forecasts (ICC = 0.22, M = 0.74, SD = 0.41), we also failed to find that participants were able to predict the magnitude of demand effects, β = 0.07, 95% CI [-0.06, 0.21], $t(12.66) = 1.14$, $p = .274$.

Table 2. Participant rating moderator analyses

Moderator	*s*	*k*	*β*	95% CI	*F*	*p*
predicted demand effects	36	151	0.07	[-0.06, 0.21]	1.31	.274
understanding of study hypothesis	36	151	0.14	[-0.02, 0.31]	3.96	.081
motivation to adjust responses	36	151	0.01	[-0.21, 0.22]	0.01	.932
opportunity to adjust responses	36	151	0.04	[-0.02, 0.1]	2.43	.155
belief in communicated hypothesis	36	151	0.06	[-0.05, 0.18]	1.46	.252

Note. s = number of studies; k = number of effect size estimates; β = estimated linear relationship between participant ratings and observed Hedge's g scores; 95% CI corresponds to the estimated value of β.

Publication bias analyses. Overall, publication bias analyses were inconclusive. Both precision-effect tests with 3LMA (β = 0.71, 95% CI [0.06, 1.37], p = .033) and aggregated dependencies (β = 0.31, 95% CI [-0.60, 1.21], p = .507) estimated that publication bias that favored hypothesis-consistent shifts in participants' responses. The estimate, however, was only significant when using 3LMA. The bias-corrected overall effect size estimates with both 3LMA (g = 0.05, 95% CI [-0.12, 0.23], p = .562) and aggregated dependencies (g = 0.12, 95% CI [-0.05, 0.30], p = .175) did not significantly vary from zero. In other words, precision-effect tests did not consistently uncover evidence of publication bias, but did consistently indicate that the overall effect size may not be robust if publication bias does exist. Further complicating matters is the unusual distribution of the funnel plots, especially in regards to two unusually large aggregated effect size estimates (see Figure 4).

Examining aggregated effect sizes using weight-function modeling – as opposed to precision effect tests – yields a different pattern: better fit is achieved in a model where publication bias favored non-significant or hypothesis-inconsistent shifts in participants' responses, $\chi^2(1) = 6.50$, p = .01. The bias-corrected overall effect size was thus upward-adjusted, g = 0.32, 95% CI [0.15, 0.49], p < .001. The discrepancy between precision-effect tests and weight-function modeling may be driven by the unusual distribution of the funnel plots (see Figure 4).

We did not find significant differences in the magnitude of demand effects between published (g = 0.22, 95% CI [0.13, 0.32], p = < .001) and unpublished (g = 0.17, 95% CI [-0.17, 0.51], p = .287) studies, $F(1, 14.38) = 0.11$, p = .748. If there is a biased selection of instances where participants responses shift in a hypothesis consistent manner, sensitivity analyses indicated that it would have to be extreme selection pressure to make the effect size non-

significant (Mathur & VanderWeele, 2020b). Even if hypothesis-consistent shifts were 10000000 times more likely to be published, the overall effect would still be 0.07, 95% CI [0.01, 0.12], p = .019.

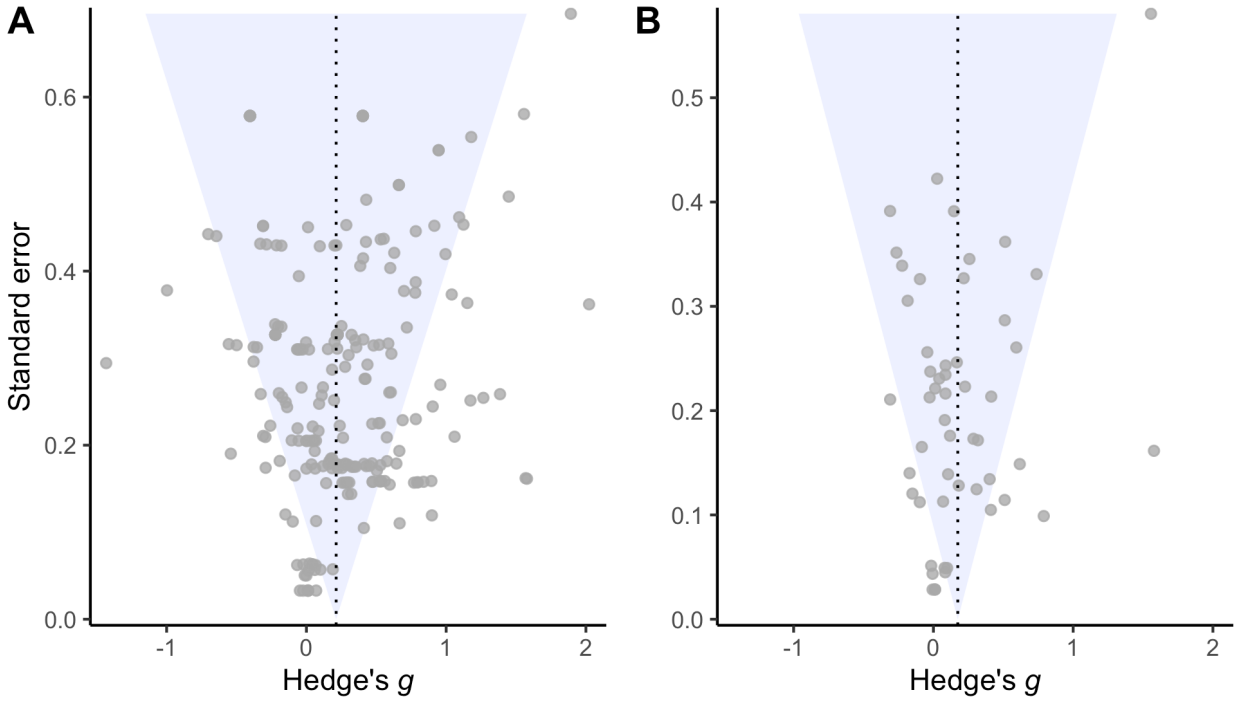

Figure 4. Raw (Panel A) or aggregated (Panel B) effect sizes plotted against their corresponding standard errors. Funnel plot is inverted to illustrate correspondence with slope estimates from precision-effect tests.

Discussion

In the *Introduction*, we described a fictitious discipline that we suspect would be met with extreme skepticism – one plagued by a methodological artifact that (a) can lead to both false positives and false negatives, (b) can create both upward bias and downward bias, (c) has unreliable effects, and (d) is difficult to explain. If one agrees that such a characterization is problematic, we argue they face an uncomfortable observation: our meta-analysis suggests that this characterization also currently applies to experimental psychology.

Since Orne popularized the concept in the mid-1900's, demand characteristics have become a literal textbook methodological concern in experimental psychology. We synthesized a subset of this literature, focusing on 252 effect sizes from 52 studies that provided experimental tests of demand effects by explicitly manipulating cues about the study hypothesis. Consistent with an influential framework developed by Rosnow and colleagues (Rosnow & Aiken, 1973; Rosnow & Rosenthal, 1997; Strohmetz, 2008), the observed and estimated true distribution of these effects suggest that demand characteristics can create false positives (Orne, 1959), false negatives (Hayes & King, 1967), and upward and downward bias (Coles et al., 2022). Such heterogeneity leads to a wide prediction interval, ranging from g = 0.89 (a medium-sized decrease in hypothesis-consistent responding) to g = -0.46 (a medium-sized *increase* in hypothesis-consistent responding).

Observing demand effects that are strong, inferentially consequential, and heterogenous would be less concerning if researchers had an explanation for how such effects operate. Unfortunately, we found little explanation in our own synthesis. Coded study features failed to generate profound insights – revealing only that demand effects tend to be larger when studies are run in-person and include comparisons between two different demand characteristic conditions.

Although such insights are certainly useful, they only explained an estimated 15.52% of in-sample variability in demand effects. Given that in-sample (vs. out-of-sample) estimates of R^2 are often inflated by overfitting (De Rooij & Weeda, 2020), we suspect that the true proportion of explained variability is even lower.

We next examined a popular and influential framework developed by Rosnow and colleagues, who successfully predicted that demand effects are heterogeneous (Rosnow & Aiken, 1973; Rosnow & Rosenthal, 1997; Strohmetz, 2008). However, we found few attempts to test their proposed explanation for such heterogeneity: differences in the extent to which participants are (a) receptive, (b) motivated, and (c) able to respond to demand characteristics. (One exception is an *unpublished* record by Coles, Wyatt, & Frank, 2023)

Contrary to early advice by Orne (1969), we did not find that much clarity emerged when consulting participants themselves. When we provided a large set of naïve participants with summaries of the studies in our meta-analysis, we found that their predictions about demand effects and their underlying mechanisms were unreliable. Indeed, participants generally indicated that they would be receptive to demand characteristics, agnostic about the communicated effects, capable of adjusting their responses, but not motivated to do so. They also generally predicted that other participants would not respond to demand characteristics. Such predictions are clearly at odds with the demand effects observed in our meta-analysis.

Even when averaging across a large number of participant judgments, we failed to find that they were able to predict or explain the mechanisms underlying demand effects. Demand effects were not significantly predicted by the extent to which they (a) correctly identified the communicated hypothesis, (b) reported they would be motivated to adjust responses, (c) reported

they would be able to adjust responses, and (d) reported they would expect the hypothesized effect to emerge.

We currently lack a satisfying explanation for our results. One possibility is that demand characteristics are not driven by receptivity, motivation, opportunity, and/or belief in the experimenter's hypothesis. However, we find this explanation unlikely given the face validity of those proposed mechanisms. Indeed, Rosnow and Rosenthal (1997) argued that opportunity was the least important mechanism because "…not many experimenters would design a study so that a participant would be *incapable* of responding to cues closely tied to the experimenter's own expectation." A second possibility is that these mechanisms are not as essential as Rosnow and colleagues expected. For example, Coles et al. (2022) argued that demand characteristics may activate mechanisms that do not require motivation or direct ability to adjust responses (e.g., conditioned responses). If demand effects are multiply determined, the effect of any single moderator may be weaker than previously expected. A third possibility is that participants are not able to accurately reflect on these mechanisms. For example, Corneille and Lush (2023) suggested that participants may occasionally be unaware that they were motivated to adjust their responses – e.g., in cases of phenomenological control. A fourth possibility, of course, is that our own methodological limitations inhibited our ability to detect the effects of these moderators. We discuss these limitations next.

Limitations

Our meta-analysis is, of course, not without limitations. It still remains unclear why participants were generally unable to predict and explain demand effects. It is unclear if our inclusion criteria were too narrow – or not narrow enough. And we suspect that there are many

supplemental analyses that can be performed on our openly-available data to further probe the nature of demand effects.

Orne (1969) suggested that participants themselves may help researchers understand demand effects. At first glance, this assumption seems reasonable. Participants are capable of predicting a variety of effects in psychology when exposed to information about the study procedures (Corneille & Béna, 2023) – and this very procedure is often used to raise concerns about demand characteristics (Bartels, 2019). We, however, failed to find that similar procedures could be used to predict or explain demand effects at the meta-analytic level. Yet, it is unclear whether this is a valid and important insight in itself – or indicative of our own methodological shortcomings. For example, perhaps participants are too different than the original participants (Gergen, 1973), perhaps they need to experience the study context first-hand (Orne, 1969), and perhaps they need better measures of the psychological mechanisms that may underlie demand effects (Flake & Fried, 2020).

Broad definitions of the demand characteristics construct presented us with yet another challenge. At its broadest, demand characteristics are defined as almost *any* cue that may impact participants' understanding of the purpose of the study, including instructions, rumors, and experimenter behavior (Orne, 1962). However, such a definition arguably creates a boundless conceptual space where any systematic change in a research design or setting might be considered a threat to scientific inferences. We focused our meta-analysis on a subset of the conceptual space that is more amenable to precise definition and study: explicit cues of the study hypothesis. Although we do not reject broader definitions of demand characteristics, we suspect that such broadening will only further deepen the mystery surrounding their effects.

Even with our relatively narrow subset of the demand characteristics literature, there are commensurability challenges. Researchers have tested the effects of explicit hypothesis cues on a variety of outcomes, including hypnosis symptoms (e.g., Orne & Scheibe, 1964), eating behavior (e.g., Kersbergen, Whitelock, Haynes, Schroor, & Robinson, 2019), visual judgments (e.g., Durgin, Klein, Spiegel, Strawser, & Williams, 2012), relationship satisfaction (e.g., Cramer, 2005), mood (e.g., Coles et al., 2022), policy support (e.g., Mummolo & Peterson, 2019), test scores (e.g., Veitch, Gifford, & Hine, 1991), and so on. Researchers also varied in how they conducted their investigations – e.g., in whether they (a) conducted their studies in-person (e.g., Orne & Scheibe, 1964) vs. online (e.g., Mummolo & Peterson, 2019), (b) sampled students (e.g., Rose, Geers, Fowler, & Rasinski, 2014) vs. non-students (e.g., Terhune & Smith, 2006), and (c) manipulated hypothesis cues within- (e.g., Martin, Sackur, & Dienes, 2018) vs. between-subjects (e.g., Coles et al., 2022).[8] We generally failed to uncover evidence that such methodological differences explain a meaningful proportion of variability in demand effects. Nonetheless, it is possible that such a large number of [often unsystematic] differences between studies limits power to detect meaningful moderators in the demand characteristics literature. However, we

[8] Whether meta-analysts should combine effects from within and between-subjects design has sparked considerable debate (Morris, 2002). We felt that combining such effects was justified given that (a) effect sizes were converted into a common metric using design-specific estimates of sampling variance, (b) sensitivity analyses of assumed within-subject correlations produced virtually no change in our overall effect size estimate, and (c) we did not detect significant differences between studies that manipulated explicit hypothesis cues within- vs. between-subjects.

note that high heterogeneity was also observed in a six-lab investigation of demand effects in conformity research (Coles et al., 2024)[9].

Last, although we performed a large number of robustness checks, these checks certainly were not exhaustive. Future researchers may wish to consider alternative search strategies, inclusion criteria, approaches to quantifying effects, methods for estimating participant judgments, and decisions about how to model the data.

We do not deny the importance of these methodological limitations. Instead, we point out that we suspect they do little to change our conclusion: demand effects can be inferentially consequential – but are unreliable, difficult to predict, and challenging to explain.

Concluding Remarks

Since Orne (1962) famously described the idea over 50 years ago, demand characteristics have become a literal textbook methodological concern in experimental psychology (Sharpe & Whelton, 2016). Orne (1962) further suggested that demand characteristics constituted an omnipresent threat to the validity of experimental psychology, arguing that "…all experiments will have demand characteristics, and these will always have some effects" (1962, p. 779). Consequently, it is perhaps not surprising that significant effort has been dedicated to their study.

Unfortunately, it is not clear if much has been learned in these 50+ years since Orne warned of this "omnipresent threat". Our review suggests that demand effects can indeed be inferentially consequential – but are also unreliable, difficult to predict, and challenging to

[9] In a six-lab investigation of explicit demand effects in conformity research, (Coles et al., 2024) observed no overall effect ($g = 0.02$) but high heterogeneity ($\tau = 0.20$).

explain. We imagine two potential responses from psychologists. One possibility is that we revive efforts to understand demand effects – i.e., investigate the individual differences, situational factors, and mechanisms driving their heterogeneity. A second possibility is that we continue business as usual: paying lip service to demand characteristics as a fundamental methodological issue, acknowledging that it threatens the validity of experimental psychology on multiple fronts, and making little progress towards a precise understanding of its effects. Based on what we have observed from the past half century, we pessimistically hypothesize that psychologists will continue to do the latter. Ironically, though, the effects of our explicitly stated hypothesis remain unclear.

References

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

* Allen, A. P., & Smith, A. P. (2012). Demand characteristics, pre-test attitudes and time-on-task trends in the effects of chewing gum on attention and reported mood in healthy volunteers. *Appetite*, *59*(2), 349–356. https://doi.org/0.1016/j.appet.2012.05.026

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

* Balze, E. M. (1998). *The role of expectancy in treatment efficacy: The use of a therapeutic" frame" to enhance performance* (PhD thesis).

* Barabasz, M., Barabasz, A., & O'Neill, M. (1991). Effects of experimental context, demand characteristics, and situational cues: New data. *Perceptual and Motor Skills*, *73*(1), 83–92. https://doi.org/0.2466/pms.1991.73.1.83

Bartels, J. (2019). Revisiting the stanford prison experiment, again: Examining demand characteristics in the guard orientation. *The Journal of Social Psychology*, *159*(6), 780–790. https://doi.org/10.1080/00224545.2019.1596058

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

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. https://doi.org/10.1002/9780470743386

* Busch, A. M., Kanter, J. W., Sedivy, S. K., & Leonard, J. L. (2007). A follow-up analogue study on the effectiveness of the cognitive rationale. *Cognitive Therapy and Research*, *31*, 805–815. https://doi.org/0.1007/s10608-007-9121-6

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

Cohen, J. (2013). *Statistical power analysis for the behavioral sciences* (Vol. 2). New York, NY: Lawrence Erlbaum Associates. https://doi.org/10.4324/9780203771587

* 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*. https://doi.org/10.1037/pspa0000316

* Coles, N. A., McCullough, M., Oishi, S., Dang, A., McCauley, T., Pfattheicher, S., … Sarabia, V. (2024). *Six unpublished investigations of the role of demand characteristics in a conformity paradigm*.

* Coles, N. A., Wyatt, M., & Frank, M. C. (2023). *Coles, Wyatt, and Frank unpublished replication of Coles et al. 2022*.

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. https://doi.org/10.1037/h0028849

Corneille, O., & Béna, J. (2023). Instruction-based replication studies raise challenging questions for psychological science. *Collabra: Psychology*, *9*(1). https://doi.org/10.1525/collabra.82234

Corneille, O., & Lush, P. (2023). Sixty years after orne's american psychologist article: A conceptual framework for subjective experiences elicited by demand characteristics. *Personality and Social Psychology Review*, *27*(1), 81–101. https://doi.org/10.1177/10888683221104368

* Cramer, D. (2004). Effect of the destructive disagreement belief on relationship satisfaction with a romantic partner or closest friend. *Psychology and Psychotherapy: Theory, Research and Practice*, *77*(1), 121–133. https://doi.org/10.1348/147608304322874290

* Cramer, D. (2005). Effect of the destructive disagreement belief on satisfaction with one's closest friend. *The Journal of Psychology*, *139*(1), 57–66. https://doi.org/10.3200/JRLP.139.1.57-66

* Cramer, D., & Buckland, N. (1995). Effect of rational and irrational statements and demand characteristics on task anxiety. *The Journal of Psychology*, *129*(3), 269–275. https://doi.org/10.1080/00223980.1995.9914964

De Rooij, M., & Weeda, W. (2020). Cross-validation: A method every psychologist should know. *Advances in Methods and Practices in Psychological Science*, *3*(2), 248–263. https://doi.org/10.1177/2515245919898466

Downs, S. H., & Black, N. (1998). The feasibility of creating a checklist for the assessment of the methodological quality both of randomised and non-randomised studies of health care interventions. *Journal of Epidemiology & Community Health*, *52*(6), 377–384. https://doi.org/10.1136/jech.52.6.377

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. https://doi.org/10.1177/0145445516673998

* Durgin, F. H., Klein, B., Spiegel, A., Strawser, C. J., & Williams, M. (2012). The social psychology of perception experiments: Hills, backpacks, glucose, and the problem of generalizability. *Journal of Experimental Psychology: Human Perception and Performance*, *38*(6), 1582-1595. https://doi.org/10.1037/a0027805

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

Fillenbaun, S., & Frey, R. (1970). More on the" faithful" behavior of suspicious subjects. *Journal of Personality*, *38*(1), 43–51. https://doi.org/10.1111/j.1467-6494.1970.tb00636.x

Flake, J. K., & Fried, E. I. (2020). Measurement schmeasurement: Questionable measurement practices and how to avoid them. *Advances in Methods and Practices in Psychological Science*, *3*(4), 456–465. https://doi.org/10.1177/2515245920952393

Franco, A., Malhotra, N., & Simonovits, G. (2014). Publication bias in the social sciences: Unlocking the file drawer. *Science*, *345*(6203), 1502–1505. https://doi.org/10.1126/science.1255484

Frank, M. C., Braginsky, M., Cachia, J., Coles, N., Hardwicke, T., Hawkins, R., … Williams, R. (2023). *Experimentology: An open science approach to experimental psychology methods*. Boston, MA: MIT Press. https://doi.org/10.25936/3JP6-5M50

* Fresson, M., Dardenne, B., Geurten, M., Anzaldi, L., & Meulemans, T. (2017). The role of self-transcendence and cognitive processes in the response expectancy effect. *Psychologica Belgica*, *57*(2), 77–92. https://doi.org/10.5334/pb.364

Gergen, K. J. (1973). Social psychology as history. *Journal of Personality and Social Psychology*, *26*(2), 309-320. https://doi.org/10.1037/h0034436

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. https://doi.org/10.3758/BF03210266

* Hoogeveen, S., Schjoedt, U., & Elk, M. van. (2018). Did I do that? Expectancy effects of brain stimulation on error-related negativity and sense of agency. *Journal of Cognitive Neuroscience*, *30*(11), 1720–1733. https://doi.org/10.1162/jocn_a_01297

Hyman, H. H. (1954). *Interviewing in social research*. Chicago, IL: University of Chicago Press.

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

* Kanter, J. W., Kohlenberg, R. J., & Loftus, E. F. (2004). Experimental and psychotherapeutic demand characteristics and the cognitive therapy rationale: An analogue study. *Cognitive Therapy and Research*, *28*, 229–239. https://doi.org/10.1023/B:COTR.0000021542.40547.15

* Kenealy, P. (1988). Validation of a music mood induction procedure: Some preliminary findings. *Cognition & Emotion*, *2*(1), 41–48. https://doi.org/10.1080/02699938808415228

* Kersbergen, I., Whitelock, V., Haynes, A., Schroor, M., & Robinson, E. (2019). Hypothesis awareness as a demand characteristic in laboratory-based eating behaviour research: An experimental study. *Appetite*, *141*, 104318. https://doi.org/10.1016/j.appet.2019.104318

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

* Larsen, J. T., & McGraw, A. P. (2011). Further evidence for mixed emotions. *Journal of Personality and Social Psychology*, *100*(6), 1095–1110. https://doi.org/10.1037/a0021846

Lüdecke, D., Ben-Shachar, M. S., Patil, I., Waggoner, P., & Makowski, D. (2021). performance: An R package for assessment, comparison and testing of statistical models. *Journal of Open Source Software*, *6*(60), 3139. https://doi.org/10.21105/joss.03139

* Martin, J.-R., Sackur, J., & Dienes, Z. (2018). Attention or instruction: Do sustained attentional abilities really differ between high and low hypnotisable persons? *Psychological Research*, *82*(4), 700–707. https://doi.org/10.1007/s00426-017-0850-1

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

Mathur, M. B., & VanderWeele, T. J. (2020a). *PublicationBias: Sensitivity analysis for publication bias in meta-analyses*. Retrieved from https://CRAN.R-project.org/package=PublicationBias

Mathur, M. B., & VanderWeele, T. J. (2020b). Sensitivity analysis for publication bias in meta-analyses. *Journal of the Royal Statistical Society Series C: Applied Statistics*, *69*(5), 1091–1119. https://doi.org/10.1111/rssc.12440

* McGinley, H., Kaplan, M., & Kinsey, T. (1975). Subject effects and demand characteristics. *Psychological Reports*, *36*(1), 267–278. https://doi.org/10.2466/pr0.1975.36.1.267

* McGlynn, F. D., Gaynor, R., & Puhr, J. (1972). Experimental desensitization of snake-avoidance after an instructional manipulation. *Journal of Clinical Psychology*. *28*(2), 224.

Morris, S. B., & DeShon, R. P. (2002). Combining effect size estimates in meta-analysis with repeated measures and independent-groups designs. *Psychological Methods*, *7*(1), 105–125. https://doi.org/10.1037/1082-989x.7.1.105

* Mummolo, J., & Peterson, E. (2019). Demand effects in survey experiments: An empirical assessment. *American Political Science Review*, *113*(2), 517–529. https://doi.org/10.1017/S0003055418000837

Orne, M. T. (1959). The nature of hypnosis: Artifact and essence. *The Journal of Abnormal and Social Psychology*, *58*(3), 277–299. https://doi.org/10.1037/h0046128

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. https://doi.org/10.1037/h0043424

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. https://doi.org/10.1093/acprof:oso/9780195385540.003.0005

* Orne, M. T., & Scheibe, K. E. (1964). The contribution of nondeprivation factors in the production of sensory deprivation effects: The psychology of the "panic button". *The Journal of Abnormal and Social Psychology*, *68*(1), 3–12. https://doi.org/10.1037/h0048803

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., et al. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ*, *372*. https://doi.org/10.1136/bmj.n71

* Palomba, D., & Stegagno, L. (1995). Dissociation between actual and expected cardiac changes: Interoception and emotional experience. In D. Vaitl & R. Schandry (Eds.), *From the heart to the brain: The psychophysiology of circulation – brain interaction* (pp. 283–298). Peter Lang Publishing.

* Perry, D. G., Roots, R. D., & Perry, L. C. (1978). Demand awareness and participant willingness as determinants of aggressive response to film violence. *The Journal of Social Psychology*, *105*(2), 265–275. https://doi.org/10.1080/00224545.1978.9924124

* Polivy, J., & Doyle, C. (1980). Laboratory induction of mood states through the reading of self-referent mood statements: Affective changes or demand characteristics? *Journal of Abnormal Psychology*, *89*(2), 286-290. https://doi.org/10.1037/0021-843x.89.2.286

Pustejovsky, J. E., & Tipton, E. (2018). Small-sample methods for cluster-robust variance estimation and hypothesis testing in fixed effects models. *Journal of Business & Economic Statistics*, *36*(4), 672–683. https://doi.org/10.1080/07350015.2016.1247004

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

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

Rodgers, M. A., & Pustejovsky, J. E. (2021). Evaluating meta-analytic methods to detect selective reporting in the presence of dependent effect sizes. *Psychological Methods*, *26*(2), 141. https://doi.org/10.1037/met0000300

* Rose, J. P., Geers, A. L., Fowler, S. L., & Rasinski, H. M. (2014). Choice-making, expectations, and treatment positivity: How and when choosing shapes aversive experiences. *Journal of Behavioral Decision Making*, *27*(1), 1–10. https://doi.org/10.1002/bdm.1775

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. https://doi.org/10.1093/acprof:oso/9780195385540.003.0007

Rosnow, R. L., & Aiken, L. S. (1973). Mediation of artifacts in behavioral research. *Journal of Experimental Social Psychology*, *9*(3), 181–201. https://doi.org/10.1016/0022-1031(73)90009-7

Rosnow, R. L., & Rosenthal, R. (1997). *People studying people: Artifacts and ethics in behavioral research*. New York, NY: Freeman. http://dx.doi.org/10.34944/dspace/69

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. https://doi.org/10.1186/1472-6947-7-16

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

Sharpe, D., & Whelton, W. J. (2016). Frightened by an old scarecrow: The remarkable resilience of demand characteristics. *Review of General Psychology*, *20*(4), 349–368. https://doi.org/10.1037/gpr0000087

* Siegel, W. E., & Hagen, R. L. (1982). The influence of demand characteristics and expectancies in the measurement of salivary response. *Journal of Behavioral Assessment*, *4*, 179–185. https://doi.org/10.1007/BF01321391

Sigall, H., Aronson, E., & Van Hoose, T. (1970). The cooperative subject: Myth or reality? *Journal of Experimental Social Psychology*, *6*(1), 1–10. https://doi.org/10.1016/0022-1031(70)90072-7

Silverman, I., & Marcantonio, C. (1965). Demand characteristics versus dissonance reduction as determinants of failure-seeking behavior. *Journal of Personality and Social Psychology*, *2*(6), 882–884. https://doi.org/10.1037/h0022628

* Smith, J. M., Bell, P. A., & Fusco, M. E. (1986). The influence of color and demand characteristics on muscle strength and affective ratings of the environment. *The Journal of General Psychology*, *113*(3), 289–297. https://doi.org/10.1080/00221309.1986.9711040

* 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. https://doi.org/10.1002/jrsm.1095

Stewart-Williams, S., & Podd, J. (2004). The placebo effect: Dissolving the expectancy versus conditioning debate. *Psychological Bulletin*, *130*(2), 324–340. https://doi.org/10.1037/0033-2909.130.2.324

Strohmetz, D. B. (2008). Research artifacts and the social psychology of psychological experiments. *Social and Personality Psychology Compass*, *2*(2), 861–877. https://doi.org/10.1111/j.1751-9004.2007.00072.x

* Terhune, D. B., & Smith, M. D. (2006). The induction of anomalous experiences in a mirror-gazing facility: Suggestion, cognitive perceptual personality traits and phenomenological state effects. *The Journal of Nervous and Mental Disease*, *194*(6), 415–421. https://doi.org/10.1097/01.nmd.0000221318.30692.a5

* Tsai, N., Buschkuehl, M., Kamarsu, S., Shah, P., Jonides, J., & Jaeggi, S. M. (2018). (Un) great expectations: The role of placebo effects in cognitive training. *Journal of Applied Research in Memory and Cognition*, *7*(4), 564–573. https://doi.org/10.1016/j.jarmac.2018.06.001

* Veitch, J. A., Gifford, R., & Hine, D. W. (1991). Demand characteristics and full spectrum lighting effects on performance and mood. *Journal of Environmental Psychology*, *11*(1), 87–95. https://doi.org/10.1016/S0272-4944(05)80007-6

* Verpaelst, C. C., & Standing, L. G. (2007). Demand characteristics of music affect performance on the wonderlic personnel test of intelligence. *Perceptual and Motor Skills*, *104*(1), 153–154. https://doi.org/10.2466/pms.104.1.153-154

Vevea, J. L., & Hedges, L. V. (1995). A general linear model for estimating effect size in the presence of publication bias. *Psychometrika*, *60*(3), 419–435. https://doi.org/10.1007/BF02294384

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, *36*(3), 1–48. https://doi.org/10.18637/jss.v036.i03

Contributions

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

Acknowledgments

We thank (1) (anonymous for peer review; AC) for assistance with code review, and (2) (anonymous for peer review; JB) for assistance developing the initial literature search strategy.

Funding

This work was supported by the John Templeton Foundation (grant # anonymous for peer review). The funder had no role in the preparation of the manuscript or decision to publish.

Ethics statement

Ethics approval was not requested for the meta-analysis because no new data were collected. The vignette rating study was reviewed and approved by the (anonymous for peer review) IRB (protocol #: anonymous for peer review; protocol title: anonymous for peer review).

Competing Interests

The authors declare no competing interests.

Data Accessibility

The project pre-registration, materials, data, and code are openly available at https://osf.io/3hkre/?view_only=2dc92af53f194e5eab0d7aecafaf01c2.