Coles, N., Wyatt, M., & Frank, M. C. (n.d.). A Meta-Analysis of the Impact and Heterogeneity of Explicit Demand Characteristics. *Collabra: Psychology*. https://doi.org/10.1525/collabra.143005

Methodology and Research Practice

A Meta-Analysis of the Impact and Heterogeneity of Explicit Demand Characteristics

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

[1] Department of Psychology, University of Florida, Gainesville, FL, USA, [2] Center for the Study of Language and Information, Stanford University, Stanford, CA, USA, [3] Department of Psychology, Stanford University, Stanford, CA, USA

Keywords: demand characteristics, meta-analysis, methodology, confound, causal inference

https://doi.org/10.1525/collabra.143005

Collabra: Psychology

Demand characteristics are a fundamental methodological concern in experimental psychology. Yet, little is known about the direction, magnitude, and consistency of their effects. We conducted a three-level meta-analysis of 252 effect sizes from 52 studies that manipulated explicit demand characteristics (EDCs). On average, EDCs led to small overall increases in hypothesis-consistent responding ($g = 0.21$, 95% CI [0.12, 0.31]). However, the effects were heterogeneous (between-study $\tau = 0.28$; within-study $\sigma = 0.18$), 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). Consistent with previous theorizing, the observed and estimated distribution of effects suggest that demand characteristics can create false positives, false negatives, upward bias, and downward bias. These unpredictable inferential consequences suggest that further research is needed to test mechanisms theorized to underlie the effects of demand characteristics.

> "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 group of people approach you and begin telling you about a method they invented for understanding humans. They tell you that their method is useful for estimating causal relationships, but add that there is an issue: it can sometimes be thrown off by a *methodological artifact*. They explain that this artifact sometimes causes them to detect an effect that is not real, and other times miss an effect that *is* real; that it sometimes makes an effect appear bigger than it actually is, and other times smaller. And that, in general, it is unclear when, why, and to what extent this artifact impacts their conclusions.

The above scenario, we argue, serves as an abstraction of a methodological puzzle in experimental psychology: the role of *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. However, later research helped establish that 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 exaggerated *or* underestimated depending on whether the experimenter communicates expectations of positive or nil effects (Coles et al., 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, explicit manipulations of demand characteristics consistently failed to produce statistically significant changes in participants' responses (Mummolo & Peterson, 2019).

As this short review illustrates, demand characteristics presents several similarities to the methodological artifact described in the opening of this paper. Demand characteristics are a literal textbook methodological concern in experimental psychology (Sharpe & Whelton, 2016) that (a) can lead to both false positives and false negatives, (b) can create both upward bias and downward bias, but (c) don't always appear to matter. In the present work, we take stock of progress on this puzzle via a meta-analysis of a unique methodological response: experiments *on* explicit

[a] Correspondence be addressed to: ncoles@ufl.edu

demand characteristics. To begin, we review the theoretical and conceptual frameworks that guided the investigation.

How do Demand Characteristics Alter Participant Responses?

One of the most influential frameworks for conceptualizing the effects of demand characteristics was developed by Rosnow and colleagues (Rosnow & Aiken, 1973; Rosnow & Rosenthal, 1997; Strohmetz, 2008). In this framework, they identified three factors that moderate the effects of demand characteristics: (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. 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. Furthermore, even if the infant possessed the astonishing ability to read, it's possible they would misunderstand the cues (Corneille & Lush, 2023) – which could be considered another form of non-receptivity.

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. For historical context, early work on demand characteristics was marked by debates about the extent to which participants are motivated to (a) help the researcher confirm their hypothesis (Orne, 1962), (b) receive positive evaluations (Riecken, 1962; Sigall et al., 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: (b) *acquiesce* (i.e., provide hypothesis-consistent responses), (b) *non-acquiesce* (i.e., not change their responses based on knowledge about the hypothesis), 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.

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 explicitly highlighted 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 motivation-irrelevant manners — e.g., via conditioned responses or other mechanisms discussed in conceptually-related work on placebo effects (Stewart-Williams & Podd, 2004). Regardless of proposed mechanisms, these frameworks converge on a prediction: the effects of demand characteristics will be heterogeneous.

Experiments on Demand Characteristics

Rosnow and colleagues' framework emerged during a time when researchers increasingly began conducting experiments on demand characteristics (McGuire, 2009). For example, Orne and Scheibe (1964) reported that participants were more likely to report sensory deprivation side-effects (e.g., hallucinations) when told that "…Such experiences are not unusual under the conditions to which you are to be subjected". Similarly, Perry, Roots, and Perry (1978) found that participants were more likely to exhibit film-induced aggressive behavior when told that the researcher anticipated such a response.

Experiments like these manipulate what we will call *explicit demand characteristics* (EDCs). Orne (1962) defined demand characteristics broadly – as *any* cue that impacts participants' beliefs about the purpose of the study. This not only includes explicit information from the experimenter (i.e., EDCs), but also more subtle information, like rumors, television shows, and courses. These more subtle sources are, of course, important to study – but more difficult to precisely manipulate. Thus, we suspect that the choice to study EDCs is more so based on methodologically convenience.

Despite their convenience, EDCs have a notable limitation: they are not representative of the typical experimental context, wherein experimenters usually refrain from disclosing their hypotheses. Thus, the methodology has been adopted with some trepidation. For example, Orne warned that EDCs may cause participants to "lean over backwards to be honest" (i.e., non-acquiesce; 1962, p. 779) or engage in "paradoxical action" (i.e., possibly counter-acquiesce; 2009, p. 116). However, Orne also used the method in his own work – often finding that participants' responses are indeed influenced by EDCs (e.g., Orne & Scheibe, 1964). Over the next few decades, dozens more studies would follow, providing an opportunity to evaluate the magnitude, consistency, and potential moderators of such effects via meta-analysis.

Methodology

We defined the scope of the meta-analysis using the Population, Intervention, Comparison, Outcome framework for structuring research questions (Schardt et al., 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 demand characteristics (EDCs) – operationalized as scenarios where a researcher tells participants about the effect of an independent variable on a dependent variable. Our comparison-of-interest were conditions where either no hypothesis or a different hypothesis was communicated to participants. Last, the 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 Stanford University 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, but all eligible records were published in English. Our search also did not have date restrictions. In total, the search yielded 1457 records (168 of which were unpublished).

Screening

Records must have met the following criteria to be eligible for inclusion:

- The researcher manipulated what participants were told about the effect of an independent variable on a dependent variable.[1] 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.

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, 54 studies from 39 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 (Hedges' g) as our effect size index (Borenstein, 2009; Cohen, 2013).

In most cases (67%), we estimated the main effect of EDCs. For example, Coles et al. (2022) manipulated whether participants were told that posing smiles would increase happiness. Here, the main effect of EDCs was computed by comparing happiness ratings from smiling participants who were either informed or not informed of the mood-boosting effect of smiling.

In other cases (33%), we estimated the *interactive* effect of EDCs. For example, in the same Coles et al. (2022) study, participants provided happiness ratings both after smiling

[1] We excluded conditions where the researcher communicated a *non-directional* effect. We did so because we worried participants in these scenarios could not unambiguously infer how the researcher expected their response 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.

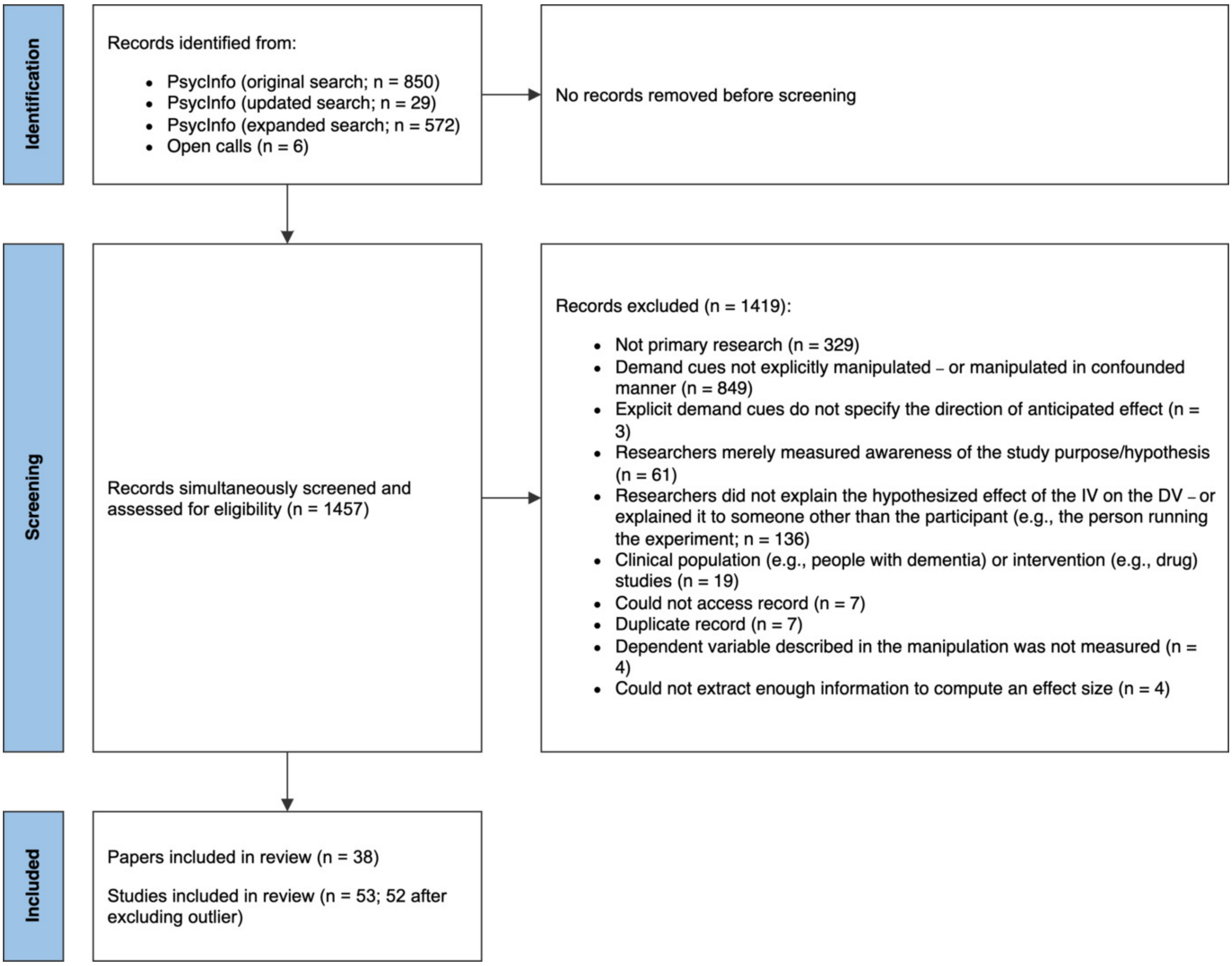
Identification
Records identified from:
• PsycInfo (original search; n = 850)
• PsycInfo (updated search; n = 29)
• PsycInfo (expanded search; n = 572)
• Open calls (n = 6)
No records removed before screening
Screening
Records simultaneously screened and assessed for eligibility (n = 1457)
Records excluded (n = 1419):
• Not primary research (n = 329)
• Demand cues not explicitly manipulated – or manipulated in confounded manner (n = 849)
• Explicit demand cues do not specify the direction of anticipated effect (n = 3)
• Researchers merely measured awareness of the study purpose/hypothesis (n = 61)
• Researchers did not explain the hypothesized effect of the IV on the DV – or explained it to someone other than the participant (e.g., the person running the experiment; n = 136)
• Clinical population (e.g., people with dementia) or intervention (e.g., drug) studies (n = 19)
• Could not access record (n = 7)
• Duplicate record (n = 7)
• Dependent variable described in the manipulation was not measured (n = 4)
• Could not extract enough information to compute an effect size (n = 4)
Included
Papers included in review (n = 38)
Studies included in review (n = 53; 52 after excluding outlier)

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

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 EDCs. In this scenario, the interactive effect of EDCs 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 would be mood boosting, an increase in mood would be coded as a positive effect. If, however, participants were told that the intervention would be mood *dampening*, that same increase in mood would be coded as a negative effect.

We calculated Hedges' g by applying a small sample correction to Cohen's d_s (for between-subject designs) and d_{rm} (for within-subject designs[2]) estimates. In 64% of cases, we used the *M*s and *SD*s reported in a paper (or extracted using WebPlotDigitizer; Drevon et al., 2017). When these values were not available, we used (in order of preference), (1) t-values (5%), (2) F-values (5%), or (4) p-values (23%). In instances where relevant information was not provided but the statistical significance and direction of the effect was described, we assumed p-values of .04 and .50 for statistically significant and non-significant effects respectively (e.g., Kenealy, 1988). In a few instances (3%), 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 et al., 2011).

Nearly all studies (74%) contained multiple effect sizes of interest. For example, the full design in Coles et al. (2022) included a positive demand, nil demand, and control

[2] 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.

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 the effects of EDCs:

- *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. Less often, 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 EDCs were manipulated 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.

For descriptive purposes, we also coded the country where the investigation was performed.

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 EDCs, 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).

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 statistical 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

We used three main approaches for assessing and correcting for potential publication bias in our estimation of the overall effect of EDCs.

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, statistically 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[3].

Second, we conducted precision-effect tests (PET). In PET, 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 (Stanley & Doucouliagos, 2014). 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.

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 statistical 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 statistically 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/. This link also contains a list of amendments/deviations we made to our pre-registration as the project evolved and we responded to reviewer feedback. Sample size was determined by the availability of relevant records. All code has been checked for reproducibility, including a script that creates a computationally reproducible manuscript using the papaja R package (Aust & Barth, 2022).

Results

In total, we analyzed 253 effect sizes from 53 studies from between the years 1964 and 2024 (M = 2003, SD = 18.71). Of these studies, 11 were unpublished.

In order of frequency, effect sizes represented a positive demand compared to a control group (k = 115), positive demand to negative demand (k = 44), negative demand to a control group (k = 43), positive demand to a nil demand group (k = 34), or nil demand to a control group (k = 17). Effect sizes tended to compare one demand condition to a control group (k = 175) – as opposed to a group exposed to a different type of demand condition (k = 78). Regardless of what type of demand manipulation was used, it was more common to manipulate the cues between (k = 209) vs. within subjects (k = 44).

Most effect sizes came from student samples (k = 160), although some samples were non-students (k = 26), 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 = 163), although some were paid (k = 50) and some were not described thoroughly enough to make a determination (k = 40). Most effect sizes came from in-person studies (k = 188), but some were from online studies (k = 52) or not described thoroughly enough to make a determination (k = 13). Most research was conducted in the United States (k = 127), United Kingdom (k = 49), Canada (k = 21), and the Netherlands (k = 18). However, occasionally, we found work conducted in France (k = 9), Kenya (k = 6), Australia (k = 3), Belgium (k = 2), Turkey, Poland, Malaysia, and Denmark (all k = 1).

Overall Results

Overall, results indicated that EDCs cause participants' responses to shift in a manner consistent with the communicated hypothesis (i.e., acquiesce), g = 0.20, 95% CI [0.11, 0.30], $t(48.13) = 4.28$, $p < .001$. 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.

As a reminder, rather than assuming that there is a *single true effect* of EDCs, 3LMA assumes a distribution of *multiple true effects*. Consistent with this assumption, observed variability in the effects of EDCs exceeded what would be expected from sampling error alone (between-study $\tau = 0.29$, $I^2 = 61.46$; within-study $\sigma = 0.18$, $I^2 = 23.85$; total $Q(252) = 978.70$, $p < .001$, total $I^2 = 85.30$). 3LMA often assumes that the multiple true effects form a normal distribution, which we recreated based on estimates of the average effect

[3] 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.

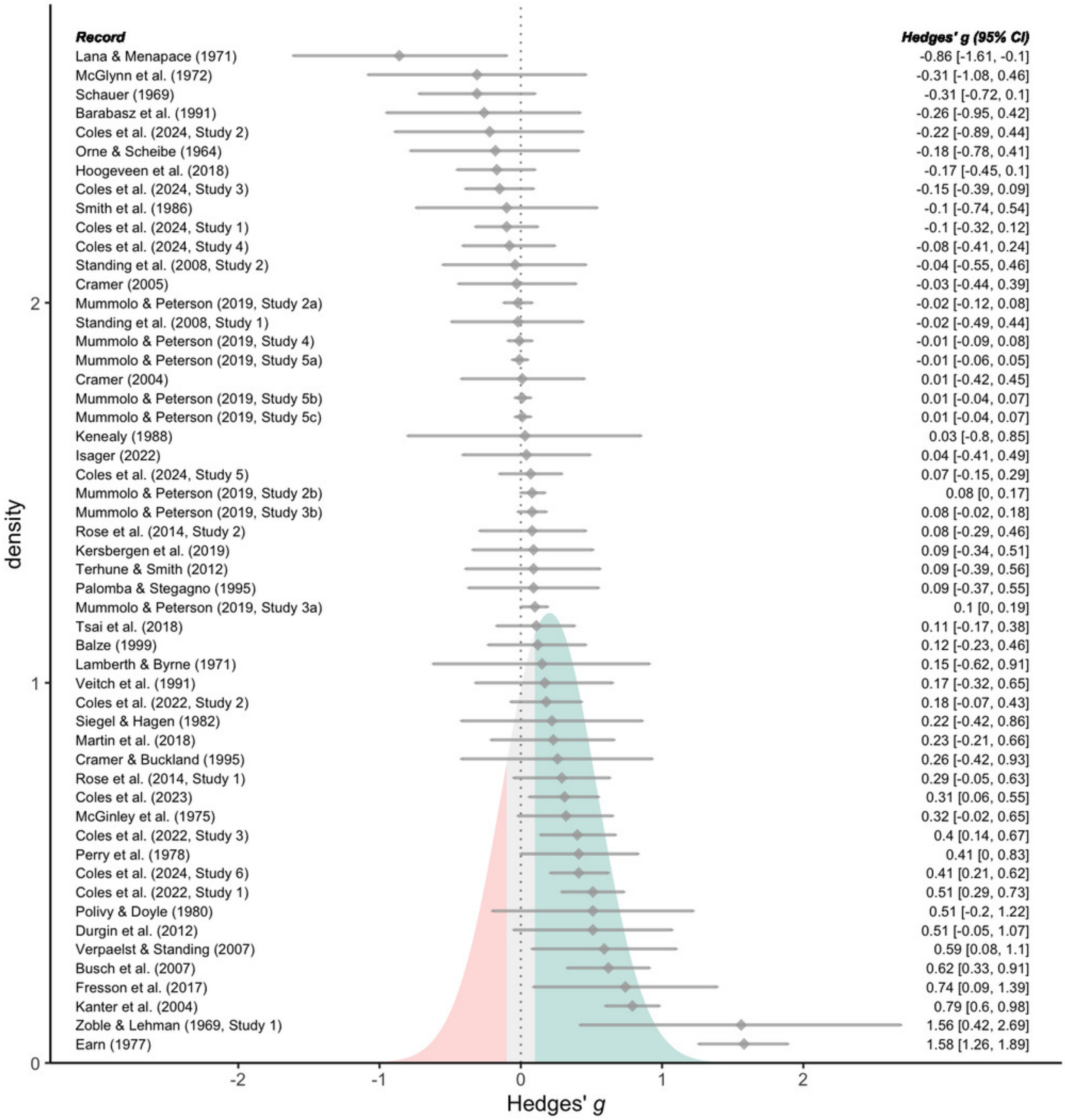

Figure 2. Forest plot of effect sizes (grey diamonds), their 95% confidence intervals (grey error bars), and their citations (left). For visualization purposes, effect sizes are aggregated within-studies (see openly-available data for non-aggregated effect sizes). The estimated effect size distribution is also shown and colored based on whether EDCs 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).

size and variability attributed to sources other than sampling error (between-study τ + within-study σ). As shown in Figure 2, this estimated distribution suggests that EDCs can have a wide range of effects. Indeed, the 95% prediction interval for a single future study ranges from g = -0.48 to g = 0.89.

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 EDCs most often produce hypothesis-consistent shifts (62%), but sometimes produce negligible shifts (20%) or shifts in the *opposite* direction of the communicated hypothesis (18%). 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.

Table 1. Study moderator and subgroup analyses.

Moderator (bolded) and level	*s*	*k*	*g*	95% CI	F	*p*
Group comparison	53	253	–	–	1.93	.287
positive vs. control	42	115	0.15	[0.04, 0.26]	7.14	.011
nil vs. control	4	17	0.22	[-0.14, 0.58]	2.91	.164
negative vs. control	17	43	0.16	[0.03, 0.29]	6.4	.021
positive vs. nil	8	34	0.36	[0.02, 0.71]	6.13	.043
positive vs. negative	16	44	0.32	[0.15, 0.5]	15.15	.001
Control vs. non-control group comparison	53	253	–	–	10.87	.008
control	45	175	0.15	[0.07, 0.24]	12.11	.001
non-control	24	78	0.33	[0.18, 0.48]	20.99	< .001
Control group comparison (levels above)	45	175	–	–	0.2	.828
Design of EDC manipulation	53	253	–	–	1.55	.240
within-subjects	14	44	0.14	[0.02, 0.25]	7.07	.020
between-subjects	45	209	0.22	[0.11, 0.34]	14.65	< .001
Participant pool	49	205	–	–	2.44	.282
students	36	160	0.26	[0.13, 0.4]	15.99	< .001
non-students	12	26	0.05	[-0.06, 0.16]	1.08	.323
mix	2	19	0.05	[-1, 1.09]	0.3	.680
Setting	50	240	–	–	4.81	.036
online	18	52	0.1	[0.01, 0.19]	5.68	.030
in-person	33	188	0.29	[0.14, 0.44]	15.49	< .001
Payment	49	213	–	–	0.55	.465
yes	13	50	0.13	[0, 0.26]	4.92	.047
no	37	163	0.19	[0.08, 0.31]	11.71	.002
Publication status	53	253	–	–	0.07	.801
published	42	240	0.21	[0.11, 0.31]	18.98	< .001
unpublished	11	13	0.17	[-0.17, 0.51]	1.26	.289

Note. *s* = number of studies; *k* = number of effect size estimates; *g* = Hedges' *g*; 95% CI corresponds to the estimated value of Hedges' *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.

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 the effects of EDCs 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 EDCs), and (2) whether the study was conducted in-person (vs. online).

The effects of EDCs were estimated to be twice as large when two demand characteristic conditions were compared (g = 0.33, 95% CI [0.18, 0.48], p < .001), as opposed to one demand characteristic condition being compared to a control group (g = 0.15, 95% CI [0.07, 0.24], p = .001), F(1, 10.37) = 10.87, p = .008. This suggests that the effects of EDCs 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, 3.52) = 1.93, p = .287.

Instances where a demand characteristic condition was compared to a control group allowed us to test whether participants' responses shift more when they expect that the researcher hypothesizes an increase (i.e., positive demand; g = 0.17, 95% CI [0.06, 0.28], p = .003), a decrease (i.e., negative demand; g = 0.19, 95% CI [0.06, 0.33], p = .007), or no change in the dependent variable (i.e., nil demand; g = 0.26, 95% CI [-0.21, 0.74], p = .178). We did not find this to be the case, F(2, 4.15) = 0.20, p = .828. We also did not find that the effects of EDCs varied depending on whether they were manipulated within- (g = 0.22, 95% CI [0.11, 0.34], p < .001) vs. between-subjects (g = 0.14, 95% CI [0.02, 0.25], p = .020), F(1, 10.48) = 1.55, p = .240

The effects of EDCs tended to be slightly more positive for in-person (g = 0.29, 95% CI [0.14, 0.44], p < .001) vs. online (g = 0.10, 95% CI [0.01, 0.19], p = .030) studies, F(1, 30.22) = 4.81, p = .036. However, we did not find that demand effects significantly varied depending on whether students (g = 0.26, 95% CI [0.13, 0.40], p < .001), non-students (g = 0.05, 95% CI [-0.06, 0.16], p = .323), or a mix of students and non-students (g = 0.05, 95% CI [-1.00, 1.09], p = .680) were sampled, F(2, 2.12) = 2.44, p = .282. We also did not find that the effects of EDCs varied depending on whether those participants were paid (g = 0.13, 95% CI [0.00, 0.26], p = .047) vs. unpaid (g = 0.19, 95% CI [0.08, 0.31], p = .002), F(1, 20.74) = 0.55, p = .465.

Residual Variability

To evaluate how much in-sample variability in the effects of EDCs 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 statistically significant moderators accounted for approximately 16.46% of in-sample variability in the effects of EDCs.

Publication Bias Analyses

Overall, publication bias analyses were inconclusive. Both PET with 3LMA (β = 0.65, 95% CI [-0.02, 1.31], p = .057) and aggregated dependencies (β = 0.09, 95% CI [-0.81, 1.00], p = .844) estimated that publication bias favored hypothesis-consistent shifts in participants' responses. The estimates, however, were not statistically significant. The bias-corrected overall effect size estimates with both 3LMA (g = 0.06, 95% CI [-0.12, 0.23], p = .535) and aggregated dependencies (g = 0.15, 95% CI [-0.03, 0.33], p = .107) 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 3).

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 statistically 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 3).

We did not find differences in the magnitude of demand effects between published (g = 0.21, 95% CI [0.11, 0.31], p = < .001) and unpublished (g = 0.17, 95% CI [-0.17, 0.51], p = .289) studies, F(1, 14.38) = 0.07, p = .801. 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 statistically non-significant (Mathur & VanderWeele, 2020b). Even if hypothesis-consistent shifts were 10,000,000 times more likely to be published, the overall effect would still be statistically significant, g = 0.06, 95% CI [0.01, 0.12], p = .026.

Discussion

In the *Introduction*, we described a methodological puzzle: researchers have found that demand characteristics (a) sometimes lead to false positive, other times to false negatives, (b) sometimes lead to exaggerated effect size estimates and other times overly conservative effect size estimates, and (c) sometimes don't seem to matter at all. The results of our meta-analysis of studies that experimentally manipulate explicit demand characteristics (EDCs) provides more formal evidence of this phenomenon. EDCs *typically* lead to small increases in hypothesis-consistent responding. However, such effects are heterogeneous – with prediction intervals ranging from a medium-sized *increase* to a medium-sized *decrease* in hypothesis-consistent responding.

Over a half century after first describing their model, the accumulation of evidence from research on EDCs supports Rosnow and colleagues' key prediction: the effects of demand characteristics are heterogeneous. Some of this heterogeneity can be linked to how researchers design their studies – e.g., whether they run participants in-person and/or test multiple sets of demand characteristics. However, in our results, study-level moderators only explained approximately 15% of in-sample variability in the effects of EDCs. Researchers may ultimately explain more heterogeneity by measuring their proposed underlying mechanisms: receptivity to cues, motivation to adjust responses, and opportunity to adjust responses. However, we note that our own attempts to do were largely inconclusive (see *Supplemental Information*). More specifically, we attempted to estimate underlying mechanisms through ratings from new participants who reviewed study summaries. However, as reviewed in the *Supplemental Information*, these ratings were neither reliable nor predictive of observed demand effect.

Rosnow and colleagues predicted that participants occasionally counter-acquiesce – i.e., adjust their responses in the opposite direction they think researchers predict (Rosnow & Aiken, 1973; Rosnow & Rosenthal, 1997; Strohmetz, 2008). This phenomenon is sometimes referred to as the "screw you effect" (Masling, 1966). Our results provide mixed evidence for such an effect. In 252 tests of EDCs, only 2 (< 1%) yielded statistically significant evidence of counter-acquiescence. Furthermore, when aggregating dependent effect sizes, no test yielded statistically significant evidence of counter-acquiescence. Thus, although the distribution modeled by our meta-analysis suggests that counter-acquiescence effects should occur with some regularity, such effects have been rarely observed in work with EDCs.

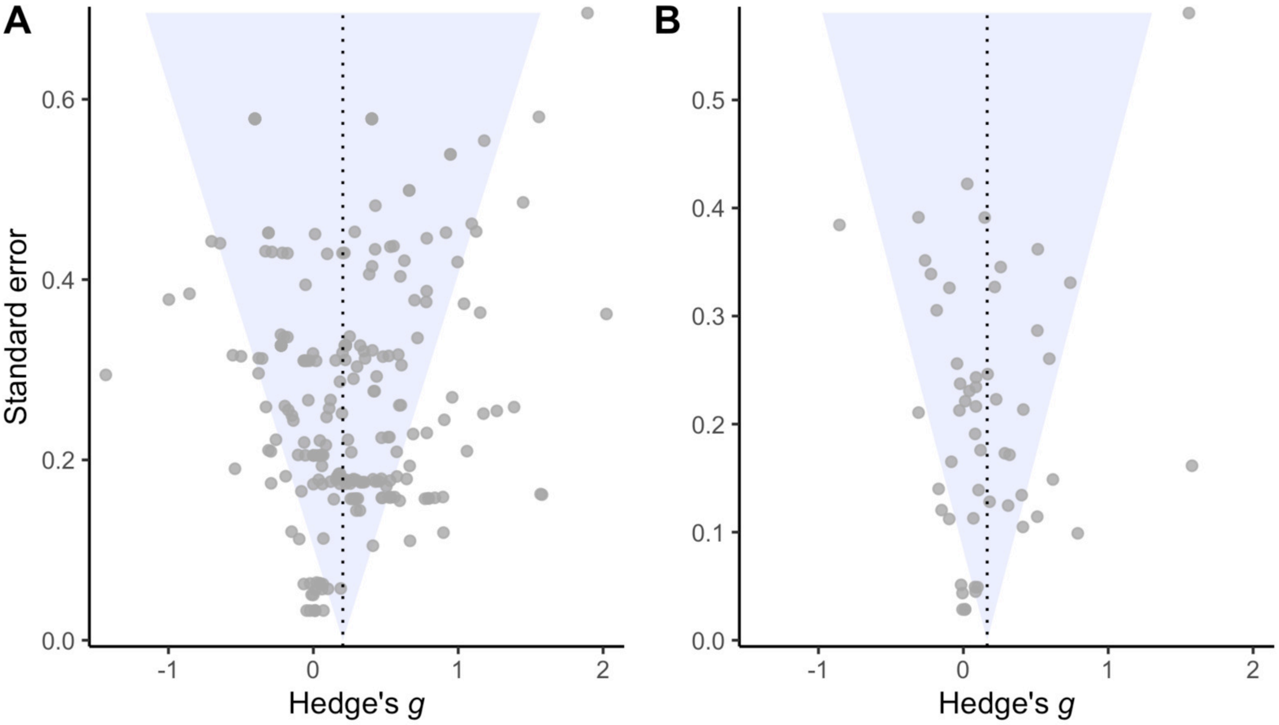

Figure 3. Raw (Panel A) or aggregated (Panel B) effect sizes plotted against their corresponding standard errors. Funnel plot is inverted to illustrate correspondence with slope estimates from precision-effect tests.

Limitations and Future Directions

As we worked to refine our own understanding of the methodological puzzle presented by demand characteristics, we encountered two major conceptual challenges: (1) varied operationalizations of demand characteristics, and (2) the commensurability of the existing evidence base.

Operationalizing Demand Characteristics

At their 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). One benefit of such a broad definition is that it highlights multiple reasons to reject an alternative methodological assumption: that human subjects enter studies as relatively blank slates. For historical context, Silverman & Schulman (1970) remarked that "…we came to regard putting input into a human subject as something akin to putting chemicals into a test tube. We are now coming to full awareness that the analogy holds only with the profound qualification that we are inevitably working with an unclean test tube."

Unfortunately, one drawback of accepting a broad definition of demand characteristics is that you are left with a test tube believed to be contaminated by virtually everything. By focusing on explicit demand characteristics (EDCs), researchers have furthered their understanding of a specific type of contaminant. However, EDCs are not representative of the contaminants typically encountered in research with human subjects – where researchers often go through great lengths to *not* explicitly reveal their hypothesis. Our meta-analysis clearly rejects previous concerns that participants ignore (Orne, 1962) or counter-acquiesce (Orne, 2009) against EDCs. Indeed, we found that participants most commonly respond to EDCs the same way Orne suggested they respond to other (less explicit) demand characteristics: by helping the experimenter confirm the hypothesis. Nonetheless, studying a combination of relatively implicit and explicit demand characteristics would bolster confidence in the generalizability of the evidence base.

Commensurability

Even with our relatively narrow subset of the demand characteristics literature, we encountered commensurability challenges. Researchers have tested the effects of EDCs on a variety of outcomes, including hypnosis symptoms (e.g., Orne & Scheibe, 1964), eating behavior (e.g., Kersbergen et al., 2019), visual judgments (e.g., Durgin et al., 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 et al., 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 et al., 2014) vs. non-students (e.g., Terhune & Smith, 2006), and (c) manipulated hypothesis cues within- (e.g., Martin et al., 2018) vs. between-subjects (e.g., Coles et al., 2022).

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. Manipulating such differences systematically in the future (e.g., in a single experimental design) would help clarify which (if any) of these methodological decisions are most impactful.

Conclusion

Since Orne (1962) famously described the idea over 60 years ago, demand characteristics have become a literal textbook methodological concern in experimental psychology (Frank et al., 2025; Sharpe & Whelton, 2016). Over these past 50 years, a clearer picture of the puzzle has emerged. Our meta-analysis on the effects of explicit demand characteristics suggests that they can produce false positives, false negatives, overestimated effect sizes, and underestimated effect sizes. Yet, while corroborating some theoretical frameworks (e.g., Rosnow & Rosenthal, 1997), results highlight that large parts of this puzzle remain unsolved. In general, explicit demand characteristics cause participants to change their responses in a hypothesis-consistent manner – on average. However, sometimes participants seem to ignore the researcher's hypothesis – and perhaps in extremely rare scenarios, behave in the opposite manner. Orne (1962) characterized this issue as an omnipresent threat, arguing that "…all experiments will have demand characteristics" (p. 779). If true, unresolved questions about when, why, and how demand characteristics impact participants' responses are of fundamental import.

Data Accessibility Statement

All materials, data, code, and pre-registrations are openly available at https://osf.io/3hkre/.

Funding

This work was supported by the John Templeton Foundation (grant #62295). The funder had no role in the preparation of the manuscript or decision to publish.

Competing Interests

Authors have no reported conflicts of interest. We thank Anjie Cao, Gabrielle Grant, and Joao Francisco Goes Braga Takayanagi for assistance with code review.

Author Contributions

NAC: Conceptualization, Data Curation, Formal Analysis, Investigation, Methodology, Project administration, Software, Supervision, Visualization, Writing - Original Draft Preparation, Writing - Review & Editing

MW: Data Curation, Investigation, Project administration, Software, Writing - Review & Editing

MCF: Formal Analysis, Investigation, Methodology, Project administration, Resources, Software, Supervision, Visualization, Writing - Review & Editing.

Editors: Louis Moses (Associate Editor)

Submitted: December 31, 2024 PDT. Accepted: July 22, 2025 PDT.

References

Allen, A. P., & Smith, A. P. (2012). Demand characteristics, pre-test attitudes and time-on-task trends in the effects of chewing gum on attention and reported mood in healthy volunteers. *Appetite*, *59*(2), 349–356. https://doi.org/10.1016/j.appet.2012.05.026

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

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). Russell Sage Foundation.

Borenstein, M., Hedges, L. V., Higgins, J. P., & Rothstein, H. R. (2011). *Introduction to meta-analysis*. John Wiley & Sons.

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

Cohen, J. (2013). *Statistical power analysis for the behavioral sciences* (Vol. 2). 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*.

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., & 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. (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

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.

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

Frank, M. C., Braginsky, M., Cachia, J., Coles, N., Hardwicke, T., Hawkins, R., ... Williams, R. (2025). *Experimentology: An open science approach to experimental psychology methods*. MIT Press. https://doi.org/10.7551/mitpress/14810.001.0001

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

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

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

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*. 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

McGuire, W. J. (2009). Suspiciousness of experimenter's intent. In *Artifacts in Behavioral Research* (pp. 15–47). Oxford. https://doi.org/10.1093/acprof:oso/9780195385540.003.0002

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. (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., & 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. https://doi.org/10.1037/h0048803

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., ... Mulrow, C. D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ*, 372.

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

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*. R Foundation for Statistical Computing. 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). 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

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*. Freeman.

Schardt, C., Adams, M. B., Owens, T., Keitz, S., & Fontelo, P. (2007). Utilization of the PICO framework to improve searching PubMed for clinical questions. *BMC Medical Informatics and Decision Making*, *7*(1), 1–6. https://doi.org/10.1186/1472-6947-7-16

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

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.

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

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

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

Supplementary Materials

Supplemental Materials

Download: https://collabra.scholasticahq.com/article/143005-a-meta-analysis-of-the-impact-and-heterogeneity-of-explicit-demand-characteristics/attachment/296916.docx?auth_token=Hs91T9jwWbOU4jmeo2WQ

Peer Review Communication

Download: https://collabra.scholasticahq.com/article/143005-a-meta-analysis-of-the-impact-and-heterogeneity-of-explicit-demand-characteristics/attachment/296917.docx?auth_token=Hs91T9jwWbOU4jmeo2WQ