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

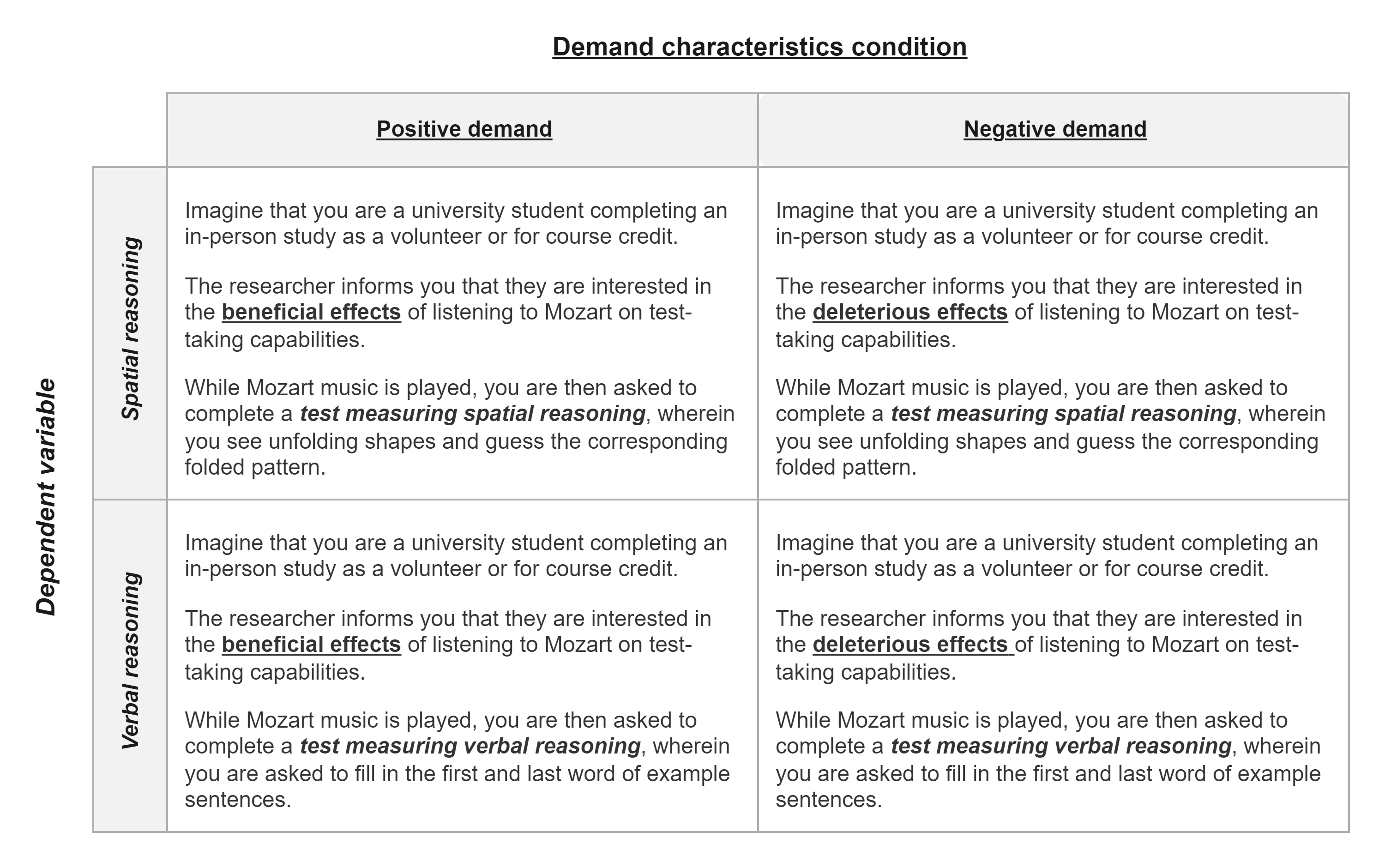
# Can research participants help us understand demand effects?

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[[1]](#footnote-1), 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 (Supplemental Figure 1). 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.



*Supplemental Figure 1*. 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) were 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.[[2]](#footnote-2) 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, 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 ( = 13.82).

### Accounting for different demand comparisons.

Hedge’s 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.

## Vignette study results

#### 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 and/or measurement difficulties.

As shown in *Supplemental Table 1*, 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, ( = 0.14, 95% CI [-0.02, 0.31], (8.23) = 1.99, = .081), (b) ratings of motivation to adjust responses ( = 0.01, 95% CI [-0.21, 0.22], (11.18) = 0.09, = .932), (c) ratings of opportunity to adjust responses ( = 0.04, 95% CI [-0.02, 0.10], (8.66) = 1.56, = .155), and (d) rated belief in the hypothesized effect ( = 0.06, 95% CI [-0.05, 0.18], (11.11) = 1.21, = .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, = 0.07, 95% CI [-0.06, 0.21], (12.66) = 1.14, = .274.

*Supplemental Table 1*. 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*.

## Vignette study limitations.

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 (Flake & Fried, 2020) of the psychological mechanisms that may underlie demand effects.

1. 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. [↑](#footnote-ref-1)
2. 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. [↑](#footnote-ref-2)