Survey Design for Mediation Analysis

April 25, 2019

Abstract

Causal mediation analysis (CMA) requires measurement of an outcome with and without some treatment, plus a set of mediator variables. There is no consensus on whether survey-based studies should measure potentially mediating variables before or after the outcome variable(s). We show that order can be consequential. Survey design decisions affect conclusions drawn from mediation analysis. We demonstrate this by replicating a recent study. Randomizing order is often prudent, but best practice depends on details, so there is no one-size-fits-all optimal survey design for valid CMA.

Word count: 5,441

The use of causal mediation analysis (CMA) has seen explosive growth. At its core, CMA seeks to pinpoint which mechanisms, of multiple plausible alternatives, generate treatment effects. Researchers conducting survey experiments have been the most enthusiastic adapters of these techniques. Generally, experimental researchers are interested in the effect of a particular treatment, such as a vignette, on an outcome of interest, such as support for a policy choice, mediated through a particular mechanism or channel. To assess the strength of these channels, the researcher measures a quantity related to that channel and then assesses the degree to which treatment effects can be attributed to changes in that measured quantity and the outcome.[1]

Researchers decompose their estimated treatment effects into such quantities as the Average Causal Mediation Effect (ACME), which "represents the indirect effects of the treatment on the outcome *through the mediating variable*" and the Average Direct Effect (ADE), which represents "the causal effect of the treatment on the outcome that is not transmitted by the hypothesized mediator" (Imai et al., 2011, p 769, emphasis in original). From these quantities, one calculates the Proportion Mediated (PM), the proportion of the treatment effect that is transmitted through a particular mediator. Different mediators are often linked to different theoretically informed mechanisms, so comparing PMs is one way researchers adjudicate between rival theories.

However, using CMA for survey experiments entails an often un-scrutinized choice of survey design: whether to measure mediators before or after the outcome of interest. Survey instruments either measure outcomes *then* mediators (an "OM" design) or mediators *then* outcomes (an "MO" design). This choice is generally viewed as innocuous, and is rarely discussed. Yet, it can have large effects on the realizations of those quantities and therefore, the inferences drawn from results.

Ex ante, it is unclear why either design would be preferable. Measuring mediators before measuring the outcome more closely resembles the causal pathway that the researcher has in mind,

[1]Recent examples include (Tomz and Weeks, 2013) described below, (Baker, 2015) on attitudes towards foreign aid, and (Lupu, 2013) on polarization.

where treatment affects the value of a mediator, which in turn affects the outcome. But researchers are often most concerned with the direct treatment effect, and therefore choose to measure the outcome directly after treatment, to avoid the possibility that asking other questions in the interim will moderate — magnify, mute, or alter — treatment effects.

We make three related points about the choice between OM and MO designs.

1. OM vs. MO can have a direct effect on the distribution of outcome variables and mediators.

2. OM vs. MO can alter the effect of treatment on outcome variables and mediators.

3. OM vs MO can affect mediation results even when it doesn't moderate the overall treatment effect.

The first and second effects arise because survey order can shift the means of outcome or mediator variables, as well as affect the likelihood of respondents choosing particular values or options for questions. Any given question can alter how respondents understand later questions, by "priming" latent thoughts or "framing" an issue. Even absent such substantive links, simple survey fatigue can induce "satisficing" on later questions Holbrook et al. (2007). In OM designs, that pattern could attenuate the relationship between treatment and mediator; it can attenuate the effect of treatment on outcomes in MO designs.

The third effect is more subtle, and arises as a consequence of the first two. Even if the overall effect of treatment on outcome is similar in OM and MO designs, the portion of treatment effects attributed to particular mediation channels can nonetheless change in meaningful ways between designs. OM and MO designs can yield results that place different weights on different mediation channels, even for the same basic experimental setup, and even if the estimated treatment effects are similar across both designs.

We demonstrate these effects by replicating a recent survey experiment regarding democratic-peace theory. In addition to randomizing treatment assignment, following the original study, we

randomly assigned respondents to an OM or MO survey instrument. The two modules proved similar in regard to distributions of key variables, and in direct treatment effect. However, which mediating variables stood out as the most important mechanisms associated differed across designs.

Unfortunately, using one design or the other cannot "bound" a result. In other words, neither an OM nor MO approach necessarily produces a lower or upper bound on results for mediation effects or the proportion of treatment through a particular mediator. Dependence across responses to survey items can take so many possible forms that it is not practical to propose a general rule for choosing between, or averaging, discrepant findings.

We conclude with a recommendation that researchers be aware of this issue, and discuss it in the particular context of their research. They should assess the degree to which OM versus MO designs can affect their results, based on contextual knowledge. Ultimately, randomly assigning respondents to an OM or MO version of a particular experiment will often prove to be the optimal approach.

1 Why MO vs. OM Might Matter

We surveyed 33 published articles and papers from 2012-17 that used CMA with a survey experiment: 8 employed an OM design; 8 used an MO design; for 15 of the papers, we could not tell from the descriptions in the text or supplementary analyses what design was used. Only 2 papers discussed the OM-MO choice, and conducted experiments using both setups.[2]

How might survey design decisions affect empirical results? Figure 1 shows all of the theoretical ways that the choice of OM versus MO can affect mediation results. The causal diagram of interest in mediation analysis concerns the bottom three nodes and the solid black lines. Assignment to a particular treatment condition (T) can have a direct effect on the outcome (O), and can also have an indirect effect, by altering levels of mediators (M), which, in turn, affect the out-

[2]Tomz and Weeks (2013) and Huddleston and Weller (2017).

come. For example, the regime type of a target country (T) can affect the perceived threat from that country (M), which then affects the respondent's approval of a military strike on the target (O).

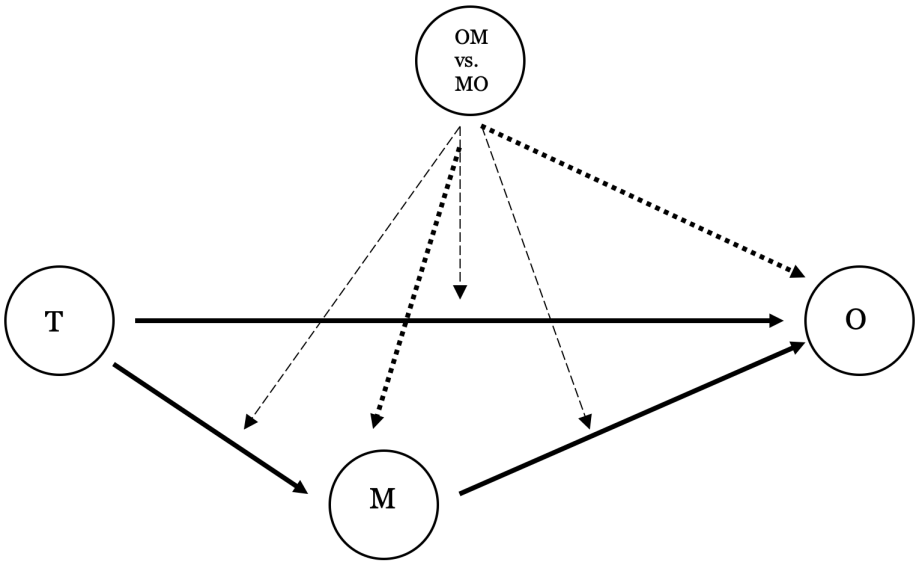

Figure 1: Ways that MO versus OM can affect mediation results

We represent the choice of research design with the top node. Just as a researcher assigns respondents to a treatment or control condition, the choice of the precise survey instrument structure can potentially affect quantities of interest.

First, this choice can have direct effects on the quantities of interest that are directly measured in mediation analysis, as represented by the dotted lines pointing to the mediator's value and the value of the outcome. In their studies of surveys with multiple experiments, Transue, Lee and Aldrich (2009) and Gaines, Kuklinski and Quirk (2007) identify how treatment assignment in one experiment can affect realizations of the outcome variable for a later assignment, which the former call "mean bias." A similar phenomenon is possible even in single-experiment CMA studies, since the presence or absence of a mediator question can affect responses to an outcome question, and vice-versa.

Existing research in survey design and psychology suggests numerous ways that the OM versus

MO choice could have these direct effects. It has long been recognized that question order can affect responses, as can response order within questions (Sudman and Bradburn, 1974). The two can even interact: Holbrook et al. (2007) conclude, from an analysis of over 500 Gallup items, that questions later in a survey induce more "satisficing" wherein respondents' likelihood of selecting first or last options increases, particularly for complicated questions and for less well-educated respondents.

These order effects can be so strong that, sometimes, question order is deliberately varied as an experimental treatment to explore framing or moderating effects, eg Sniderman and Carmines (1997). When question order is not deliberately manipulated in search of cue- or priming effects of this sort, many would say that best practice is rotation of response options.

Second, the choice of OM versus MO can moderate the relationships of interest in mediation analysis as represented by the dashed lines. The OM/MO choice can alter the way that treatment affects the mediator, the way that the mediator relates to the outcome, or the way that treatment affects the outcome through channels other than those in the mediation analysis. In the context of surveys with multiple experiments, Transue, Lee and Aldrich (2009) refer to a similar phenomenon, "inference bias," where treatment assignment for an experiment can affect estimated treatment effects for later embedded experiments.

Each of the survey order-studies described above also provides intuition for how these moderating effects can arise. If question ordering pushes some items further along in the survey, then respondent fatigue means that the OM/MO decision can attenuate relationships between treatment, outcome, and mediators. For example, in an MO design, a long list of mediators might attenuate the effect of treatment on outcome. Similarly, in an OM design, a set of multiple outcome measures might attenuate the relationship between treatment and mediators.

Survey design can also have effects that are not simply attenuation, since questions asked first can prime respondents or frame issues in particular ways. Asking mediator questions can frame the issue at hand, potentially magnifying or muting treatment effects on the outcome. In an MO

design, asking questions about the mediators potentially causes the respondent to consider the issue in a particular light, which can have unpredictable effects on the relationship between treatment and outcome Tourangeau, Rips and Rasinski (2000).

For example, in a study of audience costs, Huddleston (forthcoming) finds that the presence or absence of a question about the costs of military action has both of the effects described above. Asking about costs has a direct effect on approval of a military intervention. Additionally, a question about the costs of intervention ameliorated the effects of the main treatment of interest. He attributes these differences to the priming effect of the costs question, which alters the time horizon respondents use to assess interventions.

1.1 Effect on CMA estimates

If the OM/MO choice has any of these effects, then it can theoretically change the quantities of interest in CMA. Researchers applying CMA often focus on four things: the average treatment effect (ATE), average causal mediation effect (ACME), the average direct effect (ADE), and the proportion mediated (PM).

Define $Y_i(t, m)$ as the potential outcome for individual i, when she is assigned to treatment t and when the mediating variable for her equals m. Since treatment can also affect mediating variables, let $M_i(t)$ denote the potential value for the mediating variable when individual i is assigned to treatment t. For simplicity, let t be binary, with 0 indicating assignment to control and 1 to treatment.

The ATE refers to the difference in outcomes under assignment to treatment versus control, and is not unique to CMA. Since treatment is randomly assigned in experimental designs, this quantity can be calculated in straightforward ways, such as simple comparisons of means in outcomes across treatment conditions.

The ACME, in Equation 1, refers to the difference in the potential outcomes when the mediator takes on the value induced from being assigned to treatment versus assignment to control. Any

survey design that affects the relationship between treatment and the mediating variables has a direct effect on this quantity, since it can change $M_i(1)$ and/or $M_i(0)$, the values of the mediator for individual i when treatment equals 1 or 0 respectively. As Imai et al. (2011) note, "If the treatment has no effect on the mediator, ie $M_i(1) = M_i(0)$, then the causal mediation effects are zero" (p. 769). By extension, design choices that alter this relationship also alter the estimated ACME.

$$\delta_i(t) = Y_i(t, M_i(1)) - Y_i(t, M_i(0)) \tag{1}$$

The ADE, in Equation 2, refers to the effect of treatment on outcome, holding fixed the value of the mediating variable. Here too, any survey design choice that affects the relationship between treatment and outcome can alter this quantity. Even if survey design has no effect on the relationship between treatment and mediator, $M_i(t)$, if it affects the relationship between treatment and Y_i, then it can change the ADE.

$$\zeta_i(t) = Y_i(1, M_i(t)) - Y_i(0, M_i(t)) \tag{2}$$

Researchers often estimate the proportion of the treatment effect that goes "through" the mediating variable. Usually, this quantity refers to the ratio of the ADE to the ATE. Since the ADE refers to the part of the ATE that is *not* through the mediator, the remaining portion of the ATE can be attributed to the effect of treatment through the mediator. This identity can also be seen by observing that the ATE equals the ADE plus the ACME.

$$PM = 1 - \frac{ADE}{ADE + ACME} \tag{3}$$

Consider the effect of simple survey fatigue on the proportion mediated. In an OM design, the mediator question comes later, and is prone to being measured with more error. In turn, the estimated effect of treatment on the mediator can decrease, decreasing the ACME, and lowering the

proportion of the treatment effect attributed to that particular mediator. In an MO design, survey fatigue can have the opposite effect by lowering the ADE.

It would take strong theoretical expectations to specify, *a priori*, whether an MO or an OM result is more plausible, or least biased. The possibility that certain items "frame" later items (i.e. cue respondents to interpret them in a different manner) is extremely general: the effects might be on any moment of the later distribution, and can alter associations across variables in any direction. Which items are most difficult to measure, and most prone to fatigue-based error can be hard to pre-specify.

If one is particularly concerned about mediator items framing outcome items, and vice versa, then the distributions of mediator questions from an MO battery and of the outcome variables from an OM battery are preferable. It is possible to identify bounds on the correlation between variables from only marginal distributions, and, in the absence of a joint distribution, one can sometimes improve on these bounds by employing linear, multiple-variable models of each of the variables of interest. So, in theory, one approach to discrepant findings would be to use the best (earliest) measurement from each half of a split-sample survey. We do not pursue this approach hereafter because bounds on correlation are typically of minimal interest, and are often wide even with seemingly promising candidates for covariates.[3] We conjecture that the effect of at-all-wide bounds on correlation, carried through a complicated CMA, would almost always be extreme (unhelpful) uncertainty about the computed quantities that lend themselves to substantive interpretation.

2 Replication Study

Tomz and Weeks (2013) use a survey experiment to provide micro-foundations for one of the most prominent arguments in political science, the democratic peace. Empirically, militarized conflict

[3]For instance, for the costs mediator variable (x) and the uncollapsed favorability outcome variable (y) discussed below, we observed $r_{xy|OM} = 0.13$ and $r_{xy|MO} = 0.22$. Using covariates from the analysis below, we computed auxiliary regressions that convert to decidedly unhelpful bounds of $-0.83 < r_{x|MO,y|OM} < 0.84$.

between democratic countries rarely occurs. Some theoretical explanations for this phenomenon argue that citizens in democratic countries have preferences against fighting other democracies. To assess the effects of an adversary's regime type on citizen preferences, Tomz and Weeks surveyed approximately 2,000 respondents in waves fielded in both the United States and the United Kingdom in 2010. The survey used a vignette to describe a hypothetical situation wherein the US or UK government was confronted by a foreign country that was developing nuclear weapons. The authors randomized whether the foreign country was described as "a democracy" or "not a democracy." The outcome of interest was measured by asking whether the respondent would favor using armed force to attack the foreign country's nuclear development sites. Respondents chose from five ordered options: "favor strongly"; "favor"; "neither favor nor oppose"; "oppose"; and, "oppose strongly".

The authors also investigated several theoretically motivated channels through which an adversary's regime type might affect preferences over conflict. They measure respondents' perceptions of the *costs* associated with attack, *threats* from not attacking, the likelihood of attack *succeeding* at deterrence, and the *(im)morality* of a preemptive strike. Each of these mediators was grounded in existing research on the democratic peace as a plausible explanation for how democracy might affect conflict attitudes.

The authors hypothesized, and found consistent evidence for, a negative overall treatment effect: respondents were significantly less supportive of a military strike against foreign democracies, as compared to non-democracies. Point estimates of this effect, in bivariate analysis, were 11-12 percentage points They further investigated whether the main treatment effect of democracy operated through one or more of the four mediating channels: *threat*, *morality*, *costs*, and *success*. They first assessed whether treatment affected each of the mediators. They then calculated the proportions of the treatment effect that can be attributed to each of the various mediation channels.

They found that approximately 34% of the treatment effect of democracy could be attributed to its effect through changing perceptions of threats, and another 15% could be assigned to democracy

status changing beliefs about morality of attack. They found little evidence for mediation through the costs or likelihood of success, with only 4% and 6% of the treatment effect attributable to those effects, respectively. The balance of the democracy effect, about 40%, is direct and unmediated.

They conclude that the most prominent mechanism behind this finding concerns threat perceptions: "The finding that democracies view other democracies as less threatening, which in turn reduces support for using force, accords with major works on the democratic peace that emphasize threat perception." Because they find that little of the treatment effect can be attributed to costs and success, they discount explanations based on those potential channels. And since morality accounts for a larger proportion of the treatment effects, they conclude that morality as a mechanism behind the democratic peace "should be a major topic of future research" (862).

To be clear, here we broach no complaints about the authors' survey instrument, analysis of data, or conclusions. We chose this study because it was among the *most* diligent in its discussion of potential survey-design issues, and because it is highly cited. The original study used an OM design, measuring the outcome variable first and then the mediators. Recognizing that this choice could potentially affect results, the authors took the rare, costly, and admirable step to replicate their survey using an MO design in which they instead measured mediators immediately after treatment, before measuring the outcome (Tomz and Weeks, 2013: fn. 13). They fielded that survey to an additional 797 US respondents less than a year after the original study and found that the overall treatment effect of democracy was similar, with democracy lowering support for a strike by 11.7%. That the point estimate of the main treatment effect was so similar is reassuring, but it leaves open the possibility that their other principal conclusions, concerning mediation, might be affected by survey design.

3 Replication Results

To replicate the study in Tomz and Weeks (2013) ("TW" hereafter), we reconstructed the survey instrument, altering question order selectively to create both OM and MO versions of the survey.[4] We then fielded the survey using an online sample of 1,041 respondents recruited though Amazon's Mechanical Turk (AMT) from June 5–7, 2018. Respondents were randomly assigned to either the OM or MO setup and then randomly assigned to one of the two treatment groups, democracy or non-democracy. Respondents completed the survey with a median time of approximately six minutes and received $1.00 in compensation.

TW also included a variety of other control variables describing the respondents' characteristics (e.g. gender, race, education) and opinions (e.g. their preferences for internationalism or militarism). We measured and included these, as well, using the same wording and operationalizations.

Our survey differed from that of TW in two ways. First, we held fixed the value of TW's other treatments, which randomized features of the foreign country's alliances and military strength. Second, TW used both an across- and within-subject design depending on the wave of their survey. For the within-subject design, respondents read two scenarios about a foreign country, one a democracy and the other a non-democracy, several days apart. Our design used only an across-subject design, partly for simplicity and partly because the majority of survey experiments are not multi-wave panels. We do not expect these differences, either alone or in concert, to induce or increase a difference between OM and MO designs.

3.1 Effect on Outcome and Mediator Distributions

Our respondents were generally less supportive of an attack on the hypothetical nuclearizing nation than were the respondents in TW's US studies. Our estimated treatment effect (the gap between

[4] We maintained a constant order *within* the battery of questions gauging the potential mediators, eschewing, for now, exploration of possible order effects between mediators.

support for attacking a democracy and a non-democracy) was similar to that found in TW.

	TW,UK.OM	TW,US.OM	TW,US.MO	OM	MO
democracy	20.9	41.9		17.7	19.3
not democracy	34.2	53.3		26.5	28.6
difference	-13.3	-11.4	-11.7	-8.8	-9.3
95% CI	(-7.0, -19.6)	(-5.9, -16.8)		(-1.7, -16.0)	(-2.0, -16.7)
N.dem	364	639		260	254
N.not	398	634		260	262
N	762	1273	797	520	516

Table 1: Democracy's Effect on Favorability to Attack

To construct the "threats" mediating variable, respondents were instructed to select how many of a list of six events would "have more than a 50 percent change of happening If the U.S. did not attack the country's nuclear sites"; e.g. one event was that "the country would build nuclear weapons and threaten to use them against another country."

Table 2 shows that, as compared to TW, somewhat fewer of our respondents perceived most of the threats stemming from non-attack as likely. In general, respondents perceived democracies as less threatening than non-democracies. For individual items, the effect of democracy on perceived threat varied across OM versus MO designs. However, when summing the number of threats, the total-threats counts for the two groups were very similar. A chi-squared test supports pooling ($p = 0.89$), with about 10 percent seeing no threats as more likely than not, then 37, 23, 15, 5, 7 and 3 percent judging as plausible 1 through 6 of these threats, respectively.

TW measure respondent perception of the morality of a strike by asking "Do you think it would be morally wrong for the US military to attack the country's nuclear weapons sites?" When their respondents contemplated a democracy, about 38% said yes; when the country was not a democracy, the value fell to 31%. Our respondents were more likely to view an attack as immoral, with 65% and 52% saying that attacking a non-democracy or a democracy, respectively, would be immoral.

	TW.dem	TW.not	OM.dem	OM.not	MO.dem	MO.not
build	72	75	88	85	78	87
threaten...	38	52	41	54	35	55
threaten US	34	45	19	24	15	28
attack...	26	34	22	20	18	26
attack US	24	30	12	14	10	18
lose prestige			15	17	15	18
none			8	9	14	8
N	972	972	246	239	233	243

Table 2: Threat Levels, Democracy Status, and Survey Design

The OM/MO design appears to have mattered for these levels, as there was a statistically significant 6-percentage-point gap between the proportions declaring attack immoral (61% for MO versus 55% for OM). In addition, the OM/MO design affected the relationship between democracy and perceived morality of an attack. In the OM design, democracy increased the perceived immorality of an attack by 10%, from 50% to 60%. In the MO design, the effect of democracy was stronger, increasing the perceived immorality of an attack by 15%, from 54% to 69%.

For the costs associated with an attack, TW asked which among a set of four possible effects had a greater than 50% chance of happening, in the event of an attack, e.g. "the U.S. military would suffer many casualties." The likelihood of success was measured with two items asking whether, following an attack, there would be a greater than 50% chance that this attack "would prevent the country from making nuclear weapons in the near future" and "... in the long run." Table 3 shows the percentage of respondents in a particular study and treatment condition that indicated that a particular cost was likely. For the costs mediator, our OM and MO respondents differed very little from one another in these distributions.

3.2 Effect on Estimated Treatment Effects

Following TW, Table 4 shows results from probit models with attitude towards attacking the hypothetical nation dichotomized into support (strong or not) versus opposition (strong or not) or

	TW.dem	TW.not	OM.dem	OM.not	MO.dem	MO.not
retaliate	39	39	61	53	58	60
military casualties	33	32	53	47	51	51
economy suffer	31	31	45	39	47	44
relations suffer	53	49	66	63	65	64
near-term deterence	61	66	72	72	72	79
long-term deterence	25	30	30	35	34	37
none			6	4	5	2
N	972	972	256	247	245	250

Table 3: Cost and Success Levels, Democracy Status, and Survey Design

neutrality. Our results resemble those of TW, broadly, with a few discrepancies. Our comparatively small Ns, given the between-subject design, conspire against statistical significance in the coefficient capturing the direct treatment (democracy) effect for the distinct modules. The very similar coefficients for the democracy indicator for the OM and MO modules convert into fairly similar marginal probability shifts, -4.4 percent for the MO data and -5.8 percent for the OM, computed using mean values of all other variables. Not surprisingly, then, the impact is smaller than the simple difference of proportions displayed in Table 1. Exactly one control variable had a statistically significant impact, by the $p < 0.05$ criterion, for each data set, but which one mattered differed. The mediators all had statistically significant coefficients, in expected directions and roughly in the vicinity of those in the estimation reported in Tomz and Weeks (2013).

3.3 Effect on Mediation Analysis

To assess mediation, we matched our analysis as closely as possible to that of TW and employed the most commonly used statistical packages. For each mediator, we evaluate the percentage of the democracy treatment effect that travels through the relevant mediation channel by simulation, from two models. Here, we employ the dichotomous model of attack favorability, broached above, that includes all mediators as predictors. We then estimate a model of each mediator variable, all with the same covariates as the attack model, except that we omit the other mediators. We performed

	TW (OM)	pooled (OM,MO)	OM	MO
treatments				
democracy	**-0.18**	**-0.27**	-0.30	-0.27
ally	-0.06			
trade	-0.05			
mediators				
threats	**0.30**	**0.34**	**0.40**	**0.31**
costs	**-0.21**	**-0.21**	**-0.27**	**-0.18**
success	**0.23**	**0.34**	**0.31**	**0.36**
immorality	**-1.12**	**-1.15**	**-0.93**	**-1.57**
controls				
militarism	-0.02	0.12	0.12	0.04
internationalism	0.02	**0.17**	-0.02	**0.44**
Republicanism	0.10	-0.02	0.10	-0.03
ethnocentrism	0.09	0.09	-0.09	0.23
religiosity	-0.03	**0.29**	**0.39**	0.10
male	0.05	0.06	-0.03	0.22
white	-0.18	0.07	0.10	0.03
age	**0.01**	-0.01	-0.01	0.00
education	-0.06	-0.00	-0.12	0.12
intercept	-0.30	-0.55	-0.08	**-1.06**
N	972×2	939	475	464

Table 4: probit coefficients, dependent variable is an indicator for supporting (strongly or not) military strike, rather than opposing or declining to support or oppose; bold indicates $p \leq 0.05$; TW estimates based on within variance for 972 respondents (two responses each, one for democracy and one for non-democracy), replication estimates are from between variance (one response each).

the calculations to decompose the democracy effect into a direct effect and mediated effects, we used the `mediate` function from Tingley et al.'s `mediation` package for R.

Table 5 reveals significant differences between the OM and MO estimates. The proportion of the democracy effect attributed to each mediator varies notably across designs. In the OM design, the threats mediator accounts for 16% of the democracy effect, compared to 41% in the MO design. The immorality mediator accounts for 26% of the democracy effect in the OM design, but this quantity almost doubles to 53% in the MO design. In contrast, the relative importance of the costs mediator was higher in the OM design, 13%, compared to 0% in the MO design.

	average effect via...	% of total effect of democracy
Tomz and Weeks (OM)		
threats	**-4.0**	34
costs	-0.4	4
success	**-0.7**	6
immorality	**-1.7**	15
OM		
threats	-1.1	16
costs	-0.9	13
success	0.5	6
immorality	**-2.1**	26
MO		
threats	**-3.2**	41
costs	0.0	0
success	-0.5	7
immorality	**-5.1**	52

Table 5: Results of Mediation Analysis by Survey Design

The rankings of relative importance for each mediator was largely unchanged across OM and MO designs, with immorality and threats being the first and second strongest in both.

With no clear, sharp threshold for what constitutes an important mediator, researchers could draw different conclusions between OM and MO designs from the results reported in 5. Roughly speaking, the MO module points to two strong channels (the same ones identified by Tomz and Weeks) while the other suggests one medium-strength channel and two weaker channels. Note that in the Tomz and Weeks article, immorality accounted for 15% of the treatment effect, which led them to conclude that it was an important, under-studied channel for explaining the democratic peace. In these results, the costs channel reaches a similar threshold in the OM design, 13%, but does not reach that threshold in the MO design.

The results also differ substantially in the degree to which the mediators together explain the democracy effect. In the OM design, the four mediators explain 61% of the democracy effect, implying that 39% resides in other channels. In the MO design, two mediators explain almost the entirety of the democracy effect, leaving little room for explanatory power from unexplored

channels.

4 Discussion

In Zaller's highly influential theory of attitude formation and expression, survey responses reflect a mix of considerations that are "immediately salient or accessible" (Zaller (1992): 49). These considerations are the atoms of opinion, and everybody is understood to hold a wide variety of them, that need not cohere, since they vary from being dormant to being at the top-of-the-head. In turn, opinions are not so much fixed points as distributions—a given respondent who chooses, say, 5 on a 7-point scale is not fully revealing his true ideal point, but, rather, providing partial information about his range of possible answers, conditional on a host of short-term factors. One such factor of relevance is what questions this individual has recently been asked and answered.

Variables that seem likely mediators between treatments and outcomes are, necessarily, at least potentially connected to both. Judgments about morality of a hypothetical attack or about costs and benefits of such an attack are likely to reflect many of the same considerations that are averaged together in a complicated, unobserved mental process to crystallize into a favorability answer. So, in Zaller's terms, it should not be surprising if planting a morality or costs frame around the hypothetical scenario alters the favorability that would be expressed absent such queries, for at least some respondents.

Practically speaking, this means that survey design decisions in causal mediation analysis that appear innocuous may, in fact, affect the inferences drawn. We have highlighted several ways that the decision of whether to use an MO or OM survey design can affect responses to mediator and outcome items as well as their relationship to a randomly assigned treatment. We have shown how these effects can change the substantive conclusions that a researcher might draw about the relative weight to assign to different channels or mechanisms of mediation.

We make two recommendations. The first is for greater transparency about the survey designs

employed. Researchers should always specify what type of design they used. Moreover, they should use knowledge of the context of the research to hypothesize about likely effects of MO versus OM design choices on results. A researcher might argue that an MO design is justified because she believes that the mediator questions are unlikely to frame or influence responses to the outcome item. She might also use her knowledge of the overall survey design, arguing that an OM design is justified because the survey includes a large number of mediator items which could influence outcome responses or attenuate treatment effects. Describing, explaining, and defending design choices would be welcome.

Second, we recommend that researchers directly assess whether findings are conditional on the sequence of survey questions, by randomly assigning respondents to an OM or MO version of a particular experiment. Since CMA applications draw important substantive inferences from the proportion of a treatment effect that flows through particular mediator(s), a concrete assessment of whether survey design affects those proportions is likely to be important. This is especially true given that the theoretical ways that OM versus MO decisions can influence results are complex and potentially unknown to the researcher.

Of course, there is some cost to this approach, but we think this cost is justified. If analyses from the OM and MO modules agree, this validates pooling data from both designs, minimizing any loss of power. If analyses do not agree, however, the researcher has valuable information about the complexity of the associations between variables that can help inform interpretation of results and further inquiry.

5 Appendix

In Table 4, we reported models of a collapsed, dichotomous version of favorability towards attacking the hypothetical nuclearizing nation, following the precedent from Tomz and Weeks (2013). Below are alternative models: an ordered probit fit to the five-category favorability response, not collapsed; and an OLS model on that same variable, assuming intervality. In both cases, we obtain statistical significance at the conventional $p < 0.05$ level for the democracy treatment, in contrast to the probit model reported in 4. These estimates also exhibit further heterogeneity in the behavior of control variables.

	OM	MO
treatments		
democracy	**0.29**	**0.36**
mediators		
threats	**-0.36**	**- 0.31**
costs	**0.16**	**0.22**
success	**-0.24**	**- 0.36**
immorality	**0.79**	**1.26**
controls		
militarism	0.01	**0.24**
internationalism	-0.01	**-0.18**
Republicanism	-0.10	-0.06
ethnocentrism	-0.01	-0.18
religiosity	**-0.24**	-0.12
male	0.16	**-0.22**
white	0.18	-0.09
age	0.00	0.01
education	0.12	-0.05
intercept.1|2	**-1.86**	**-2.61**
intercept.2|3	-0.33	**-0.67**
intercept.3|4	0.44	0.19
intercept.4|5	**1.75**	**1.41**
N	475	468

Table 6: Notes: ordered probit coefficients, dependent variable is five-category support for military strike (strong support, support, unsure, opposition, strong opposition); bold indicates $p \leq 0.05$.

	OM	MO
treatments		
democracy	**0.22**	**0.22**
mediators		
threats	**-0.29**	**- 0.23**
costs	**0.13**	**0.15**
success	**-0.19**	**- 0.24**
immorality	**0.65**	**1.01**
controls		
militarism	0.01	**0.16**
internationalism	0.00	**-0.12**
Republicanism	-0.07	-0.03
ethnocentrism	-0.00	**-0.14**
religiosity	**-0.20**	-0.09
male	0.12	-0.14
white	0.11	-0.04
age	0.00	0.00
education	**0.10**	-0.02
intercept	**2.97**	**3.19**
N	475	468
$adj.R^2$	0.45	0.56
RSE	0.85	0.78

Table 7: Notes: ordinary least squares coefficients, dependent variable is five-category support for military strike, treated as interval-valued (1. strong support, 2. support, 3. unsure, 4. opposition, 5. strong opposition) ; bold indicates $p \leq 0.05$.

References

Baker, Andy. 2015. "Race, Paternalism, and Foreign Aid: Evidence from US Public Opinion." *American Political Science Review* 109(1):93–109.

Gaines, Brian J., James H. Kuklinski and Paul J. Quirk. 2007. "The Logic of the Survey Experiment Reexamined." *Political Analysis* 15(1):1–20.

Holbrook, Allyson L., Jon A. Krosnick, David Moore and Roger Tourangeau. 2007. "Response Order Effects in Dichotomous Categorical Questions Presented Orally: The Impact of Question and Respondent Attributes." *Public Opinion Quarterly* 71(3):325–348.

Huddleston, R. Joseph. forthcoming. "Think Ahead: Cost Discounting and External Validity in Foreign Policy Survey Experiments." *Journal of Experimental Political Science* .

Huddleston, R. Joseph and Nicholas Weller. 2017. "Unintended Causal Pathways: Probing Experimental Mechanisms Through Mediation Analysis.".

Imai, Kosuke, Luke Keele, Dustin Tingley and Teppei Yamamoto. 2011. "Unpacking the Black Box of Causality: Learning about Causal Mechanisms from Experimental and Observational Studies." *American Political Science Review* 105(04):765–789.

Lupu, Noam. 2013. "Party Brands and Partisanship: Theory with Evidence from a Survey Experiment in Argentina." *American Journal of Political Science* 57(1):49–64.

Sniderman, Paul M. and Edward G. Carmines. 1997. *Reaching Beyond Race*. Harvard University Press.

Sudman, Seymour and Norman M. Bradburn. 1974. *Response Effects in Surveys: A Review and Synthesis*. Aldine/NORC.

Tomz, Michael R. and Jessica L.P. Weeks. 2013. "Public Opinion and the Democratic Peace." *American Political Science Review* 107(4):849–865.

Tourangeau, Roger, Lance J. Rips and Kenneth Rasinski. 2000. *The Psychology of Survey Response*. Cambridge University Press.

Transue, John E., Daniel J. Lee and John H. Aldrich. 2009. "Treatment Spillover Effects Across Survey Experiments." *Political Analysis* 17(2):143–161.

Zaller, John R. 1992. *The Nature and Origins of Public Opinion*. Cambridge University Press.