Being Careful with Conjoints: Accounting for Inattentiveness in Conjoint Experiments

John V. Kane and Mia Costa

October 15, 2024

Abstract

In typical survey experiments–i.e., experiments that involve a relatively limited amount of manipulated content–respondent inattentiveness tends to bias treatment effect estimates toward zero. Such bias may be even more pronounced in conjoint experiments, which require respondents to attend to an even larger amount of manipulated content. And yet, little research has investigated strategies to account for inattentiveness in conjoint experiments specifically. In this study, we explore potential ways to both measure–and account for–respondent inattentiveness when estimating causal effects in both single- and two-profile conjoint experiments. Replicating published conjoint experiments with large national samples, we demonstrate how researchers can implement a simple strategy using pre-treatment measures of attentiveness. Toward this end, we propose a novel method– "conjoint attention checks" (CACs)–to both measure respondents' attentiveness to conjoint profiles and to provide for more robust tests of hypotheses in conjoint experiments.

Author Affiliations. John V. Kane is Associate Professor at the Center for Global Affairs, New York University; john.kane@nyu.edu. Mia Costa is Assistant Professor of Government, Dartmouth College; mia.costa@dartmouth.edu.

Survey-based online experiments have become increasingly popular in the social sciences (see e.g. Thomas, 2024). The flexibility and scalability of online platforms make it possible to administer complex experimental designs, efficiently gather large amounts of data from diverse populations, and provide insights into a wide range of topics. In short, the methodological shift to online surveys has opened new avenues for investigation into causal relationships and the effects of various treatments on respondent attitudes and behaviors (Druckman, 2022; Sniderman, 2018).

Alongside these advantages, the widespread adoption of online experiments has also introduced significant challenges. Chief among these is the concern that respondents may be less motivated to pay close attention than in more controlled, in-person environments. Whether due to distractions or a desire to complete the survey quickly for payment, respondents often do not fully engage with survey items. As a form of "satisficing"—i.e., a behavior whereby respondents give suboptimal answers to finish faster— inattentiveness reduces the quality of responses and adds statistical noise (Berinsky, Margolis and Sances, 2016; Krosnick, 1991; Silber, Danner and Rammstedt, 2019). Perhaps even more importantly, such behavior often biases the observed strength of experimental effects toward zero, making it more difficult to draw reliable conclusions from one's results (Bailey, 2021, 145-146).

In response, scholars have developed several strategies to identify and account for inattentive respondents in surveys (Berinsky et al., 2024). For example, instructional manipulation checks (IMCs) include hidden instructions in questions to provide a specific response (Berinsky, Margolis and Sances, 2014; Oppenheimer, Meyvis and Davidenko, 2009). Factual manipulation checks (FMCs) ask respondents questions to verify specific elements of an experimental vignette experiment (Kane and Barabas, 2019). Mock vignette checks (MVCs) contain information in a format similar to standard experimental vignettes, and then ask respondents factual questions to assess their attentiveness to the content (Kane, Velez and Barabas, 2023).

At the same time, conjoint experiments have rapidly gained traction in the social sciences, particularly within political science, due to their ability to randomize multiple factors simultaneously (see Bansak, Hainmueller, Hopkins, Yamamoto, Druckman and Green, 2021; Hainmueller, Hopkins and Yamamoto, 2014). Unlike traditional survey experiments, where respondents are typically asked to evaluate a single scenario (i.e., vignette) with a limited amount of manipulated content, conjoint experiments involve presenting respondents with one or two profiles, each comprising lists of multiple attributes that vary randomly. Respondents are then asked to make choices, or rate preferences, based on these profiles–a task that is repeated multiple times for each respondent.

How can researchers be more confident that respondents are attending to all of this information? While standard attention check measures are effective for traditional survey questions or vignette experiments, such measures may not be well-suited for gauging attentiveness to conjoint experiments. In traditional vignette experiments, the focus is typically on a single dimension or a limited set of factors, often presented in the form of a narrative vignette (e.g., a mock news article), making it potentially easier for respondents to remain engaged. In contrast, conjoint experiments require respondents to process and evaluate a large amount of manipulated content, presented in the form of one or more columns of attributes/attribute-levels, which may lead to higher levels of inattentiveness and more measurement error.

In short, given their particularly large quantity of manipulated content, conjoint experiments may be especially vulnerable to the aforementioned consequences of respondent inattentiveness. Moreover, respondents who are motivated to attend to a narrative-style vignette (as in a typical survey experiment) may not be the same respondents who are motivated to attend to an anodyne table of attributes and attribute levels. And yet, guidance for how to specifically deal with inattentiveness in conjoint experiments remains limited.

Building on existing research, we offer a simple approach for accounting for inattentiveness in conjoint experiments. This approach relies upon the use of pre-treatment measures of attentiveness. It can easily be incorporated into one's survey, can be applied to both single-and two-profile conjoint designs, avoids the risk of inducing post-treatment bias (Montgomery, Nyhan and Torres, 2018), and allows for straightforward analyses and hypothesis testing. Within this general approach, we explore several specific measures that can be used to account for pre-treatment inattentiveness in conjoints, including an original attention measure we call a "conjoint attention check" (CAC). This technique involves asking factual questions specifically about conjoint content, enabling researchers to better gauge how well respondents are attending to the conjoint experiment itself.

We test the performance of CACs, as well as other pre-treatment attention checks, and demonstrate how they can be implemented using both a single- and two-profile conjoint design. Consistent with previous research, such checks reveal that attentive respondents tend to show substantially larger effects than inattentive ones and, thus, that failing to account for inattentiveness will tend to yield smaller average marginal component effects (AMCEs) in conjoint experiments. Overall, we find that CACs, compared to other checks, may be particularly well-suited for identifying inattentiveness in both single- and two-profile conjoint experiments, potentially offering a more reliable and informative estimates of treatment effects.

That said, we caution that CACs are not intended to be a substitute for estimating AMCEs for the sample as a whole. Rather, we regard the use of CACs as a kind of robustness test, enabling conjoint experimentalists to more robustly test their hypotheses and, ultimately, learn more from their studies.

Measuring Attentiveness in Surveys: Common Approaches

There are several methods to measure and account for inattentiveness in surveys (see Berinsky et al., 2024, for a review). Indirect measures rely on behavioral indicators, such as page timers, where unusually fast responses may suggest that a respondent is rushing through the survey (Wood et al., 2017). "Bogus items" ask respondents to agree or disagree with obvious, factual statements (e.g., asking whether one agrees that "water is wet") (Gummer, Roßmann and Silber, 2021; Meade and Craig, 2012).

Other direct measures instruct respondents to engage in a specific way. For example, instructed response items (IRIs) explicitly tell respondents to select a particular answer within a grid of items (Gummer, Roßmann and Silber, 2021). Instructional manipulation checks (IMCs) embed such instructions within an ostensible survey question, asking respondents to ignore the main content and instead follow a specific instruction (Oppenheimer, Meyvis and Davidenko, 2009). While these measures are widely used for detecting whether respondents are reading instructions carefully, they do not measure attentiveness to experiments in particular (i.e., the content of IRIs and IMCs is unrelated to the experiment).

As for measuring attention in survey experiments, there are several options. Factual manipulation checks (FMCs) assess whether respondents have understood and retained key information from an experiment (Kane and Barabas, 2019). These checks involve asking objective questions about specific elements of the experimental content. However, these items are, by definition, post-treatment and should not be used to re-estimate experimental effects (Montgomery et al. 2018; Varaine 2023).

To avoid post-treatment bias, a mock vignette check (MVC) can be used to assess attention to an experimental vignette by presenting a vignette similar to the actual experiment and then asking factual questions about it (Kane et al. 2023). For example, after reading a mock vignette about a policy proposal, respondents might be asked to recall specific details about the proposal. This approach thus avoids post-treatment bias while also capturing whether respondents are paying attention to content that is stylistically similar to a vignette-based experiment.

Yet while these latter approaches are highly applicable to typical survey experiments, for the reasons discussed above, they may not be well-suited for measuring attentiveness to a conjoint experiments. Indeed, little work has attempted to develop measures of attentiveness to conjoints specifically. A notable exception is that of Clayton et al. (2023). These authors make a significant contribution by introducing Intra-Respondent Reliability (IRR) as a method to measure the consistency of responses specifically in conjoint experiments. In this approach, a two-profile task is repeated again (in reverse order) as the last task, and the proportion of agreement between these two identical tasks is calculated. This therefore provides researchers with a binary indicator of whether or not a given respondent chose the same profile in the final task as in the first and, ultimately, an estimate of the overall sample's attentiveness that can be used to re-estimate results. That said, this approach is exclusive to two-profile, forced-choice conjoint designs. Additionally, it requires using (in effect) a post-treatment measure of attentiveness, and is somewhat complex in terms of significance testing.

An Additional Approach: Conjoint Attention Checks

How else might researchers account for attentiveness in conjoint experiments? To first understand the extent to which researchers attempt to address inattentiveness in conjoint experiments, we reviewed 101 articles that feature conjoints. These articles were published in eight high-ranking journals in political science (*American Political Science Review*, *American Journal of Political Science*, *Journal of Politics*, *British Journal of Political Science*, *Political Behavior*, *Public Opinion Quarterly*, *Political Science Research and Methods*, and *Journal of Experimental Political Science*).

We find that only 13% of make explicit mention of any attempt to measure or address inattentiveness. Given the degree to which inattentiveness can bias experimental studies toward substantively smaller and non-significant effect sizes, this pattern suggests that conjoint experimentalists may not yet have sufficient strategies for addressing inattentive responses. Again, this may be due to conjoint experiments' large amount of manipulated content. And while some research has found only limited degradation of response quality as the number of attributes or tasks increases (Bansak et al., 2018; Bansak, Hainmueller, Hopkins and

Yamamoto, 2021), this does not address respondent inattentiveness that exists before the conjoint experiments even begins. In short, a respondent may be inattentive to one's conjoint experiment regardless of the number of attributes and tasks. This possibility highlights the need for more targeted methods to measure and mitigate the effects of inattentiveness in conjoint experiments.

As noted above, Clayton et al. (2023) offer one approach, specifically for two-profile, forced-choice designs. Calculating the IRR enables scholars to assess the extent of measurement error, which Clayton et al. (2023) note has been largely overlooked in conjoint analysis. The method provides a clear indicator of the degree to which respondents' choices are reliable. That said, the IRR is not necessarily meant to capture respondent "attentiveness" *per se*, though implementing their correction substantially increases the precision of AMCEs. To be sure, the goal of improving precision often comes with the trade off of sacrificing ecological validity and how people make decisions in the real world (often, inconsistently).

On one hand, the MVC approach proposed by Kane, Velez and Barabas (2023) might be sufficient for capturing attentiveness immediately before a conjoint experiment begins, and should substantially condition treatment effects just as they do in conventional survey experiments. On the other hand, compared to a typical survey experiment vignette, conjoint experiments differ in two key respects. First, compared to the simpler presentation of information in a survey experiment (e.g., a paragraph of text written in a narrative style), a conjoint experiment presents in an arguably less engaging format–i.e., as one (or multiple) column(s) of attributes and their respective levels. Second, conjoint experiments often contain a larger quantity of (treatment-relevant) information. Whereas a survey experiment may manipulate only a single sentence, or perhaps even a single word, virtually *all* content in a conjoint experiment is manipulated. By virtue of its simpler presentation style and quantity of content, a mock vignette may therefore not be capable of adequately identifying respondents who do, and do not, attend to a conjoint experimental design.

An improved approach, we reason, would be to ask respondents factual questions about what is seen in the conjoint vignette itself. There are two important considerations, however. First, conjoint experiments involve *multiple* tasks per respondent, meaning that a question designed to gauge respondent i's attentiveness must be asked about a single conjoint task, not after multiple tasks, as the correct response will vary by task.

Second, as noted above, measured attention to manipulated content is, by definition, post-treatment, and therefore risks inducing post-treatment bias if the measure is included in the analysis (Montgomery, Nyhan and Torres, 2018). [1]

To overcome these two considerations, we propose that researchers instead ask factual questions *after a first task that does not vary across respondents.* Specifically, after the first measure of the outcome (a forced choice, a rating scale, etc.), respondents are asked three closed-ended questions about three attributes that appeared in the first task (e.g., a target's race/ethnicity, gender, and education level).[2] This first task is identical for all respondents, meaning that the correct answer for each of these *conjoint attention checks* (CACs) is the same for each respondent. A correct answer to each CAC would be indicative of sufficient attentiveness to the conjoint-experimental content and, in that way, closely resembles a (treatment-relevant) factual manipulation check (Kane and Barabas, 2019). Then, similar to the MVC method, the three binary indicators of whether a respondent answered each CAC correctly are combined to create an additive scale that, subsequently, can be used to (1) subset the data on the most versus least attentive respondents (e.g., respondents who answered two or more CACs correctly versus one or zero CACs correctly) to estimate AMCEs for each group, and/or (2) interact with any predictor variables of interest.

We use *three* questions to balance between reliability and respondent burden. Asking

[1]Relative to traditional survey experiments, the risk of post-treatment bias may be lower in conjoint experiments insofar as respondents will view *multiple* levels of each attribute. Nevertheless, respondents will not receive identical combinations of attributes, nor can researchers rule out the possibility that the *order* in which different attribute levels were seen affected respondents' ability to answer a factual question about the conjoint content. For these reasons, placing an attention check before any randomized content helps ensure, even in a conjoint experiment, that analyses are free of post-treatment bias.

[2]In addition to reducing measurement error, the use of three CACs effectively reduces the probability of attaining a maximum attentiveness score (3) by random chance to $<2\%$.

more than three questions increases both the time required to complete the survey and the cognitive load on respondents, which can lead to greater fatigue and reduced data quality. On the other hand, asking fewer than three questions risks compromising the reliability of the attention scale. Given the number of attributes typically included in conjoint designs, using only one or two questions is insufficient to accurately gauge attentiveness across the profile.

It is worth emphasizing the benefits of fixing the first conjoint vignette to be *static* across respondents. First, this helps to prevent the possibility that the same CAC is more difficult for some respondents by virtue of having seen a different profile (or set of profiles). For example, if Respondent A views two immigrants (Immigrant 1 and Immigrant 2) that have the *same* level of education, a CAC about the relative education levels shown in the profiles might be easier to answer than for Respondent B, who saw one immigrant with a higher education than the other immigrant. Second, because the first task does not randomly vary between respondents, the CACs are, by definition, not post-treatment. However, as the first task contains the same attributes that will be featured in subsequent tasks, this first task can still be included in the analysis to maximize statistical power. Put differently, even though each respondent views the same set of attribute-levels in the first task, these are legitimate attribute-levels that will randomly vary, both between and within respondents, in subsequent tasks. For any given respondent, then, the response on the first outcome measure(s) is just as valid as if attribute-levels had been randomly assigned in the first task.

When selecting which questions to include as CACs, we focus on simple, easily distinguishable attributes that appear across the entire conjoint profile. This means selecting one attribute from near the top of the table, one from the middle, and one from near the bottom, ensuring that the CACs measure whether respondents are paying attention throughout the task. This approach is consistent with findings that respondents tend to focus on certain parts of conjoint tables while selectively ignoring others to reduce cognitive processing costs Jenke et al. (2021). Researchers may also wish to avoid asking about attributes that are central to the study's main hypotheses, as this could inadvertently prime respondents to place more emphasis on those attributes. That said, we test for such priming effects and find no substantial evidence that CACs cause respondents to prioritize certain attributes over others (see Appendix E).

See Table 1 for a summary on how we implemented the CACs in the studies discussed below.

Compared to the Clayton et al. (2023) approach noted above, the CAC technique has several notable strengths. First, the CAC technique can just as easily be applied to single-profile conjoint experiments as it can two-profile conjoint experiments, and can be used even if researchers do not require respondents to make a forced choice. Second, the CAC technique is perhaps simpler to implement, particularly in terms of estimating quantities for hypothesis-testing purposes (e.g., standard errors and p-values). Third, the CAC technique does not include a post-treatment measure. [3] Fourth, the CAC technique does not assume that inattentiveness will be equally consequential for all attribute levels–i.e., it allows inattentiveness to matter for some attributes more than others.[4]

Given conjoints' large quantity of manipulated information and lack of narrative style, though, it is possible that even attentive respondents may struggle to answer CACs correctly. This could lead to more measurement error than the MVC approach, for example, and thus a weaker ability to distinguish effects among attentive vis-a-vis inattentive respondents. But as this is ultimately an empirical question, we can investigate it by examining how CACs perform relative to MVCs, as well as to the common practice of using a (pre-treatment) IMC.

[3] Per Clayton et al., the first and final tasks are used to estimate intra-respondent reliability (IRR), which is then used to re-estimate AMCEs.

[4] The logic here is that, for attributes (and attribute-levels) that truly have no effect, the degree to which a respondent is attentive to it should–in expectation–matter little for the estimated AMCE (i.e., it should be approximately zero regardless). However, for attributes that, in reality, *do* have a large effect, attentiveness to it should matter a great deal for the estimated AMCE (i.e., the AMCE's size will depend much upon a respondent's attentiveness to it). Thus, we see good reason to allow inattentiveness to heterogeneously impact AMCEs.

Table 1: Details Regarding Implementation of Conjoint Attention Checks (CACs)

Constructing CACs	
Basic Implementation	A table of attributes takes the same form as tables in subsequent tasks. Crucially, the attribute levels do not vary between respondents. (The table can be a static image or programmed to show a particular, non-varying level for each attribute.)
Attention Questions	This first table is followed by the outcome measure(s) and then, on a separate screen, three CACs. The first CAC asks about an attribute featured toward the top; the second about an attribute featured in the middle; the third about an attribute featured toward the bottom (we did not randomize the order of CACs). In a one-profile design, CACs ask about attribute levels; in a two-profile design, CACs ask about which profile had a particular attribute level (for categorical attributes) or had the higher/lower attribute level (for ordinal/continuous attributes). One CAC always asks about an attribute that was not featured. The questions are written to be as simple as possible given that a respondent has attended to the profiles.
Response Options	Each CAC has four response options. In a one-profile design, each option refers to a possible attribute level of the profile viewed. In a two-profile design, each option refers to a profile, explicitly specifying the profile "on the left" or "on the right." A third response option is that the two profiles had equal levels. In both the one- and two-profile designs, the final option is always that the attribute was not featured. (The response options were not randomized.)
Using CACs in Analyses	
Combining CACs	The three CACs are first coded as either correct or incorrect. These three binary variables are then combined into a single additive scale ranging from 0 to 3.
Estimating AMCEs	After presenting AMCE results for the sample as a whole, researchers can estimate AMCEs (1) at high versus low levels of attentiveness (e.g., 2 or more correct CACs versus 1 or zero correct CACs), or (2) via interactions between attribute levels and the continuous CAC scale. Substantially larger AMCEs at higher levels of attentiveness provide additional evidence for a non-zero effect, whereas similarly-sized AMCEs across varying levels of attentiveness provide additional evidence for a negligible and/or non-significant effect.

Note: This table displays the guidelines followed in the present study when constructing and implementing conjoint attention checks (CACs). Researchers may of course choose to modify these principles for their own studies (e.g., feature additional CACs), though it is important to be mindful of potential trade-offs (e.g., additional questions causing greater respondent fatigue).

Another concern is whether CACs might induce greater attentiveness in subsequent tasks, as respondents may think they will be asked about the tasks again. This could lead to a type of demand effect where respondents become more attentive due to the checks themselves. If

CACs were causing this, we would expect to see little difference in treatment effects between high- and low-performing respondents on the CACs. However, as we show below, this is not the case. We observe significant variation in the experiment outcomes based on CAC performance, indicating that the checks measure attentiveness rather than induce it.

Ultimately, each method has different trade-offs but all can potentially be effective strategies to account for inattentiveness. Indeed, as we demonstrate below, the methods are *not* mutually exclusive—researchers could employ several to understand how inattentiveness impacts treatment effects and report the results from each approach. This strategy can offer a more comprehensive understanding of how attention influences experimental outcomes, as well as how efficacious particular attributes may be on a given outcome.

In the following section, we provide two applications using common conjoint designs first outlined in (Hainmueller, Hopkins and Yamamoto, 2014). We find substantial variability in attentiveness among respondents, and that performance on CACs strongly correlates with performance on other attentiveness measures (MVCs, IMCs, and IRR). We also find substantial differences in AMCE magnitude when comparing respondents who performed relatively well on CACs (2 or more correct) versus those who performed relatively poorly (1 or 0 correct). More broadly, we find that accounting for attentiveness using any pre-treatment measure tends to yield larger AMCEs, but CACs in particular perform slightly better than these other measures. Specifically, CACs yield larger conditional average treatment effects (CATEs) across conjoint attributes in both applications, and also produce the most substantial improvements in model fit when interactions between attributes and attentiveness are specified.

Applications

Immigration Forced-Choice Conjoint

We fielded our first study in the Fall of 2023 online via Lucid. As with other Lucid surveys, the sample is diverse in terms of U.S. geographic region, race/ethnicity, gender, and age. A total of 2,094 respondents completed the conjoint portion of our study, which replicated the Hainmueller and Hopkins (2015) two-profile immigration experiment. Importantly, respondents who did not (1) consent to participate, and (2) correctly answer a "CAPTCHA" question, were not permitted to proceed. Thus, any inattentiveness found in our study may be an *underestimate* of the true level of inattentiveness among Lucid respondents (Stagnaro et al., 2024).

This experiment involved presenting respondents with a series of pairs of (hypothetical) immigrants who desire to "move to the United States." For each immigrant, the same set of nine attributes is listed, though the specific level shown (for each attribute) is randomly assigned. Table 2 displays the attributes as well as each of the levels the attribute can take on.

After being provided with details about both immigrants in the pair, respondents were asked to choose "which of the two immigrants [they] would personally prefer to see admitted to the United States." This constitutes our *Choice* outcome, which (per common practice) is modeled using OLS with standard errors clustered at the respondent-level.[5]

Importantly, the study featured several measures of attentiveness. Prior to the conjoint experiment, respondents were asked three mock vignette checks (MVCs; Kane, Velez and Barabas, 2023). Respondents were also asked an instructional manipulation check (IMC, or "screener"; Berinsky, Margolis and Sances, 2014; Oppenheimer, Meyvis and Davidenko,

[5]Following Hainmueller and Hopkins (2015), respondents were also asked to indicate the degree to which they would be willing to admit each immigrant. This *Rating* outcome was measured on a seven-point scale ranging from "Absolutely not admit" (1) to "Definitely admit" (7). Analyses of this outcome are featured in the Supplemental Appendix.

Table 2: Attributes for Immigration Experiment

Attributes	Levels
Prior Trips to the U.S.	Never been to the U.S. Entered the U.S. once before on a tourist visa Entered the U.S. once before without legal authorization Has visited the U.S. many times before on tourist visas Spent six months with family members in the U.S.
Reason for Application	Reunite with family members already in U.S. Seek better job in U.S. Escape political/religious persecution
Country of Origin	Germany; France; Mexico; Philippines; Poland; India; China; Sudan; Somalia; Iraq
Language Skills	During admission interview, this applicant spoke fluent English During admission interview, this applicant spoke broken English; During admission interview, this applicant tried to speak English but was unable During admission interview, this applicant spoke through an interpreter
Profession	Gardener; Waiter; Nurse; Teacher; Child care provider; Janitor; Construction worker; Financial analyst; Research scientist; Doctor; Computer programmer
Job Experience	No job training or prior experience One to two years Three to five years More than five years
Employment Plans	Has a contract with a U.S. employer Does not have a contract with a U.S. employer, but has done job interviews Will look for work after arriving in the U.S. Has no plans to look for work at this time
Education Level	No formal education Equivalent to completing fourth grade in the U.S. Equivalent to completing eighth grade in the U.S. Equivalent to completing high school in the U.S. Equivalent to completing two years at college in the U.S. Equivalent to completing a college degree in the U.S. Equivalent to completing a graduate degree in the U.S.
Gender	Female; Male

Note: Each level was randomized across tasks for each respondent to generate the conjoint profiles. This experiment is based on Hainmueller and Hopkins (2015).

2009). Both of these types of items have only one correct response: respondents who answer them correctly are identified as attentive, while those who answer incorrectly are identified as not attentive. The three MVCs were combined into an additive scale that ranged from 0 to 3, while the IMC is coded as incorrect (0) or correct (1).

Our design also employed Clayton et al.'s (2023) technique to calculate the sample's intra-respondent reliability (IRR). Respondents were shown a total of seven pairs of immigrants. However, the survey was programmed such that the first and seventh pairs of immigrants were identical but for a reversal in ordering (the immigrant that appeared on the left, in the first pairing, was shown on the right in the seventh pairing, and vice versa). This feature allows us to not only calculate the IRR for the sample as a whole but also use whether the respondent correctly switched their choice (yes=1, no=0) as a binary *individual-level* measure of attentiveness to investigate the predictive validity of the CAC approach.[6]

Finally, we also employed the conjoint attention check (CAC) approach. To reiterate, this method involved using a static—i.e., non-varying—table of attributes for the first pair of immigrants, such that all respondents saw the exact same table. Following the *Choice* and *Rating* measures, respondents were asked three factual questions—conjoint attention checks (CACs)—about the pair of immigrants they saw in the table.

To standardize the presentation and difficulty of the CACs, each CAC asked a question about a specific attribute and always included (essentially) the same four response options [7] The three CACs, and their respective response options, were presented in a fixed order, with the correct answer to one CAC being "Immigrant 1," the correct answer to another CAC being "Immigrant 2," and the correct answer to the third CAC being the fourth choice (i.e., that the attribute was not mentioned).

Great care was taken to help ensure that the questions would be as simple as possible given that a respondent attended to the attributes. Toward this end, CACs asked questions for which the contrast between profiles was especially stark. For example, the two profiles differed greatly with respect to educational attainment, with one immigrant having attained

[6]Clayton et al. only employ the IRR/swapping error as a sample-level measure, not an individual-level measure.

[7]The use of four options (rather than two or three) is to further reduce the probability that a respondent could answer correctly by chance to just 25%. With three CACs, therefore, answering all three CACs correctly by chance alone falls to less than 2 %.

a "fourth grade" education while the other immigrant completed a college degree. One CAC therefore asked, "From the previous screen, which immigrant had more education?" Respondents could then choose one of four response options: (1) Immigrant 1 (the one shown on the left); (2) Immigrant 2 (the one shown on the right); (3) The two immigrants had an equal amount of education; and, (4) Education was not mentioned.

Because the seventh pair was primarily used for calculating the IRR, respondents' answers on the final task were excluded from the analyses discussed below. This results in 2,127 respondents evaluating up to twelve unique immigrant profiles, yielding a total n of 24,910.

Candidate Rating-Based Conjoint

We fielded a second study in July of 2024 also via Lucid that included the same consent and CAPTCHA questions at the start of the survey (again, these likely screened out a substantial share of inattentive respondents (Stagnaro et al., 2024)). The experiment, which was also modeled from (Hainmueller, Hopkins and Yamamoto, 2014) –included total of 1,046 respondents. With six tasks, the experiment yielded a total of 6,260 observations.

In contrast to the two-profile *Immigrant* experiment above, this experiment involved single profiles of potential political candidates. Respondents were presented with a list of attributes about candidates running for the U.S. House of Representatives.[8] These attributes, and their respective levels, are shown in Table 3. Notably, we included one additional attribute beyond those featured in the original study: support from political Independents.[9]

After viewing each profile, respondents were asked to rate the candidate in terms of how likely the respondent would be to "vote for the candidate in an election." This *Rating* outcome was measured on a seven-point scale ranging from (1) "Very unlikely to vote for them" to (7) "Very likely to vote for them." This rating was then rescaled to range from 0

[8]Note that (Hainmueller, Hopkins and Yamamoto, 2014) present the candidates as competing in a presidential election. Given the prevalence of conjoint experiments on congressional elections, we chose to present respondents with hypothetical House candidates.

[9]Also in contrast to the original study, we also model the candidate's age as continuous, rather than categorical, given that the levels are in roughly equal intervals and for ease of exposition.

Table 3: Attributes for Candidate Experiment

Attributes	Levels
Religion	Catholic; Evangelical Protestant; Mainline Protestant; Mormon; Jewish; None
Profession	Lawyer; High school teacher; Business owner; Farmer; Doctor; Car dealer; Council member
Age	Random integer between 36 and 75
Annual Income	$32,000; $54,000; $65,000; $92,000; $210,000; $5,100,000
Race/Ethnicity	Hispanic; White; Black; Asian American; Native American
Gender	Man; Woman
Military Service	Served; Did not serve
College Education	No BA; BA from a community college; BA from a Baptist college; BA from an Ivy League college; BA from a state university; BA from a small college
Support Among Independents	20-25%; 50-55%; 75-80%

Note: Each level was randomized across tasks for each respondent to generate the conjoint profiles. This experiment is based on Hainmueller, Hopkins and Yamamoto (2014), with one exception: Support Among Independents, which was an original attribute for this study and did not appear in Hainmueller, Hopkins and Yamamoto (2014).

to 1, and modeled using OLS with standard errors clustered at the respondent level.

To measure attentiveness, we used the exact same MVC and IMC items that were featured in the previous study. As in the previous study, these items appeared *prior to* the conjoint experiment.

We again employed the CAC approach. This involved showing all respondents the same candidate profile in the first task and then, following the *Rating* measure, asking respondents factual questions about this profile before proceeding to the second task. For reasons explained below, respondents saw one of two (randomly selected) sets of CACs, each of which contained three questions. Respondents assigned to the first set of CACs were asked to identify the candidate's: (1) religion, (2) gender, and (3) political party. Respondents assigned to the second set were asked to identify the candidate's: (1) profession, (2) level of education, and (3) political party.

Again, the CAC questions were designed to be simple given that respondents attended to the attributes. For example, language used in the response options was identical to the language used in the single conjoint table. Further, we ensured that correct CAC responses were qualitatively different from the incorrect response options. For example, the candidate shown in the static conjoint table attained a "BA from a small college." Therefore, only one CAC response option mentioned a "BA"–other response options featured qualitatively distinct degrees. As in the previous study, each CAC question had four response options, one of which was always that the item was "not mentioned." In this case, our final CAC question was about the candidate's "political party" which did not actually appear in the conjoint table. The CAC text, and list of response options for each CAC, are included in the Supplemental Appendix.

Similar to the *Immigrant* experiment, the profile that appeared before the CACs was repeated again as the last task. Since there are no two profiles to "swap" in this study, the first and last tasks were simply identical. This allows for calculating a modified version of the IRR: a respondent demonstrates reliability if they give the same rating for the first profile as they did for the last profile. We emphasize, however, that this is a modified version of the IRR since the original authors propose this technique only for a two-profile design that uses a forced-choice outcome rather than a scale. Nevertheless, we believe this modified IRR still represents a reasonable metric by which to assess respondents' attentiveness to the conjoint experiment insofar as inattentive respondents (compared to attentive ones) should be at least somewhat less likely to give the same rating to identical profiles.

We again exclude the final tasks from the analysis on the grounds that the final task simply repeated the first task as a means of calculating the IRR.

Results

Investigating the Validity of the CAC Approach

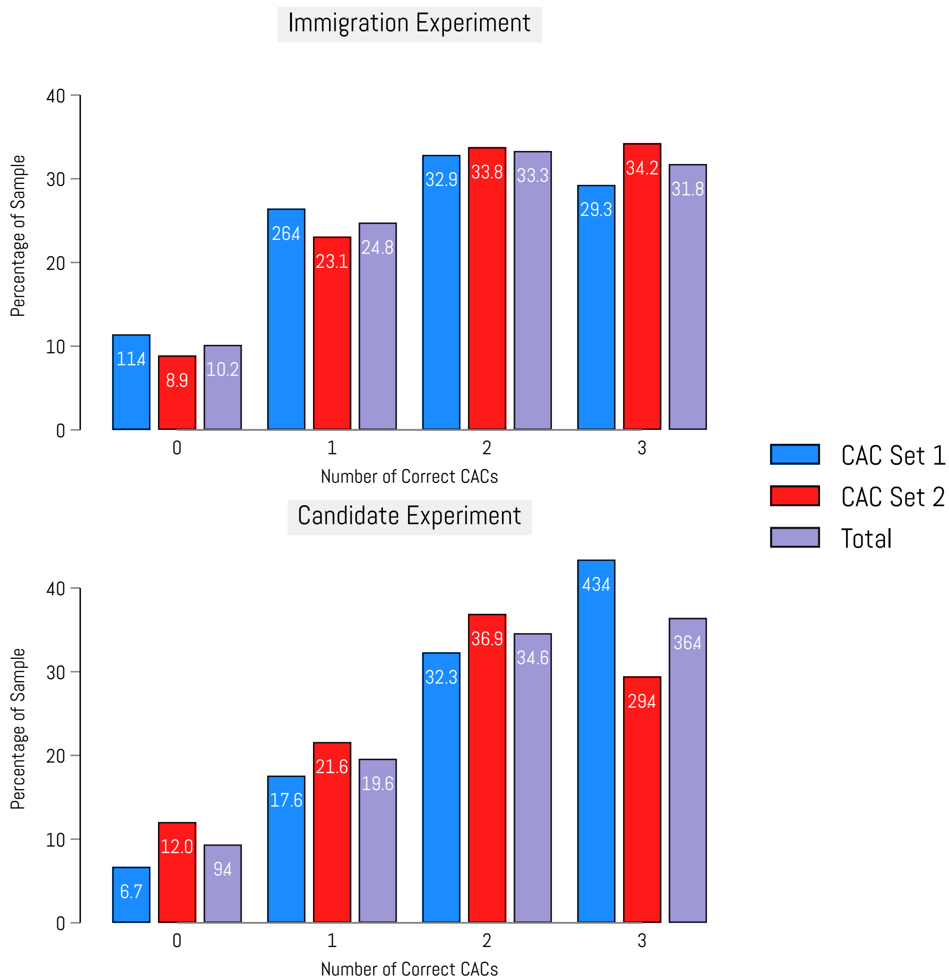
A reasonable initial concern is that answering factual questions (i.e., the CACs) about one or two columns' worth of information may be overly challenging for most respondents. As such, it is crucial to first investigate respondents' performance on CACs. For reasons discussed below, we randomly assigned respondents–in both experiments–to answer one of two possible sets of CACs. Each CAC set contained three questions, with the third question in each set being identical. Figure 1 therefore presents, for each CAC set as well as for the sample as a whole (see "Total" category), the percentage of the sample that answered zero, one, two, or all three CACs correctly. The top panel of Figure 1 displays the results for the *Immigration* experiment, while the bottom panel displays the results for the *Candidate* experiment.

Beginning with the top panel of Figure 1, several patterns are worth highlighting. First, the "Total" performance is encouraging, with nearly one-third of the sample answering all three CACs correctly. Only 10% of the sample answered all three CACs incorrectly, while 25% answered only one CAC correctly. Such figures are quite comparable to performance on other attentiveness measures, such as factual manipulation checks, screeners, and MVCs (e.g. Berinsky, Margolis and Sances, 2014; Kane and Barabas, 2019).

Second, each individual CAC was answered correctly at rates that are significantly better than chance (in all cases, $p < .001$). Additionally, we see no discernible tendency for respondents to answer the "Age" CAC substantially better or worse than other CACs, which is instructive given that no information about the immigrants' ages was provided. In short, this result suggests that respondents were not simply assuming that the answer must be one of the two immigrants and, therefore, randomly selecting between the first two response options.[10]

[10]Had this been the case, then the expected probability of answering a CAC correctly would be far closer to 50% than 25%.

Figure 1: Performance on Conjoint Attention Checks (CACs)



Note: Percentage of respondents who answered zero, one, two, or all three CACs correctly in both the Immigration Experiment (top panel) and Candidate Experiment (bottom panel), categorized by which CAC set they saw and for the entire sample ("Total").

The results for the single-profile *Candidate* study (shown in the bottom panel of Figure 1) are broadly similar to the results for the *Immigration* study. First, large majorities answered at least two of the three CACs correctly. In fact, performance is slightly higher in the *Candidate* study than in the *Immigration* study, potentially because a single-profile conjoint contains less information to attend to. Second, respondents again answered each CAC significantly better than chance alone (p<.001).

Overall, Figure 1 provides initial confirmation that CACs are not an overly difficult measure of attentiveness in conjoint experiments and that, as with other measures of attentiveness, respondents exhibit a substantial amount of variation in their ability to correctly answer them.

We next investigate whether CACs are predictive of other measures of attentiveness that are used in survey experiments. To simplify presentation, Figure 2 plots performance on the three-item CAC scale against performance on the three-item MVC scale, an IMC, and intra-respondent reliability (IRR). The MVC models are OLS. The IMC and IRR models are logistic given that a respondent either answered the IMC correctly or not (both studies), either changed their choice of immigrant between the first and last (reversed) task or not (IRR in *Immigrant* experiment), and either retained their candidate rating between the first and last (identical) task or not (IRR in *Candidate* experiment). The top panel of the figure displays the results for the *Immigrant* experiment, while the bottom panel displays the results for the *Candidate* experiment.

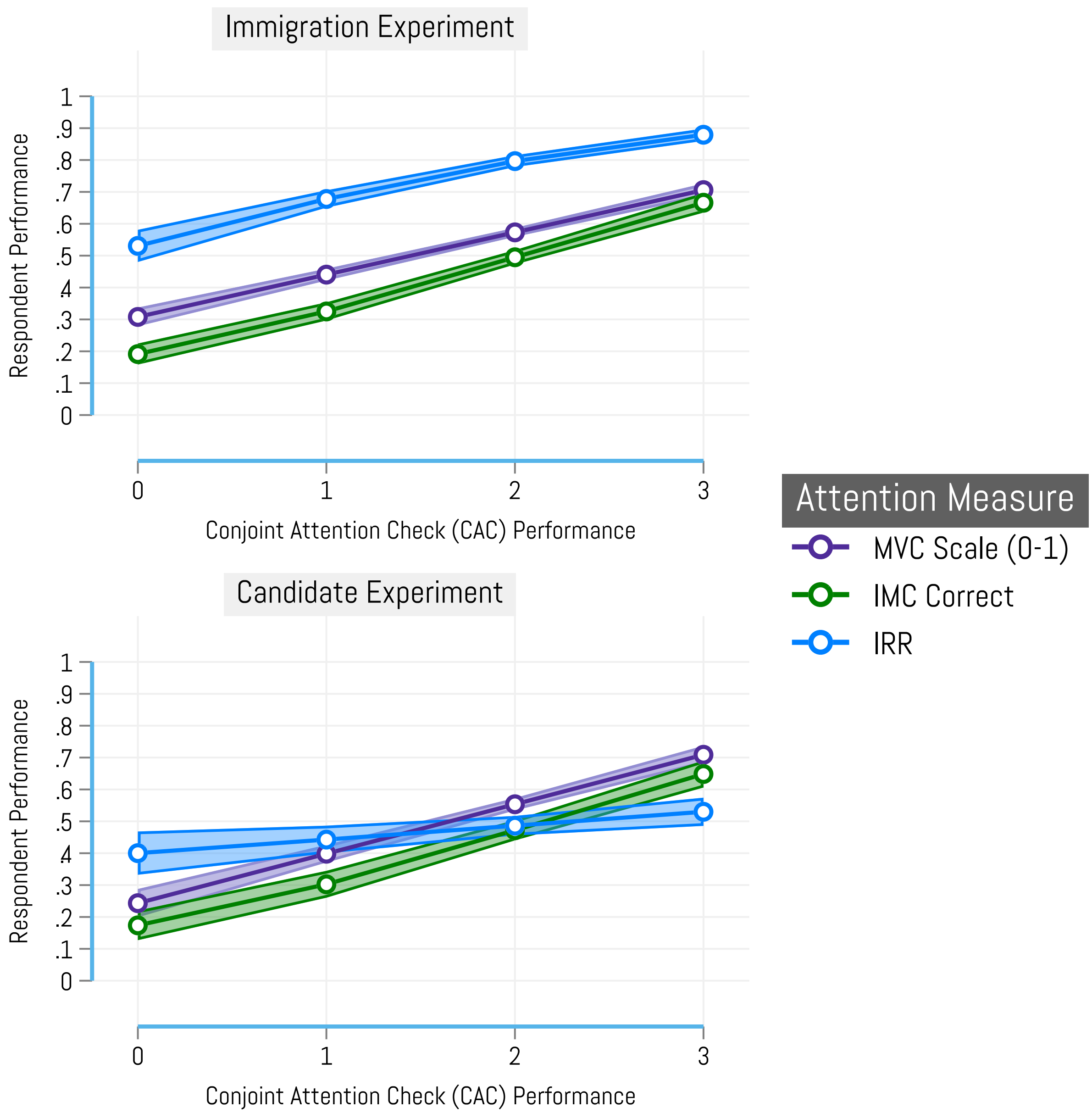
Overall, Figure 2 shows that respondents' performance on CACs is strongly associated with other measures of attentiveness. Beginning with the *Immigrant* experiment (see top panel), moving from lowest to highest CAC performance predicts a 40 percentage-point increase in MVC performance, a 47 percentage-point increase in the likelihood of answering the IMC correctly, and a 35 percentage-point increase in the likelihood of switching choices between the first and last tasks (p<.001 in all cases).[11]

The bottom panel of Figure 2 displays the results for the *Candidate* experiment. Again, as performance on the CAC scale increases, performance on other measures of attentiveness significantly increase (p<.01 in all cases). As we move from lowest to highest CAC performance, predicted performance on the MVC scale increases from .24 to .71, while the

[11] Regarding the IMC, only 47% of the sample answered it correctly. Regarding the IRR, 66.95% chose the other profile on the (reversed) last task. Thus, the Clayton et al. "intra-respondent reliability (IRR)" = .6695. And, therefore, τ is equal to:

$$\frac{1-\sqrt{1-2\cdot(1-0.6695)}}{2} = 0.20888147 \tag{1}$$

Figure 2: Predictive Validity of Conjoint Attention Checks (CACs)



Note: Performance on the three-item CAC scale in the Immigrant Experiment (top panel) and Candidate Experiment (bottom panel) against performance on the three-item Mock Vignette Check (MVC) scale, the Instructional Manipulation Check (IMC), and Intra-Respondent Reliability (IRR). MVC models estimated using ordinary least squares (OLS); IMC and IRR models estimated using logistic regression.

predicted probability of correctly answering the IMC increases from .17 to .65. More modestly, the predicted probability of giving the same rating on the first and last profile (IRR), which were identical, increases from .40 to .53.[12]

In summary, Figures 1 and 2 confirm several key points. First, there is substantial variability in respondent attentiveness to conjoint experiments, with a non-negligible share of respondents appearing to be inattentive to key details regarding the target(s) featured. This presents a serious challenge to researchers insofar as inattentiveness will likely attenuate AMCEs (just as it attenuates treatment effects in typical survey experiments (Bailey, 2021; Kane, 2024). Second, a majority of respondents are capable of answering at least two out of three CACs correctly, indicating that these items are not overly challenging and that overall, respondents can generally recall much of the information presented in a conjoint. Third, performance on CACs is strongly associated with performance on a variety of other attentiveness measures. This serves as a key test of CACs' validity, demonstrating that CACs are indeed capturing respondent attentiveness, but with the added benefit that they measure attentiveness to the conjoint experiment specifically.

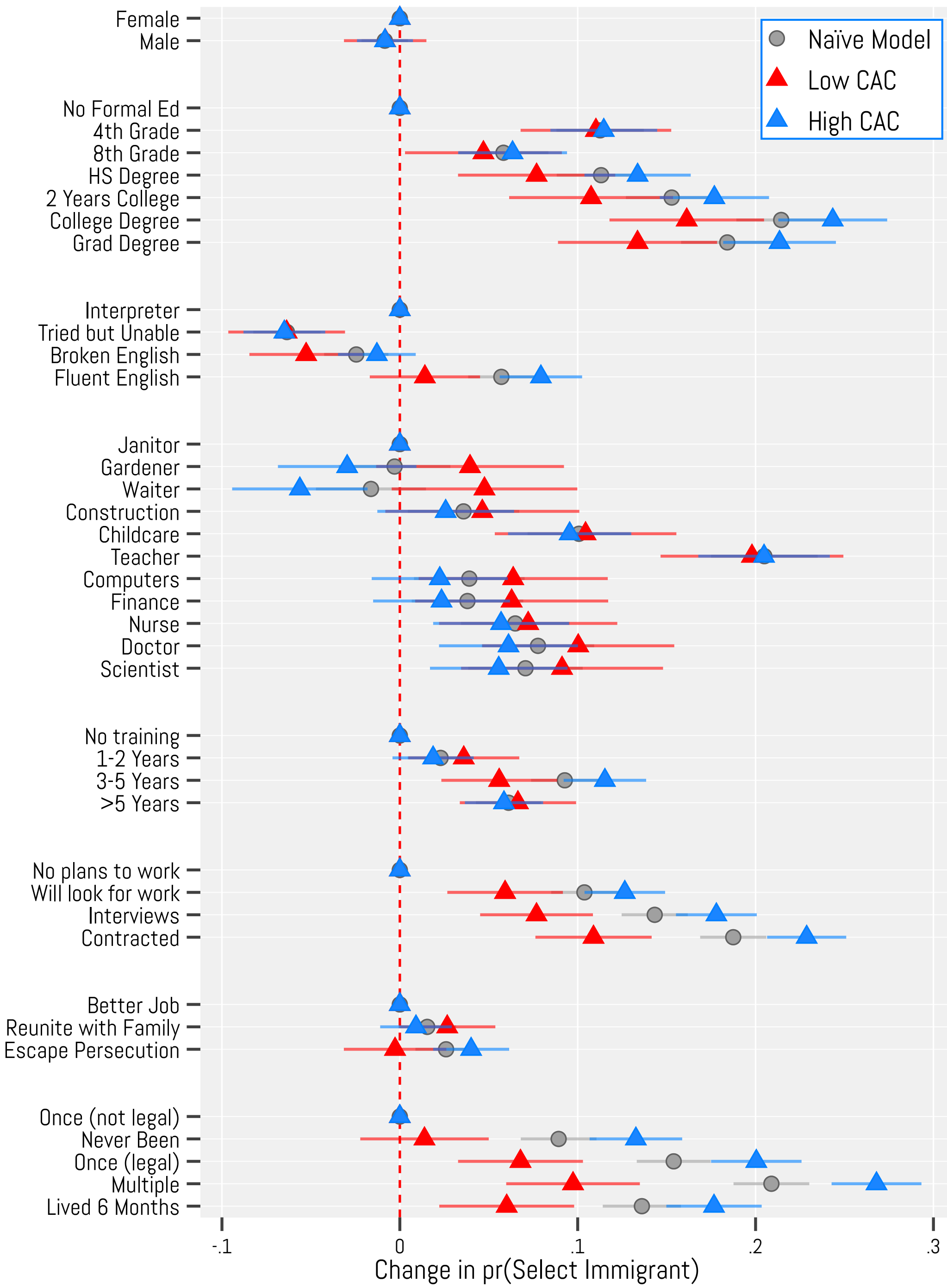
We next explore how CACs perform in terms of their ability to distinguish average marginal component effects (AMCEs) among less attentive respondents (which should tend to be smaller vis-à-vis more attentive respondents (Kane, Velez and Barabas, 2023)).

CAC Performance & Changes in AMCE Estimates

The key question of interest in this section is whether AMCEs are substantively different among those with higher, versus lower, performance on CACs, particularly for the attribute-levels that do exhibit a meaningful effect on the outcome.

[12]This may be partly due to some inattentive respondents giving the same rating for *every* task. Indeed, we find that 10% of respondents had zero variation in their ratings across all seven tasks. Again, given the single-profile nature of this experiment, this is an alternative measure of the IRR than the one proposed by Clayton et al. (2023).

Figure 3: AMCEs at Low & High CAC Performance in Immigration Experiment



Notes: The figure displays AMCEs from a naïve model that does not account for attentiveness, as well as AMCEs for respondents with low performance on the CACs (<=1 correct CAC) and high perfomance on the CACs (>=2 correct CACs). Each AMCE indicates the estimated change in probability of an immigrant being selected. The underlying model includes all attributes (with standard errors clustered by respondent) but the figure excludes the constant and the targets' country of origin to conserve space. The underlying model clusters standard errors by respondent). Total n=24,910 (2,127 individual respondents). Data from Qualtrics.

Importantly, for any attributes that are unimportant/irrelevant to respondents, we would expect smaller CATE estimates. Conversely, for attributes that do exert meaningful effects, we should observe larger CATEs insofar as attentive respondents are likely to exhibit increasingly large effects relative to inattentive respondents.

Figure 3 displays the results of the *Immigrant* experiment. The figure features results from three separate models: (1) a "naïve" model that does not account for attentiveness (gray, circular point estimates); (2) a "Low" model featuring respondents who answered only 1 or fewer CACs correctly (red, triangular point estimates); and, (3) a "High" model featuring respondents who answered 2 or more CACs correctly (blue, triangular point estimates). The "Low" model includes 35% of respondents, while the "High" model includes the other 65%.[13]

Looking at the naïve model results, we observe many of the same effects as found in the original Hainmueller et al. study. For example, an immigrant's education, ability to speak English, and job plans can all matter a great deal for being selected.

That said, a crucial pattern revealed here is that many of the largest AMCEs from the naïve model are (1) substantially smaller in magnitude for respondents with the lowest performance on the CACs, and (2) larger for the respondents with the highest performance on the CACs.

This pattern is precisely what we should expect if CACs are indeed capturing attentiveness, and is perfectly consistent with survey-experimental research that shows weaker treatment effects among less (versus more) attentive respondents (Bailey, 2021; Kane, Velez and Barabas, 2023)

For example, the estimated effect of an immigrant being able to speak fluent English (compared to requiring an interpreter) in the naïve model is 5.7 percentage points (p<.001). However, by accounting for attentiveness, we observe that the estimated AMCE is 7.9 percentage

[13]In this as well as in the subsequent study, we chose to define "Low" and "High" in this way insofar as (1) it evenly divides the 0-3 CAC performance scale, and (2) the resulting groups were both of a substantial size. That said, researchers might choose to use an alternative grouping strategy if, for example, CAC performance is less evenly distributed.

points among those who were highly attentive to the conjoint experiment (p<.001)—a sizable increase, and even stronger evidence that immigrants' language ability is consequential for immigration attitudes. Conversely, the estimated AMCE is only 1.4 percentage points (p=.37) for respondents who exhibited low attentiveness to the conjoint experiment.

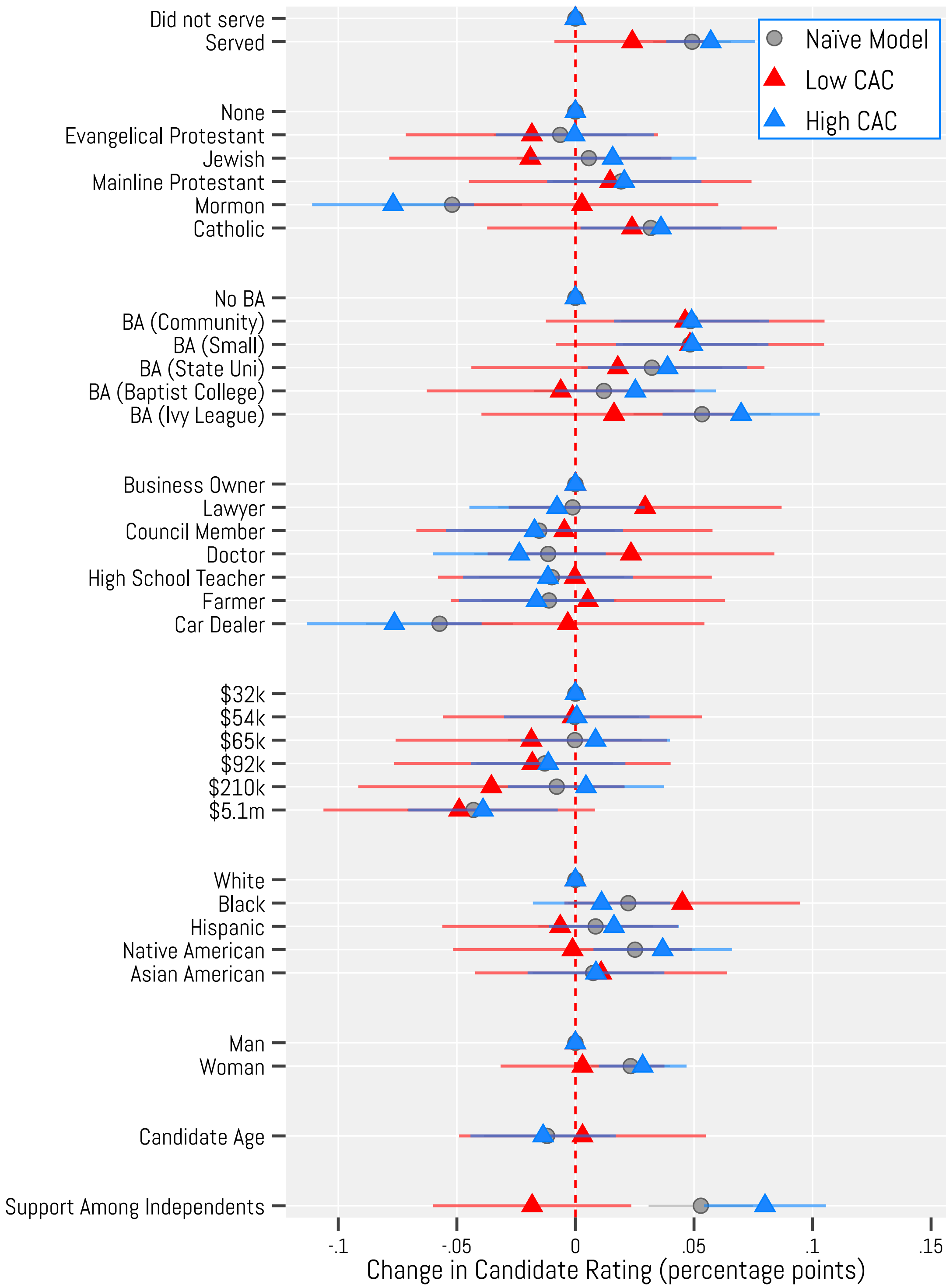
Similar patterns exist for other large effects found in the original study. For example, a college degree (compared to "No Formal Education") yields an AMCE that is 18 percentage points in the naïve model, but is 24 (16) percentage points among the highly (low) attentive; being contracted to work (versus having no plans for work) yields an AMCE that is 19 percentage points in the naïve model, but is 23 (11) percentage points among the highly (low) attentive; reason for emigrating is to escape persecution (compared to looking for a better job) yields an AMCE that is 3 percentage points in the naïve model, but is 4 (0) percentage points among the highly (low) attentive; and, experience living the U.S. for six months (compared to visiting once illegally) yields an AMCE that is 14 percentage points in the naïve model, yet is 18 (6) percentage points among the highly (low) attentive.

In the Supplemental Appendix, we feature the same set of analyses for the seven-point *Rating* outcome that was also included in this *Immigrant* experiment. The pattern is virtually identical, with highly attentive respondents tending to exhibit larger AMCEs than in the naïve movel, and the naïve model, in turn, tending to exhibit larger AMCEs than the low-attention respondents. The *Immigration* experiment therefore provides strong evidence for the utility of CACs. Specifically, the figure demonstrates how accounting for inattentiveness via CACs can assist researchers with identifying AMCEs among those who are sufficiently attentive to their two-profile conjoint design. For the attributes that exhibited the largest AMCEs–especially immigrants' education, work plans, and time spent in/visits to the US–we indeed find that the more (versus less) attentive exhibit even stronger effects than in the naïve model.

The same set of analyses were conducted for the single-profile *Candidate* experiment and are shown in Figure 4. First, while some differences exist with the original study, many of the most notable effects in the naïve model remain similar. For example, the naïve model finds that a candidate who has (versus has not) served in the military yields an AMCE of 5 percentage points; being a Mormon candidate (versus no religious identification) yields an AMCE of approximately -5 percentage points; a college degree (compared to no degree) yields positive AMCEs (e.g., 5 percentage-points in the case of a B.A. from an Ivy League University); and, being a "Car Dealer" (versus "Business Owner") yields an AMCE of -6 percentage points. We also see that going from low to high support among political Independents yields an AMCE of 5 percentage points.

But for any given attribute, does performance on CACs tend to correspond with substantially different AMCEs? As with the previous experiment, the AMCEs tend to be stronger (weaker) in magnitude among the more (less) attentive respondents. For example, the AMCE for "Car Dealer" (as the candidate's profession): highly attentive respondents exhibit an AMCE larger than the AMCE in the naïve model (-8 percentage points (p<.001)), while low-attention respondents exhibit an AMCE of only -0.3 percentage points (p=.91). Thus, inattentiveness again appears to be downwardly biasing the AMCE in the naïve model. By accounting for attentiveness, the *High* model exhibits even stronger evidence that a candidate's profession is consequential for their electoral prospects.

Figure 4: Naïve Model and CATEs Using CACs in Candidate Experiment



Notes: The figure displays AMCEs from a naïve model that does not account for attentiveness, as well as AMCEs for respondents with low performance on the CACs (<=1 correct CAC) and high perfomance on the CACs (>=2 correct CACs). Each AMCE indicates the estimated change in the target candidate's rating (measured on a seven-point scale, with higher values indicating higher favorability). The underlying model clusters standard errors by respondent). Total n=6,260 (1,046 individual respondents). Data from Lucid Theorem.

Similar patterns can be seen for other attributes. For example, the effect of military service is estimated to be 6 (2) percentage points in the *High* (*Low*) model; the effect of being a Mormon candidate is estimated to be -8 (0) percentage points in the *High* (*Low*) model; the effect of having a B.A. from an Ivy League University is (also) estimated to be 7 (2) percentage points in the *High* (*Low*) model ; and, the effect of support from political Independents is estimated to be 8 (-2) percentage points in the *High* (*Low*) model. Again, these all represent substantial differences in AMCE sizes compared to the results of the naïve model, further confirming the importance of accounting for respondent attentiveness in conjoint experiments (whether single-profile or two-profile).[14]

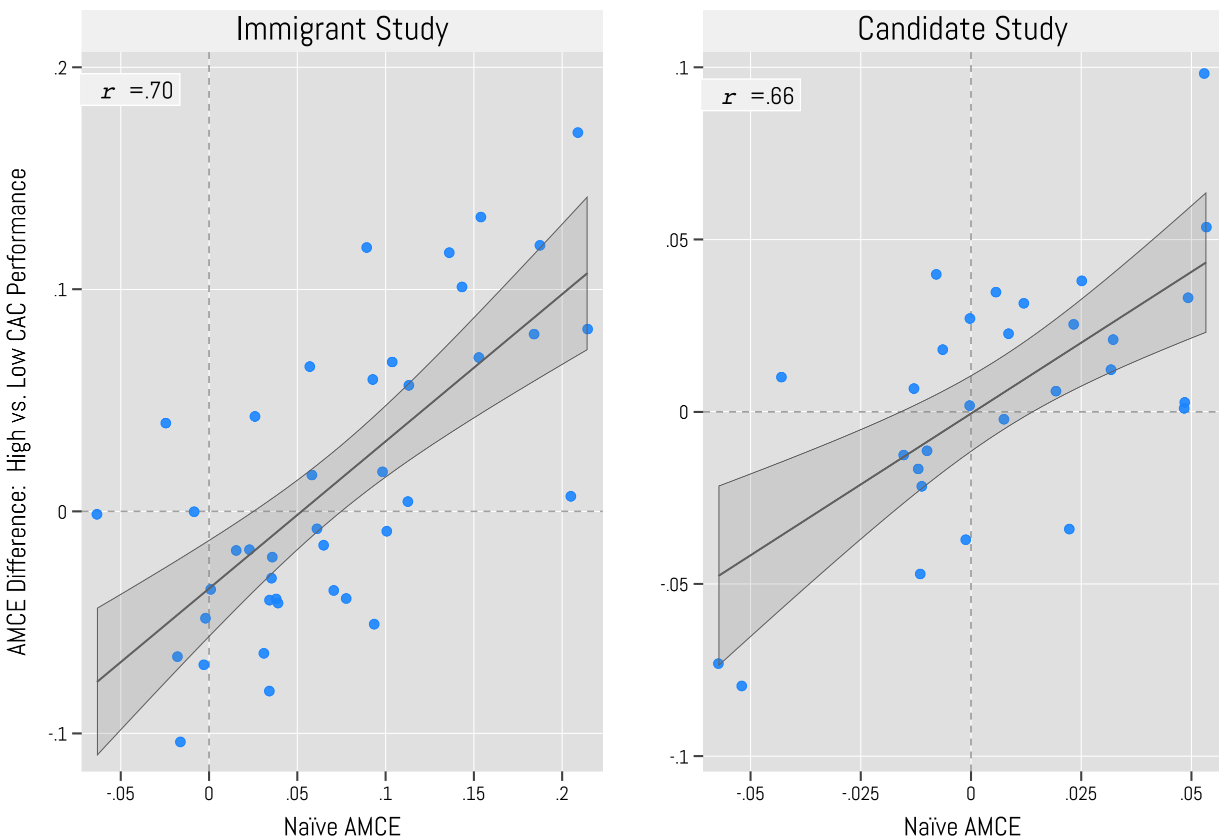
Because it is only is subtly revealed in the preceding figures, a final point worth emphasizing is that respondent inattentiveness should increasingly (decreasingly) matter for AMCE estimates as an attribute-level exerts a larger (smaller) effect. If an attribute-level would exert a near-zero effect even in a fully attentive sample, in other words, then high- and low-attention respondents should exhibit similar AMCEs. But if an attribute-level would exert a strong effect in a fully attentive sample, we should observe even larger differences between high- and low-attention respondents in the present sample.

Figure 5 demonstrates exactly this point, with results from the *Immigrant* study shown in the left panel and results from the *Candidate* study shown in the right panel. Each observation represents one attribute-level from the preceding figures. For each attribute-level, the y-axis shows the total difference between "High" and "Low" AMCE estimate, while the x-axis shows the estimated AMCE for this attribute-level in the naïve model. Despite the two experiments being different in a variety of ways (e.g., the number of profiles, the topic, and

[14]In the Supplemental Appendix, we analyze the data from this and the previous experiment in an alternative fashion–i.e., specifying interactions between each attribute-level and the CAC Performance scale. While both analytical strategies are perfectly valid, we opted to separate respondents into "High" and "Low" groups here for ease of exposition. The interaction-based approach, alternatively, avoids having to group respondents together based upon some level of CAC performance. Yet it should be noted that specifying an interaction for each attribute level will yield relatively larger standard errors for each conditional AMCE estimate. Nevertheless, in terms of the point estimates themselves, the patterns observed in those analyses are perfectly consistent with what we present here.

the CAC items themselves), the results are remarkably consistent: as the naïve AMCE gets larger in magnitude, so, too, does the *difference* in AMCE for the high- versus low-attentive respondents. The observed correlation (r) for the *Immigrant* (*Candidate*) experiment is .70 (.66) (p<.001 in both cases). These patterns not only provide added validity for the CAC approach, but also imply that using a measure of attentiveness is increasingly valuable to researchers as the absolute size of a given AMCE grows larger.

Figure 5: Larger Naïve AMCEs Associated with Larger High-vs.-Low AMCE Differences



Note: The figure displays each attribute-level from both the *Immigrant* (left) and *Candidate* (right) conjoint experiments. The x-axis shows the AMCE for each attribute level from the naïve model. The y-axis shows the *difference* in AMCE size between the highly attentive (>=2 CACs correct) and low-attentive (<=1 CAC correct). Vertical (horizontal) dashed lines indicate a naïve (differenced) AMCE estimate equal to 0..

Exploring Alternative Measures of Attentiveness

We also examined how MVCs and an IMC performed with respect to accounting for inattentiveness in conjoint experiments (see Supplemental Appendix for full results of these analyses). Overall, both MVCs and an IMC also demonstrate an ability to appreciably moderate the effects of attributes. That is, using either one represents a substantial improvement over a naïve model that ignores respondent inattentiveness.

However, we find evidence that CACs perform slightly better than these two alternative approaches in two key respects. First, we examined the conditional average treatment effect (CATE) sizes that result from interacting each attribute with CAC, MVC, and IMC performance separately.[15] Specifically, each measure was recoded to range from 0 to 1. We then ran three separate models, each interacting one of these three attentiveness measures with each attribute-level. The CATE is given by the coefficient on the interaction term, effectively representing the total change in effect size moving from lowest to highest performance on the attentiveness measure. For each type of attentiveness measure, we stored the mean of the absolute values of each CATE after grouping them by attribute (e.g., in the *Candidate* study, *education* would have five CATEs, the absolute values of which would be averaged to yield a single mean, while *military service* would have just one CATE). The logic is that a better measure of attentiveness should tend to have relatively larger (absolute) CATEs, and vice versa, insofar as respondents who are deemed inattentive should tend to have substantially smaller AMCEs compared to respondents who are deemed attentive.

Figure 6 displays the results of this analysis, with the results of the *Immigration* (*Candidate*) experiment shown in the top (bottom) panel. As noted above, both MVC and IMC performance tend to moderate effects, performing roughly similarly to CAC performance. Nevertheless, we find that for six out of the nine attributes in the *Immigration* experiment,

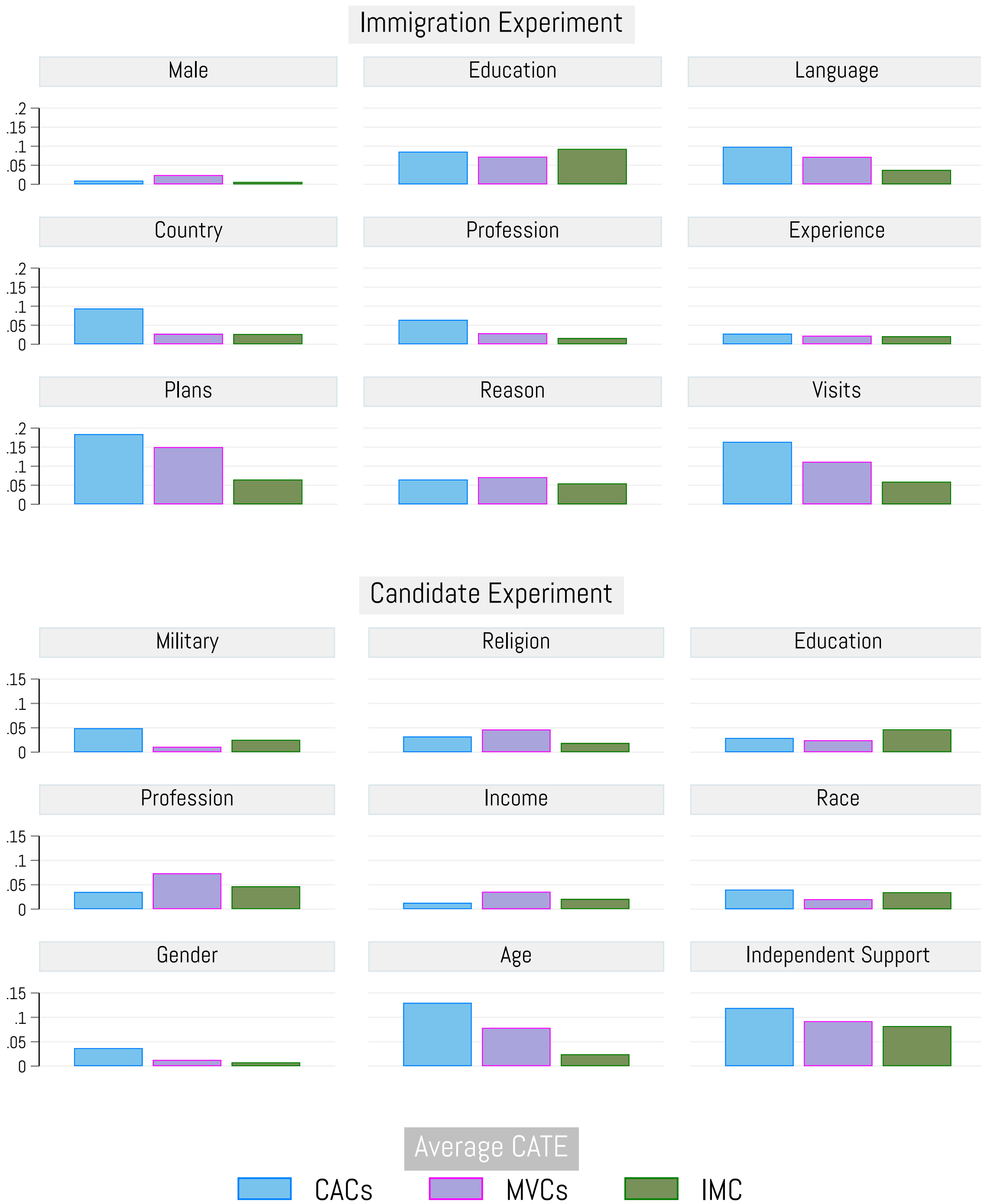
[15] Though different from the analyses shown above, the approach allows for a more straightforward comparison given that the meaning, and sample sizes, for a "High" and "Low" group would be different depending on which measure (CAC, MVC, or IMC) is being employed.

CACs have the largest average CATE value. Similarly, we find that for five out of the nine attributes in the *Candidate* experiment, CACs have the largest CATE value. This provides initial evidence that CACs may be best equipped to detect differences in how attentive vis-à-vis inattentive respondents perform in conjoint experiments.

Second, better measures of attentiveness should, when interacted with each attribute, explain more variability in the outcome measure. The logic here is that a poor measure of attentiveness, when interacted with a given attribute, should not significantly explain any more variability in the outcome than if no interaction between the two predictors had been specified. We therefore specified a series of nested regression models to determine whether the addition of interactions between the attributes and the measure of attentiveness (the second model) significantly improved model fit compared to when only the attributes and the measure of attentiveness were specified (the first model). We find that, for both experiments, the addition of interactions between attributes and CAC performance (compared to MVC and IMC performance) yields the largest percent increase in R^2 and is statistically significant in both cases. [16]

[16] We find that CACs yielded a 14% improvement in model fit in the *Immigrant* experiment ($p<.001$), compared to MVCs (4%; $p<.001$) and the IMC (2.3%; $p=.11$). In the *Candidate* experiment, CACs yielded a 23% improvement in model fit ($p=.002$), compared to MVCs (15%; $p=.28$) and the IMC (20%; $p=.08$).

Figure 6: Comparing Conditional Average Treatment Effect Sizes Across Attentiveness Measures



Note: Figure shows, for each attribute, the mean absolute values of all CATEs within that attribute. This analysis is performed separately for the Immigration Experiment (top panel) and Candidate Experiment (bottom panel).

Additional Analyses: Demographic Predictors of Attentiveness & Investigating Priming Effects

Previous research finds that some demographic groups perform better on measures of attentiveness than others (Thomas and Clifford, 2017). We therefore also examine: (1) the extent to which demographic factors predict performance on CACs, and (2) the extent to which we observe comparable patterns for other measures of attentiveness (namely, MVCs and the IMC). This approach enables us to discern qualities that may be unique to CACs vis-a-vis typical of attentiveness measures in general.

While the full results appear in the Supplemental Appendix, the key pattern across all measures of attentiveness (CACs, MVCs, the IMC, and the binary IRR measure) is that younger, non-college educated, and non-White respondents tend to perform worse on these items relative to older, college-educated, and/or White respondents. However, particularly for race and age, these patterns were somewhat weaker for CACs relative to other measures.[17] This implies that, for example, estimating effects among those who performed best on CACs (as opposed to MVCs or an IMC) may help to retain the representativeness of the sample. Secondly, in contrast to MVCs and an IMC, we do not find that using a smartphone (roughly two-thirds of each sample), relative to a desktop computer (roughly one-quarter of each sample) predicts lower performance on CACs. This is an unexpected finding, and suggests that using CACs (versus other measures) may help retain a larger sample when estimating effects among attentive respondents.

A second consideration is that CACs might incline respondents to be attentive to certain attributes in future tasks—i.e., the attributes that were asked about in the CACs. Especially, if we use CACs on the primary attributes of interest, it could appear as though a researcher is trying to augment treatment effects by priming respondents to attend to the key attributes of our study.

[17] In a similar vein, relative to MVCs and the IMC, less variance in CAC performance is explained by demographic factors.

We explore this possibility more fully in the Supplemental Appendix. Overall, however, we find limited, inconsistent evidence that CACs substantially alter AMCE estimates. This was determined by virtue of randomly assigning respondents to one of two *different* sets of CACs and comparing AMCEs across the two sets. While these differences are limited, the results nevertheless suggest that, to be conservative, researchers may want to employ CACs that ask about attributes not of primary theoretical interest. Again, this would be to minimize any risk of leading respondents to overly weight these attributes in their choices and ratings.

Discussion & Recommendations

The previous sections reveal that obtaining pre-treatment measures of attentiveness enables researchers to identify AMCEs among the more (versus less) attentive respondents. Including such measures enables the researcher to not only report the AMCEs for the sample as a whole, but to also conduct an unbiased, more robust test of one's hypotheses in conjoint-experimental designs.

Conjoint attentions checks (CACs) are one such pre-treatment measure. The purpose of CACs is to measure attentiveness to a typical conjoint presentation and, indeed, our study finds that CACs strongly predict alternative measures of attentiveness (e.g., an instructional manipulation check (IMC), a measure of intra-respondent reliability (IRR), and a mock vignette check (MVC)). CACs, as constructed here, were also not overly difficult for respondents to answer correctly, and demographic groups previously shown to exhibit relatively higher inattentiveness tended to perform relatively better on CACs than on alternative measures. Most importantly, we consistently find that CAC performance conditions AMCE estimates, with high-attention respondents tending to have substantially larger estimates than low-attention respondents. With respect to their ability to demonstrate different effect sizes for more (versus less) attentive respondents, CACs slightly outperform two common alternative approaches (an IMC and MVCs).

Thus, in both studies, we find substantial evidence for the efficacy of using pre-treatment measures of attentiveness to estimate AMCEs at varying levels of attentiveness. The technique we propose here–conjoint attention checks (CACs)–appears particularly well-suited for this task. However, there are two further points worth emphasizing.

First, any measure of attentiveness will contain some degree of error and, thus, will never permit a researcher to perfectly identify truly attentive sub-samples. Nevertheless, employing the general approach of using a pre-treatment measure of attentiveness (whether it is CAC performance, MVC performance, IMC performance, or something else) to estimate AMCEs represents a substantial improvement over estimating naïve AMCEs alone.

Table 4: Comparison between Mock Vignette Check (MVC), Instructional Manipulation Check (IMC), Inter-Respondent Reliability (IRR), and Conjoint Attention Check (CAC)

Feature	**MVC**	**IMC**	**IRR**	**CAC**
Straightforward to implement, especially for estimating standard errors and hypothesis testing	✓	✓		✓
Measure is pre-treatment, avoiding risk of post-treatment bias	✓	✓		✓
Can be used for both single-profile and two-profile conjoint experiments *and* choice and rating outcomes	✓	✓		✓
Tailored specifically for conjoint experiments			✓	✓
Allows inattentiveness to matter to varying degrees for different attributes and attribute-levels	✓	✓		✓
Allows all tasks to be used in the analysis	✓	✓		✓

Second, the approach we have outlined here is by no means mutually exclusive with other approaches (e.g., the one detailed by Clayton et al.). Indeed, employing *multiple* approaches stands to enable researchers to present a far more diverse, and therefore considerably more robust, analysis of their conjoint-experimental results. To assist researchers with adapting and implementing measures of attentiveness in their own conjoint studies, Table 4 outlines the main features of CACs compared with other techniques.

Conclusion

Failing to account for respondent inattentiveness in online experiments runs the risk of biasing treatment effects toward zero. This may be even more true for conjoint experiments, which ask respondents to carefully attend to a particularly large quantity of manipulated content, and for a particularly long amount of time.

Yet, as our aforementioned content analysis suggests (and Clayton et al. emphasize), conjoint experimentalists rarely appear to directly account for respondent inattentiveness in their experiments.

In this study, we explore several techniques for doing so and offer researchers a method that relies upon pre-treatment measures of attentiveness. The method is simple to implement in both single- and two-profile conjoint experiments, allows inattentiveness to matter to different degrees for different attribute levels, and permits straightforward analyses of one's results.

More generally, measuring attentiveness in conjoint experiments contributes to the growing literature on improving conjoint design and estimation (Abramson, Koçak and Magazinnik, 2022; Abramson et al., 2023; Clayton et al., 2023; Leeper, Hobolt and Tilley, 2020, e.g.). There has been some concern about whether respondents experience information overload (see Bansak et al., 2018; Bansak, Hainmueller, Hopkins and Yamamoto, 2021). In the studies presented here, approximately two-thirds of respondents (on average) got at least two out of three factual questions about a conjoint task correct. Knowing whether respondents are attentive to conjoint information strengthens the validity of any given experiment's results. It ensures that the conclusions drawn from the study reflect respondents' thoughtful engagement with the tasks rather than inattentiveness or random error.

In sum, the inclusion of pre-treatment measures of attentiveness accomplishes two crucial aims for researchers. First, they will be better positioned to estimate the substantive magnitude of attributes that truly matter to respondents. Second, for attributes that appear to exhibit little effect on a given outcome, researchers can better investigate whether or not this "null result" is largely due to respondent inattentiveness, e.g., via examining whether or not the AMCE becomes substantially stronger at higher levels of respondent attentiveness (see Kane, 2024). Thus, by accounting for respondent inattentiveness in this way, conjoint-experimentalists can better distinguish meaningful results from statistical noise.

References

Abramson, Scott F, Korhan Koçak and Asya Magazinnik. 2022. "What do we learn about voter preferences from conjoint experiments?" *American Journal of Political Science* 66(4):1008–1020.

Abramson, Scott F, Korhan Kocak, Asya Magazinnik and Anton Strezhnev. 2023. Detecting Preference Cycles in Forced-Choice Conjoint Experiments. Technical report Working Paper.

Bailey, Michael A. 2021. *Real stats: Using econometrics for political science and public policy.* Oxford University Press.

Bansak, Kirk, Jens Hainmueller, Daniel J Hopkins and Teppei Yamamoto. 2018. "The number of choice tasks and survey satisficing in conjoint experiments." *Political Analysis* 26(1):112–119.

Bansak, Kirk, Jens Hainmueller, Daniel J Hopkins and Teppei Yamamoto. 2021. "Beyond the breaking point? Survey satisficing in conjoint experiments." *Political Science Research and Methods* 9(1):53–71.

Bansak, Kirk, Jens Hainmueller, Daniel J Hopkins, Teppei Yamamoto, James N Druckman and Donald P Green. 2021. "Conjoint survey experiments." *Advances in experimental political science* 19:19–41.

Berinsky, Adam J, Alejandro Frydman, Michele F Margolis, Michael W Sances and Diana Camilla Valerio. 2024. "Measuring Attentiveness in Self-Administered Surveys." *Public Opinion Quarterly* 88(1):214–241.

Berinsky, Adam J, Michele F Margolis and Michael W Sances. 2014. "Separating the shirkers from the workers? Making sure respondents pay attention on self-administered surveys." *American journal of political science* 58(3):739–753.

Berinsky, Adam J, Michele F Margolis and Michael W Sances. 2016. "Can we turn shirkers into workers?" *Journal of Experimental Social Psychology* 66:20–28.

Clayton, Katherine, Yusaku Horiuchi, Aaron R Kaufman, Gary King, Mayya Komisarchik, Danny Ebanks, Jonathan N Katz, Gary King, Georgina Evans, Gary King et al. 2023. "Correcting measurement error bias in conjoint survey experiments." *American Journal of Political Science* 12(B2):1–11.

Druckman, James N. 2022. *Experimental thinking.* Cambridge University Press.

Gummer, Tobias, Joss Roßmann and Henning Silber. 2021. "Using instructed response items as attention checks in web surveys: Properties and implementation." *Sociological Methods & Research* 50(1):238–264.

Hainmueller, Jens and Daniel J Hopkins. 2015. "The hidden American immigration consensus: A conjoint analysis of attitudes toward immigrants." *American journal of political science* 59(3):529–548.

Hainmueller, Jens, Daniel J Hopkins and Teppei Yamamoto. 2014. "Causal inference in conjoint analysis: Understanding multidimensional choices via stated preference experiments." *Political analysis* 22(1):1–30.

Jenke, Libby, Kirk Bansak, Jens Hainmueller and Dominik Hangartner. 2021. "Using eye-tracking to understand decision-making in conjoint experiments." *Political Analysis* 29(1):75–101.

Kane, John V. 2024. "More than meets the ITT: A guide for anticipating and investigating nonsignificant results in survey experiments." *Journal of Experimental Political Science* pp. 1–16.

Kane, John V and Jason Barabas. 2019. "No harm in checking: Using factual manipulation checks to assess attentiveness in experiments." *American Journal of Political Science* 63(1):234–249.

Kane, John V, Yamil R Velez and Jason Barabas. 2023. "Analyze the attentive and bypass bias: Mock vignette checks in survey experiments." *Political Science Research and Methods* 11(2):293–310.

Krosnick, Jon A. 1991. "Response strategies for coping with the cognitive demands of attitude measures in surveys." *Applied cognitive psychology* 5(3):213–236.

Leeper, Thomas J, Sara B Hobolt and James Tilley. 2020. "Measuring subgroup preferences in conjoint experiments." *Political Analysis* 28(2):207–221.

Meade, Adam W and S Bartholomew Craig. 2012. "Identifying careless responses in survey data." *Psychological methods* 17(3):437.

Montgomery, Jacob M, Brendan Nyhan and Michelle Torres. 2018. "How conditioning on posttreatment variables can ruin your experiment and what to do about it." *American Journal of Political Science* 62(3):760–775.

Oppenheimer, Daniel M, Tom Meyvis and Nicolas Davidenko. 2009. "Instructional manipulation checks: Detecting satisficing to increase statistical power." *Journal of experimental social psychology* 45(4):867–872.

Silber, Henning, Daniel Danner and Beatrice Rammstedt. 2019. "The impact of respondent attentiveness on reliability and validity." *International Journal of Social Research Methodology* 22(2):153–164.

Sniderman, Paul M. 2018. "Some advances in the design of survey experiments." *Annual Review of Political Science* 21(1):259–275.

Stagnaro, Michael Nicholas, James Druckman, Adam J Berinsky, Antonio Alonso Arechar, Robb Willer and David Rand. 2024. "Representativeness versus attentiveness: A comparison across nine online survey samples." *PsyArXiv* .

Thomas, Kathrin. 2024. "The Advent of Survey Experiments in Politics and International Relations." *Government and Opposition* 59(1):297–320.

Thomas, Kyle A and Scott Clifford. 2017. "Validity and Mechanical Turk: An assessment of exclusion methods and interactive experiments." *Computers in Human Behavior* 77:184–197.

Wood, Dustin, Peter D Harms, Graham H Lowman and Justin A DeSimone. 2017. "Response speed and response consistency as mutually validating indicators of data quality in online samples." *Social Psychological and Personality Science* 8(4):454–464.

Appendices:

Being Careful with Conjoints:
Accounting for Inattentiveness in Conjoint Experiments

October 15, 2024

Contents

https://doi.org/10.33774/apsa-2024-xqt7d-v2 Content not peer-reviewed by APSA.

A Conjoint Attention Checks: Question Text

Immigrant Experiment

CAC Set 1

From the previous screen, which immigrant had more education?
– Immigrant 1 (the one shown on the left) (1)
– Immigrant 2 (the one shown on the right) (2)
– The two immigrants had an equal amount of education (3)
– Education was not mentioned (4)

Which immigrant had more prior trips to the U.S.?
– Immigrant 1 (the one shown on the left) (1)
– Immigrant 2 (the one shown on the right) (2)
– The two immigrants had an equal number of prior trips to the US (3)
– Number of prior trips to the US was not mentioned (4)

Which immigrant was older in age?
– Immigrant 1 (the one shown on the left) (1)
– Immigrant 2 (the one shown on the right) (2)
– The two immigrants were the same age (3)
– Age was not mentioned (4)

CAC Set 2

From the previous screen, which immigrant's profession was listed as "Child Care Provider"?
– Immigrant 1 (the one shown on the left) (1)
– Immigrant 2 (the one shown on the right) (2)
– Both immigrants' professions were listed as "Child Care Provider" (3)
– Profession was not mentioned (4)

Which immigrant had more job experience?
– Immigrant 1 (the one shown on the left) (1)
– Immigrant 2 (the one shown on the right) (2)
– Both immigrants had an equal amount of job experience (3)
– Job experience was not mentioned (4)

Which immigrant was older in age?
– Immigrant 1 (the one shown on the left) (1)
– Immigrant 2 (the one shown on the right) (2)
– The two immigrants were the same age (3)
– Age was not mentioned (4)

Candidate Experiment

CAC Set 1

From the previous screen, what was the candidate's religion?
– Evangelical Protestant (1)
– Catholic (2)
– Jewish (3)
– Religion was not mentioned (4)

What was the candidate's gender?
– Woman (1)
– Man (2)
– Non-binary (3)
– Gender was not mentioned (4)

What was the candidate's political party?
– Democrat (1)
– Republican (2)
– Independent (3)
– Political party was not mentioned (4)

CAC Set 2

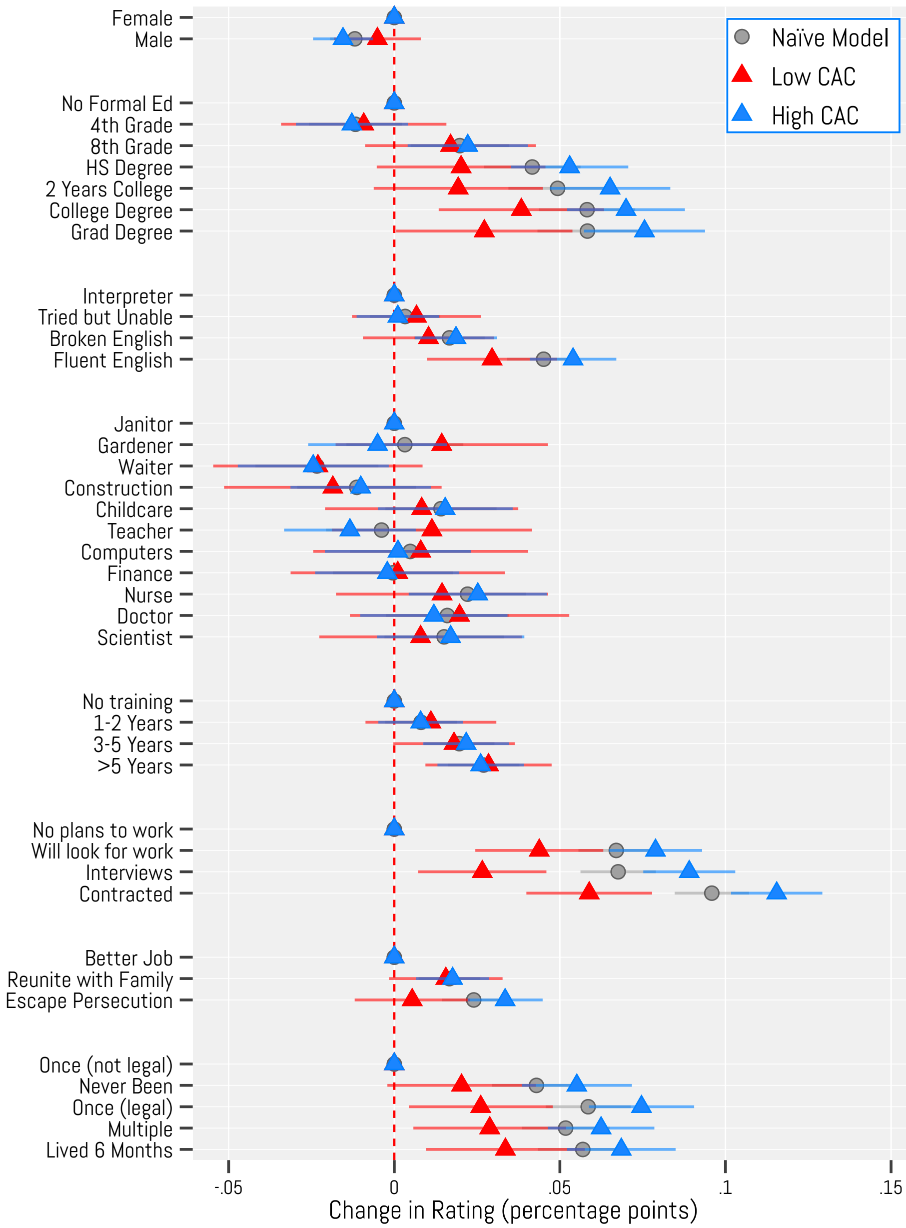
From the previous screen, what was the candidate's profession?
– Farmer (1)
– Lawyer (2)
– Doctor (3)
– Profession was not mentioned (4)

What was the candidate's level of education?
– JD (1)
– BA (2)
– PhD (3)
– Education was not mentioned (4)

What was the candidate's political party?
– Democrat (1)
– Republican (2)
– Independent (3)
– Political party was not mentioned (4)

B Rating Outcome: Immigration Experiment

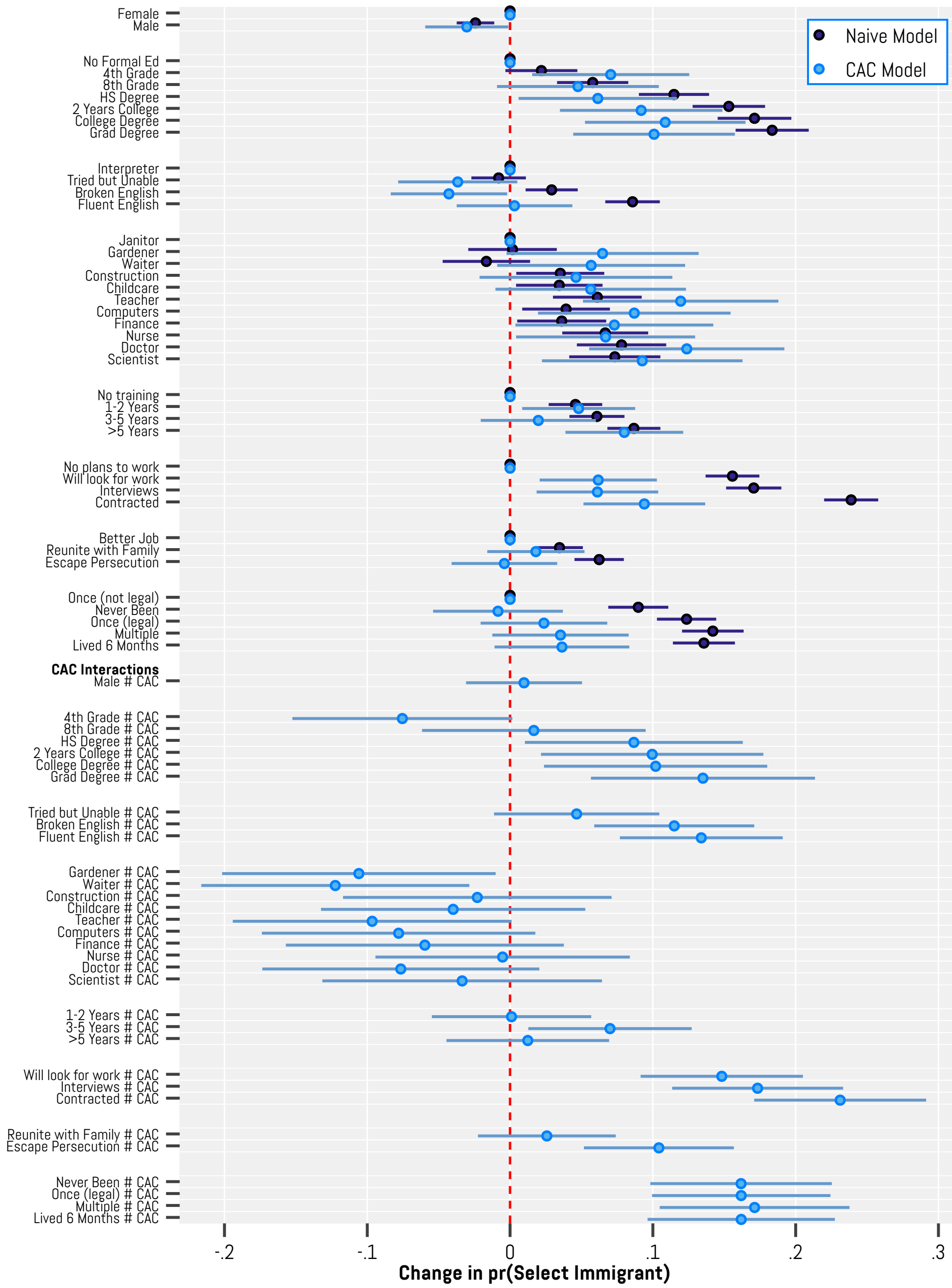
Figure B1: AMCEs at Low & High CAC Performance in Immigration Experiment (Ratings)



Note: The figure displays AMCEs from naïve model that does not account for attentiveness, as well as AMCEs for respondents with low performance on the CACs (<=1 correct CAC) and high perfomance on the CACs (>=2 correct CACs). Each AMCE indicates the estimated change in rating on a given immigrant. The rating was originally a seven-point scale ranging from 1=Definitely would not admit to 7 Definitely would admit, and was rescaled to range from 0 to 1 to allow AMCEs to be interpreted in percentage-point changes. The underlying model includes all attributes (with standard errors clustered by respondent) but the figure excludes the constant and the targets' country of origin to conserve space. The underlying model clusters standard errors by respondent). Total n=24,910 (2,127 individual respondents). Lucid Data.

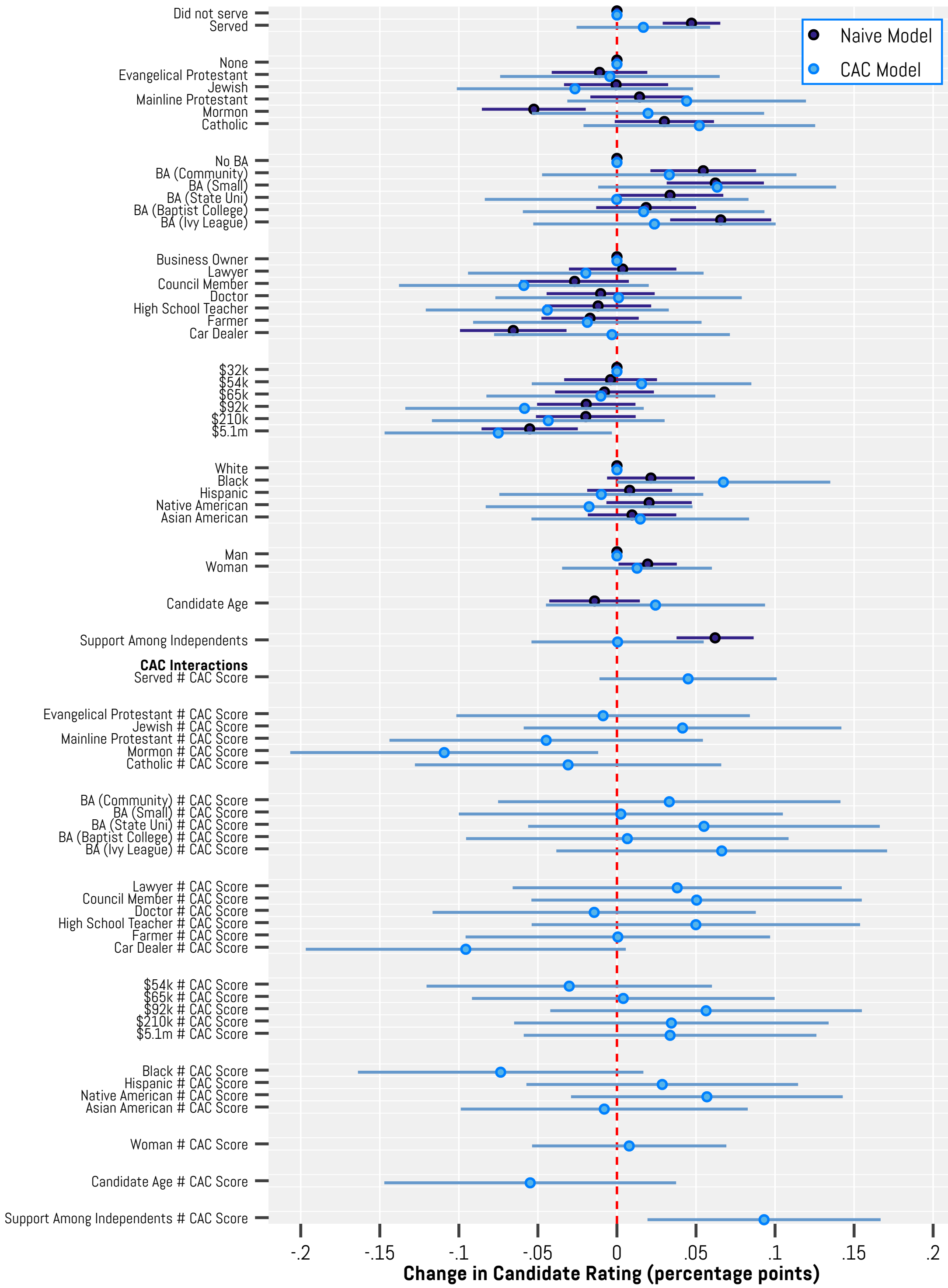
C Full CATE Results from Both Experiments

Figure C1: Naïve Model and CATEs Using CACs in Immigration Experiment



Note: The top portion presents AMCEs from a naïve model that does not account for attentiveness. The bottom portion displays conditional average treatment efefcts (CATE)s, highlighting the change in probability of an immigrant being selected. The CAC Scale was recoded to range from 0 to 1 for this analysis (interactions thus reflect the total change in each AMCE across the entire CAC scale). The model includes all attributes but the figure excludes the constant and the targets' country of origin to conserve space.

Figure C2: Naïve Model and CATEs Using CACs in Candidate Experiment



Note: The top portion presents AMCEs from a naïve model that does not account for attentiveness. The bottom portion displays conditional average treatment effects (CATE)s, highlighting the change in candidate rating. The CAC Scale was recoded to range from 0 to 1 for this analysis (interactions thus reflect the total change in each AMCE across the entire CAC scale).

D Demographic predictors of CAC performance

https://doi.org/10.33774/apsa-2024-xqt7d-v2

Table D1: Demographic Predictors of Attentiveness (Immigrant Experiment)

	(1) CAC Scale	(2) MVC Scale	(3) IMC	(4) IRR
Gender: Non-Binary	0.08	0.05	0.09	0.12
	(0.09)	(0.09)	(0.13)	(0.12)
Gender: Decline	-0.11	-0.14	0.01	-0.07
	(0.19)	(0.20)	(0.28)	(0.31)
Gender: Woman	0.05***	0.01	0.03	0.02
	(0.01)	(0.01)	(0.02)	(0.02)
Race: Asian	-0.00	-0.14***	-0.05	-0.09^
	(0.04)	(0.04)	(0.06)	(0.05)
Race: Black	-0.10***	-0.15***	-0.16***	-0.09**
	(0.02)	(0.02)	(0.03)	(0.03)
Race: Hispanic	-0.02	-0.11***	-0.12**	-0.08*
	(0.03)	(0.03)	(0.04)	(0.04)
Race: Native Am.	-0.01	-0.07	-0.12	0.02
	(0.06)	(0.06)	(0.08)	(0.08)
Race: ME / Other	0.04	0.01	-0.13	-0.14
	(0.11)	(0.11)	(0.15)	(0.14)
Race: 2 or more	-0.04	-0.08^	0.04	0.01
	(0.05)	(0.05)	(0.07)	(0.07)
Edu: HS Diploma	0.07^	0.10*	0.13*	0.06
	(0.04)	(0.04)	(0.06)	(0.06)
Edu: Some College	0.16***	0.17***	0.19**	0.15*
	(0.04)	(0.04)	(0.06)	(0.06)
Edu: 2 yr. College	0.11*	0.12**	0.14*	0.13*
	(0.05)	(0.05)	(0.07)	(0.06)
Edu: 4 yr. College	0.15***	0.18***	0.19**	0.16**
	(0.04)	(0.04)	(0.06)	(0.06)
Edu: Grad Degree	0.12**	0.18***	0.21**	0.22***
	(0.05)	(0.05)	(0.07)	(0.06)
Age	0.15***	0.37***	0.61***	0.19**
	(0.04)	(0.04)	(0.06)	(0.06)
Device: Smartphone	-0.03	-0.07***	-0.09***	-0.04^
	(0.02)	(0.02)	(0.03)	(0.02)
Device: Tablet	-0.11***	-0.06^	-0.07**	-0.02
	(0.03)	(0.03)	(0.04)	(0.04)
Constant	0.47***	0.38***	0.21**	0.60***
	(0.05)	(0.05)	(0.07)	(0.06)
Observations	2,116	2,190	2,184	2,100
R-squared	0.06	0.12	0.10	0.04

Note: All models are OLS to allow for direct comparison (SEs in parentheses). Middle East (ME) and Other racial identifications combined due to small sample sizes (¡1% total). Baseline categories are male (Gender); White (Race); No High School Diploma (Edu); Desktop computer (Device). *** p<.001; ** p<.01; * p<.05; ^ p<.10. Lucid data.

Table D2: Demographic Predictors of Attentiveness (Candidate Experiment)

	(1) CAC Scale	(2) MVC Scale	(3) IMC	(4) IRR
Gender: Non-Binary	0.05	-0.01	-0.23	-0.27
	(0.11)	(0.12)	(0.17)	(0.19)
Gender: Other	-0.05	0.06	0.20	0.05
	(0.22)	(0.24)	(0.34)	(0.36)
Gender: Decline	0.13	-0.04	-0.31	-0.42
	(0.22)	(0.24)	(0.34)	(0.36)
Gender: Woman	-0.01	0.01	0.05	0.03
	(0.02)	(0.02)	(0.03)	(0.03)
Race: Asian	-0.11*	-0.20***	-0.08	0.04
	(0.04)	(0.05)	(0.07)	(0.07)
Race: Black	-0.10***	-0.20***	-0.05	-0.05
	(0.03)	(0.03)	(0.05)	(0.05)
Race: Hispanic	-0.10**	-0.11**	-0.13*	-0.15*
	(0.04)	(0.04)	(0.06)	(0.06)
Race: Native Am.	0.08	-0.12	0.19	-0.07
	(0.09)	(0.10)	(0.14)	(0.15)
Race: Other	0.05	-0.08	-0.27	0.03
	(0.11)	(0.12)	(0.17)	(0.18)
Race: 2+	0.01	0.06	0.08	-0.12
	(0.08)	(0.09)	(0.13)	(0.14)
Edu: HS Diploma	0.06	0.06	0.08	0.05
	(0.05)	(0.05)	(0.08)	(0.08)
Edu: Some College	0.16**	0.10ˆ	0.11	0.06
	(0.05)	(0.05)	(0.08)	(0.08)
Edu: 2 yr. College	0.09ˆ	0.09	0.21*	0.05
	(0.05)	(0.06)	(0.08)	(0.09)
Edu: 4 yr. College	0.16**	0.14*	0.19*	0.09
	(0.05)	(0.05)	(0.08)	(0.08)
Edu: Grad Degree	0.13*	0.15*	0.05	0.07
	(0.05)	(0.06)	(0.09)	(0.09)
Age	0.28***	0.33***	0.46***	0.12ˆ
	(0.05)	(0.05)	(0.07)	(0.07)
Device: Smartphone	-0.04	-0.00	-0.09*	-0.05
	(0.02)	(0.03)	(0.04)	(0.04)
Device: Tablet	-0.07ˆ	0.02	-0.05	-0.09
	(0.04)	(0.04)	(0.06)	(0.06)
Constant	0.49***	0.37***	0.24**	0.42***
	(0.06)	(0.06)	(0.09)	(0.09)
Observations	1,046	1,046	1,046	1,027
R-squared	0.13	0.14	0.11	0.03

Note: All models are OLS to allow for direct comparison between models (SEs in parentheses). Baseline categories are male (Gender); White (Race); No High School Diploma (Edu); Desktop computer (Device). *** $p<0.001$; ** $p<0.01$; * $p<0.05$; ˆ $p<0.10$. Lucid data.

E Investigating Potential Priming Effects

A potential concern with CACs is that, by asking respondents about particular attribute levels after the first task, it may lead respondents to focus more on these attributes–and/or less on attributes that were *not* asked about–in subsequent tasks. In this section, therefore, we examine whether AMCE results differ substantially depending upon which CAC set was used.

Beginning with the *Immigrant* experiment, this experiment featured two different sets of CACs: Set 1 asked which target had more education, had more prior trips to the US, and was older in age (note: age was not an attribute featured in the experiment). Set 2 asked about which target had a particular profession, had more job experience, and (again) was older in age.

If CACs significantly change how people process the attributes, we should expect to see that AMCEs will differ substantially depending upon which CAC set a respondent was randomly assigned to.

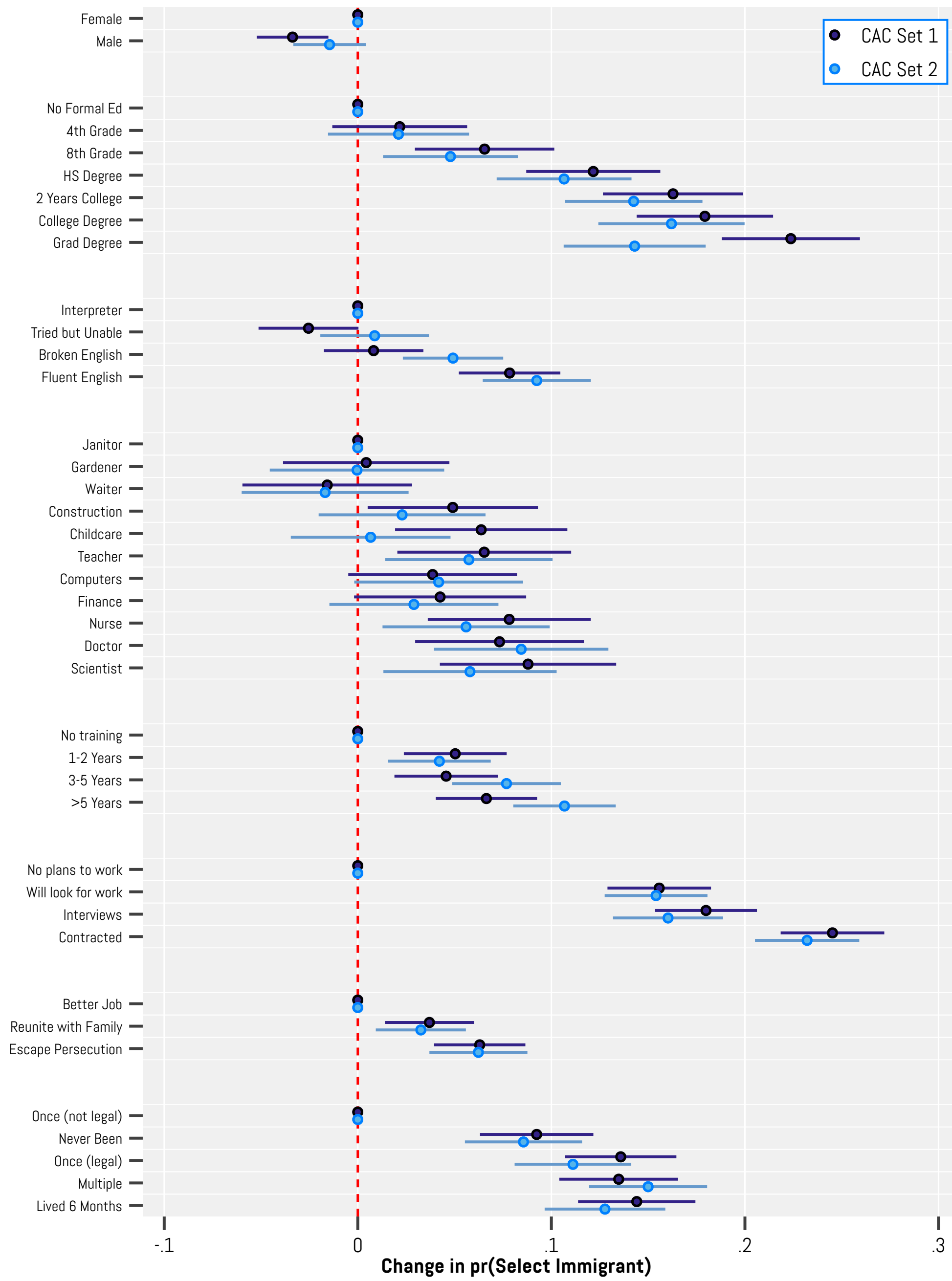
Figure E1 displays AMCEs separately for respondents assigned to CAC Set # 1 vs. Set #2.

Overall, the figure offers only slight evidence that the CACs might prime respondents to attend to particular attributes. For example, CAC Set #1 asked about the target's education level. We see that respondents who saw CAC Set #1 had, overall, slightly larger AMCEs than those who saw CAC Set #2. Notably, the one clearly significantly different AMCE—"Grad degree"—was not referenced in the CAC (respondents were asked which immigrant had more education, and neither had a graduate degree).

However, no such pattern is evident for the other CAC from Set #1, which asked about prior trips/time spent in US. These effects are substantively similar regardless of CAC Set.

CAC Set #2 asked about the target's profession. There is again no clear pattern, with AMCEs looking substantively similar regardless of which CAC Set was seen.

Figure E1: AMCEs for CAC1 and CAC2 in Immigration Experiment



CAC Set #2 also asked about the target's job experience. Here, the two highest levels of the attribute show slightly larger AMCEs for respondents who saw CAC Set 2 compared to those who saw CAC Set #1 (between 3 and 4 percentage points).

Using seemingly unrelated estimation (SUE), we confirmed that the AMCEs for the highest attribute level (i.e., more than 5 years job experience) are significantly different between the two models at conventional levels (p=.03), but the lower level (3-5 years job experience) is not (p=.11).

That said, some degree of variation in AMCEs is expected due to random sampling variation (i.e., because respondents saw either CAC Set #1 or Set #2, we should expect to observe some differences, particularly given the large number of AMCEs being estimated). On this point, we also observe some differences between the two groups on items that were *not* asked about in either CAC Set.

For example, there is a significant interaction between CAC Set and English language ability (p<.05), with the "Broken English" level exhibiting a 4 percentage-point larger AMCE for the CAC Set #2 group compared to the CAC Set #1 group. (SUE analysis also confirmed these AMCEs were significantly different (p=.03). Additionally, there were several country AMCEs (Iraq, Philippines, and Germany) that significantly interacted with CAC set.

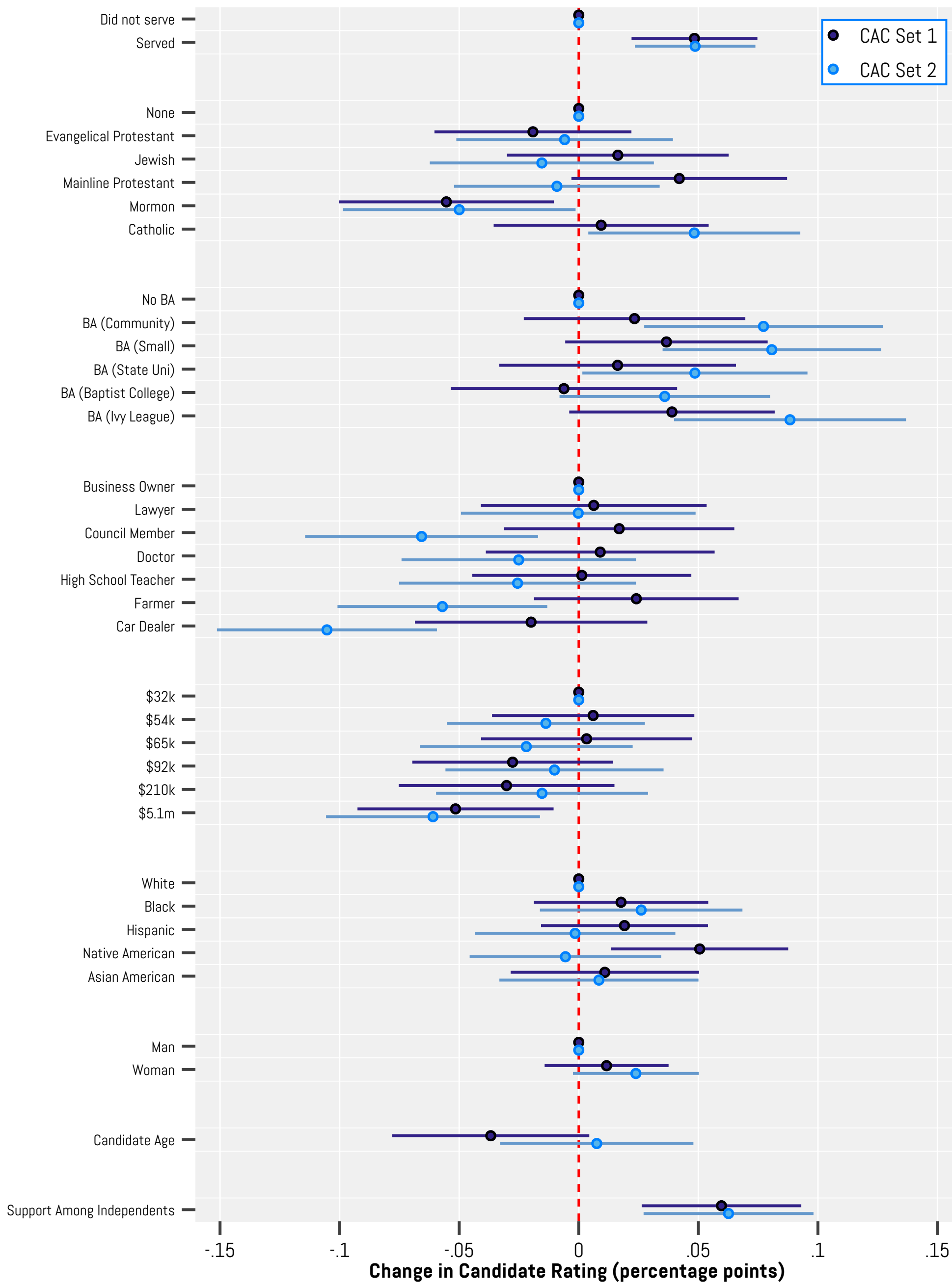
Again, because these attributes were not asked about in either CAC Set, it suggests that some differences are to be expected purely because of random sampling error and the large number of coefficients being estimated.

We also investigated potential priming using the ratings outcome of the *Immigrant* experiment. This analysis again provides only limited evidence for priming: there is a noticeable pattern of differences for education, though the pattern is opposite what was seen in the above figure–namely, the CAC Set #1 group now exhibits *smaller* AMCEs compared to the CAC Set #2 group.

CAC Set #1 group was also asked about trips/time spent in US and shows somewhat smaller effects compared to CAC Set #2 group, whereas it showed somewhat *larger* effects in the figure above.

There is only one noticeable difference on the job training item, which was asked in CAC Set #2, with the coefficient on 3-5 years being somewhat larger for the CAC Set 2 group. When interacted with CAC Set, this was the only coefficient to have obtained a statistically significant interaction at the $p<.05$ level. There was no clear pattern for the profession attribute, which was also asked of CAC Set #2.

Figure E2: AMCEs for CAC1 and CAC2 in Candidate Experiment



We conducted the same analysis for the *Candidate* experiment. The results are shown in Figure E2. In this case we do observe a pattern with respect to the attributes asked about in CAC Set #1: though not significantly different, AMCEs tend to be substantively larger in magnitude when the attribute was asked about in CAC Set #2. Namely, the AMCEs for the education attribute tend to be more positive, while the AMCEs for the profession attribute tend to be more negative. That said, the attributes asked about in CAC Set #1 (religion and gender) do not tend to show any clear pattern with respect to directionality nor magnitude.

Overall, out of 29 interactions (between the attribute levels and whether the CACs were from Set 1 or Set 2), only two interactions involving attributes featured in a CAC attained significance at the .05 level. (At the same time, one interaction that did *not* involve an attribute featured in a CAC–Native American race–also attained significance.) Thus, in this experiment there is slight, but inconsistent evidence that CACs can alter effects of attributes by asking about them. However, for items that were not asked about there is very little difference in results, again suggesting that asking about certain attributes does not reduce attention to other items.

Again, we do not find much evidence for substantial priming effects. However, though the evidence is slight, researchers may be best advised to ask CACs about attributes that are *not* central to their hypotheses. This can avoid any potential concern that the experimenter may be inducing demand effects (though see Mummolo et al. 2019) and/or systematically leading respondents to ignore other attributes. On this latter point, however, it is worth noting that other attributes—e.g., work plans and visits/time spent in the US—exhibit large effects, despite not being asked about in either CAC Set. This pattern indicates that CACs do not necessarily reduce attention to other attributes that might be more important to researchers.