Generalization Problem in Experiments Involving Multi-dimensional Decisions

Jiawei Fu* Xiaojun Li†

March 9, 2024

Abstract

Can the internally valid effects observed in experiment be generalized to real-world scenarios? This question lies at the heart of social science studies, where the ultimate concern is the real-life impact of findings. External validity primarily assesses whether experimental effects persist across different settings, including populations, treatments, outcomes, and contexts, implicitly presuming the experiment's ecological validity—that is, the consistency of experimental effects with their real-life counterparts even without dramatic varying those settings. However, we argue that this presumed consistency may not always hold, especially in experiments involving multi-dimensional decision processes, such as conjoint survey experiments. We introduce a formal model to elucidate how attention and salience effects lead to three types of inconsistencies between experimental findings and real-world phenomena: amplified effect magnitude, effect sign reversal, and effect relative importance reversal. We derive testable hypotheses from each theoretical outcome and verify these hypotheses using data from various existing conjoint experiments, in addition to conducting our own experiments. Drawing on our theoretical framework, we propose several guidelines for experimental design aimed at enhancing the generalizability of survey experiment findings.

Keywords: Conjoin Experiment, Survey Experiment, Ecological Validity, External Validity, Generalizability, Attention, Salience.

*Ph.D. Candidate, New York University jf3739@nyu.edu
†Associate Professor of Political Science, NYU Shanghai xl4335@nyu.edu

Introduction

The randomized experiment is often hailed as the gold standard for identifying and estimating causal effects within the social sciences. Researchers frequently employ conjoint survey experiments analysis to examine how individual preferences for candidates or immigrants are influenced by characteristics such as gender, race, experience, and skills (Eggers et al., 2018; Carnes and Lupu, 2016; Sanbonmatsu, 2002). However, the interest of researchers extends well beyond merely observing causal effects within an experimental setting. The ultimate goal is to ensure that these experimental estimates accurately mirror real-world effects. For instance, in a survey experiment where it is discovered that voters exhibit a preference for female candidates with an average causal effect of approximately 2%, the hope is that this finding corresponds to an actual increase in preference for female candidates in the real world. When such a correspondence is established, we regard the experimental effect as being consistent with the real-world effect. This consistency, which establishes the ecological validity of experiment, serves as the ultimate concern of social science studies.

Ecological validity represents a distinct facet of external validity (Kihlstrom 2021). As depicted in Figure 1, internal validity refers to the validity within the experimental setting. External validity concerns the generalization of results from the experimental domain to real-world contexts with varied populations, treatments, outcomes, and contexts (Cook et al. 2002). A fundamental premise is that the effects observed in an experiment should be consistent with real-life effects, even without dramatic variations in populations, treatments, outcomes, or contexts. If the effects estimated in the experiment do not align with those experienced by the same individuals in the real world, it undermines the rationale for discussing the generalizability of these results to other contexts. We demonstrate that experiments involving multi-dimensional decision-making processes, such as the extensively employed vignette and conjoint survey experiments, frequently fail to meet that ecological validity. This failure raises substantial concerns about our ability to generalize experimental insights to other real-life scenarios effectively.

Within the scope of this paper, we analyze three types of consistency: effect magnitude, effect

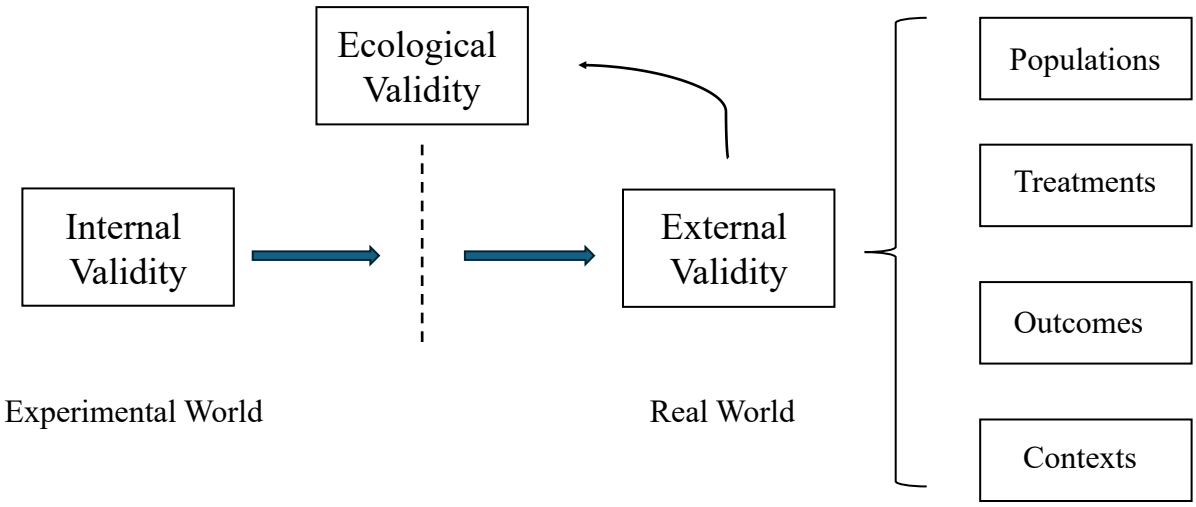

Figure 1: EXPERIMENTAL VALIDITY: ECOLOGICAL VALIDITY AS A PREMISE OR A SPECIAL SUBSET OF EXTERNAL VALIDITY

direction, and attribute importance. The *Consistency of Effect Magnitude* asserts that the observed size of an experimental effect should mirror the magnitude noted in real-world phenomena. This criterion, while offering the deepest insights into real-world applicability, stands as the most challenging to fulfill. The *Consistency of Effect Direction* posits that the sign (positive or negative) of an experimental effect should match its real-world counterpart. Achieving magnitude consistency ensures direction consistency by default, rendering the latter a less rigorous yet still crucial requirement for real-world insights. Lastly, *Consistency of Attribute Importance* emphasizes the need for experimental outcomes to truthfully represent the relative significance of different treatments or attributes. This aspect is particularly valuable for researchers aiming to identify which attributes most significantly influence variations in outcomes, marking it as another derivative of magnitude consistency.

It is not imperative for an experiment to meet all three consistency criteria simultaneously; rather, the requisite level of consistency should align with the experiment's objectives. For in-

stance, program evaluations conducted by policymakers, who require effect magnitude information to assess overall welfare changes, necessitate experiments to fulfill magnitude consistency. Conversely, when experiments aim to test predictions derived from formal theory, often only the direction consistency is essential. Similarly, attribute importance consistency suffices when the goal is to understand attribute influences and to predict outcomes[1].

We develop a formal theory elucidating why experiments involving multi-dimensional decision-making frequently fall short of fulfilling the three consistency criteria. At the heart of our theoretical framework lies the interplay between attention and salience. Conventional decision-making theories have often posited that individuals possess unlimited attention and exhibit consistent salience across contexts. Such assumptions yield straightforward theoretical predictions but have been increasingly questioned due to numerous observed anomalies and paradoxes. Notably, the Allais (1953) and Ellsberg (1961) paradoxes serve as prominent examples that challenge these traditional views. While the bulk of existing research has focused on refining decision theory to account for these discrepancies, our work pioneers the application of a behavioral theoretical approach to systematically address these inconsistencies within experimental research methodologies.

Our primary formal finding underscores that experimental effects are notably exaggerated due to the phenomena of limited attention, which is also observed by Barabas and Jerit (2010) in the context of survey experiments. We extend this observation by providing a formal framework that outlines sufficient conditions under which this phenomenon occurs. We identify the attention effect within experiments that involving multi-dimensional decision-making processes. Attention effect posits that the consideration set within an experimental context diverges significantly from that in real-world scenarios. The limitation and transience of human attention have long been recognized by social scientists. For example, consumer choices may favor option y in the presence of x, but this preference can reverse when x is not immediately apparent to the consumer (Masatlioglu et al., 2016). Chetty's field experiments demonstrate that consumers tend to under-react to

[1]In machine learning exercises, for example, predictors that account for the majority of variation are assigned greater importance and have a more significant impact on predictions.

taxes when tax-inclusive pricing is not prominently displayed, leading to an 8 percent decrease in demand when it is (Chetty et al., 2009). We formalize the idea that the experimental setting can inadvertently direct and intensify individual attention to certain attributes, resulting in skewed estimates.

To elucidate the inconsistencies in effect sign and relative importance, we delve into the salience effect. The main intuition is that in a complex decision-making environment, the salience effect can correlate otherwise independent attributes, significantly altering the weight placed on certain attributes and potentially reversing decision outcomes. To precisely model this salience effect, we draw on the psychologically grounded model of choice proposed by Bordalo et al. (2012, 2013), which posits that salience is primarily driven by the perceived intensity of each attribute.

We derive several corresponding testable hypotheses from our theoretical results. For example, a key hypothesis suggests that the observed effect diminishes or even reverses as the number of attributes increases. To empirically test these hypotheses, we analyze data from multiple existing conjoint experiments and conduct our series of conjoint experiments. The experimental findings consistently align with our hypotheses, thereby reinforcing our theoretical framework. Given the elucidation of mechanisms driving inconsistencies within our theory, we propose several experimental design recommendations in the discussion section aimed at alleviating these issues.

Our research intersects with and contributes to several vital streams of literature. At the core of our study is a pivotal inquiry into the conjoint experiment literature. Our exploration is timely and critical, considering the burgeoning body of research on conjoint analysis since the seminal work of Hainmueller et al. (2014b). Subsequent investigations have primarily honed in on statistical concerns—such as external validity (De la Cuesta et al., 2022), measurement errors (Clayton et al., 2023), power analysis (Stefanelli and Lukac, 2020), survey satisficing (Bansak et al., 2021), social desirability bias (Horiuchi et al., 2022), hypotheses testing (Ham et al., 2022), and multiple testing (Horiuchi et al., 2022)—yet, the fundamental issue of real-world consistency remains underexplored. Our findings analyze the conditions under which conjoint experiments are reliable.

Second, our work dialogues with the emerging literature on Theoretical Implication of Empir-

ical Methods (TIEM) (Slough and Tyson, 2023; De Mesquita and Tyson, 2020; Slough, 2023; Fu and Slough, 2023), which critically assesses empirical methodologies through the lens of formal theory. This approach, which contrasts with the Empirical Implications of Theoretical Models (EITM), enables a deeper understanding of the limitations and potential inconsistencies within empirical research methods. In the context of conjoint experiments, our application of decision theory and behavioral economics illuminates foundational discrepancies between experimental effects and their real-world counterparts, extending beyond the aggregate reference inconsistencies identified by Abramson et al. (2019).

Third, we contribute to the external validity, which is a core concern in experimental research, especially in lab and survey experiments (Barabas and Jerit 2010; List and Levitt 2005; Gaines et al. 2007). Egami and Hartman (2023) develop a unified formal framework to discuss four types of external validity, including population, contexts, treatments, and outcomes, as proposed by Cook et al. (2002). Slough and Tyson (2023) also analyze external validity within meta-analysis. Most studies primarily focus on whether the results can be generalized to a new population (Mullinix et al. 2015). Huang (2022) provides a sensitivity analysis for extrapolating the effect to a new population. Our paper, however, focuses on ecological validity, which compares the experimental effect and real-world effect even for the same individuals, treatments, and outcomes. We provide a formal framework to analyze ecological validity for experiments involving multi-dimensional decision processes. Hainmueller et al. (2014a) empirically find that survey experiments have high ecological validity[2]. Their results do not contradict our findings because their specific design successfully eliminates the attention effect. Therefore, their findings further strengthen our theory.

Lastly, we broaden the application of salience theory beyond its traditional confines in political science and economics, where its implications for election, party competition, agenda-setting, and judicial biases have been well-documented (Dragu and Fan, 2016; Riker et al., 1986; Ascencio and Gibilisco, 2015; Guthrie et al., 2000; Viscusi, 2001). By integrating salience theory into experimental design and applying the framework developed by Bordalo et al. (2012, 2013, 2016),

[2]In their paper, they primarily describe it as external validity. Actually, they demonstrate both external validity (comparing different individuals) and ecological validity (comparing experimental effects and real-world effects).

we highlight salience as a pivotal concern in research methodologies, extending its relevance and applicability.

1 Theoretical Model of Experiment

A decision maker (DM) evaluates alternatives/profiles within a set $A = \{A_1, A_2, ..., A_j\}$. Each alternative is characterized by K attributes, denoted by the vector $X = (X_1, ..., X_K)$. In a typical candidate-choice conjoint experiment, the DM faces two hypothetical candidates, A_1 and A_2. For each candidate, researchers provide several attributes, such as age, gender, and job experience. For DM i, the evaluation of each attribute k is denoted by $u_k(X_k, \theta_i)$, where θ_i represents the DM's characteristics that may influence individual preferences, including factors like gender, income, education, etc. In multi-dimensional decision-making, the preference for a certain alternative depends on the overall evaluation of its attributes. In other words, the DM assigns a weight α_k to each attribute X_k. We refer to this weight α as "salience." As we will demonstrate, salience can be easily manipulated or altered by the context. The discussion on how salience changes will be deferred to the section 3. The complete evaluation of alternative A_i is thus a weighted sum as follows [3]:

$$V_i(A_j) := V_i(X) = \sum_{k=1}^{K} \alpha_k u_k(X_k, \theta_i) \tag{1}$$

where $\sum_{\alpha_k \in \alpha} \alpha_k = 1$.

Within the context of the above model, we define the Individual Treatment Effect (ITE) for attribute X_1. Consistent with standard practice, the causal effects are understood in terms of potential outcomes in a hypothetical scenario. We use Table 1 to illustrate the concept. Suppose we aim to define the ITE for attribute X_1 when it changes from x_1 to $\tilde{x}_1$, while holding the other

[3] In the model, we implicitly assume that the evaluation of attribute X_k is independent of attribute $X_{k'}$ for $k \neq k'$. More generally, however, we acknowledge the possibility that they might be correlated. As we will see later, the salience effect can induce correlation among *independent* attributes. Therefore, the current utility function does not lose generality; rather, it enables us to isolate and understand the mechanism of the salience effect.

	$World1$		$World2$	
	A_1	A_2	$\tilde{A}_1$	A_2
X_1	$\boxed{x_1}$	x_1'	$\boxed{\tilde{x}_1}$	x_1'
X_2	x_2	x_2'	x_2	x_2'
X_3	x_3	x_3'	x_3	x_3'

Table 1: Illustration of Multi-dimensional Decisions

attributes constant. This involves considering two hypothetical scenarios.

In the first scenario (World 1), the DM is presented with a choice between two candidates, A_1 and A_2. Each candidate is characterized by three attributes: X_1, X_2, and X_3. The realized values of these attributes for A_1 and A_2 are the vectors (x_1, x_2, x_3) and (x_1', x_2', x_3'), respectively. For example, if X_1 represents gender, with x_1 indicating female and x_1' indicating male, the evaluations for the two candidates in World 1 are $V_i(A_1)$ and $V_i(A_2)$.

In the second scenario (World 2), the DM is presented with the same A_2, but for A_1, the attribute X_1 is realized as $\tilde{x}_1$ instead. In other words, all attribute realizations remain the same except for X_1. We thus use $\tilde{A}_1$ to differentiate this modified version of A_1 from its original. The potential evaluations in World 2 become: $V_i(\tilde{A}_1)$ and $V_i(A_2)$.

In practice, most conjoint experiments measure the forced choice outcome Y; that is, whether a candidate is chosen or not, denoted by 1 or 0. Theoretically, the DM chooses candidate A_1 over A_2 because the DM has a higher evaluation for the former. Therefore, Y is a function of the utility difference. The corresponding utility differences in World 1 are $V_i^1 = V_i(A_1) - V_i(A_2)$, and in World 2, they are $V_i^2 = V_i(\tilde{A}_1) - V_i(A_2)$ [4]. It is evident that the binary choice potential outcome can be defined as $Y_i^j = \mathbb{1}[V_i^j \geq 0]$, where $j = 1, 2$.

Because the utility difference drives the potential outcome we observe, without loss of generality, throughout the main text, we focus on V rather than Y [5]. Consequently, the individual causal effect for attribute X_1 when it changes from x_1 to $\tilde{x}_1$, given the other attributes remain constant, is defined as the difference-in-differences $V_i^1 - V_i^2$.

[4]Note that $V_i(A_2)$ in the two hypothetical worlds may differ due to the salience effect, which we will discuss in Section 3.

[5]For outcomes that are preference scores, our results also hold because score is also a function of the evaluation. In some survey experiments, the DM may only observe one alternative. In such cases, we can simply set $V_i(A_2) = 0$.

Definition 1 (ITE)**.** Given an alternative set $A = \{A_1, \tilde{A}_1, A_2\}$ characterized by attributes $X = (X_1, X_2, ..., X_K)$. Let $X^{A_1} = (x_j, x_{-j})$ represent the attributes of A_1, where x_j is the value of attribute X_j for A_1 and x_{-j} denotes the values of all other attributes except X_j. Similarly, let $\tilde{X}^{\tilde{A}_1} = (\tilde{x}_j, x_{-j})$ and $X^{A_2} = (x'_j, x'_{-j})$ represent the attributes of $\tilde{A}_1$ and A_2. For a DM i, the individual treatment effect of attribute X_j with respect to changing its value from x_j to $\tilde{x}_j$ is defined as: $ITE^i = [V_i(X^{A_1}) - V_i(X^{A_2})] - [V_i(\tilde{X}^{A_1}) - V_i(X^{A_2})]$.

This concept is also referred to as the individual component effect (ICE) by Abramson et al. (2023). The formulation aligns with methodologies employed in the conjoint experiment literature, as discussed by Hainmueller et al. (2014b). Given that alternatives are characterized by attributes X, we also use the notation $ITE^i = [V_i(A_1) - V_i(A_2)] - [V_i(\tilde{A}_1) - V_i(A_2)]$ to denote the Individual Treatment Effect (ITE) in this paper. This example of ITE underscores the notion that, within a multi-dimensional decision-making environment, the ITE is contingent not only on the treatment attribute X_1 but also on other attributes.

It is worth noting that the formal derivation of the ITE hinges on two assumptions: Stability and No Carryover Effects, and No Profile-Order Effects, which are standard in the conjoint experiment literature. A more formal discussion is available in the Supplementary Information (SI) A. The Marginal Component Effect (MCE) is defined as a weighted sum of ITEs, and the Average Marginal Component Effect (AMCE) represents the expected value of MCE across the distribution of DMs.

Our analytical framework is not confined to conjoint experiments alone; it possesses the flexibility to be applied across a spectrum of (survey) experimental designs. For example, most experiments frequently introduce background information on a given topic, which can be methodically represented through attributes and levels. Until now, our discussion has primarily operated under the presumption that DMs across both treatment and control groups are exposed to an identical array of attributes. Nevertheless, certain designs of survey experiments diverge from this norm, especially in cases where DMs in the control group are not informed about specific attributes that are subjected to manipulation in the treatment group (for a detailed discussion, see Gaines et al.

(2007)). One common scenario involves assessing the causal impact of specific information (e.g., about corruption, denoted as X_1) on individual preferences. In such studies, DMs in both the treatment and control groups might be presented with the same foundational background information, with the distinction that only participants in the treatment group receive details concerning X_1. This specific experimental setup and its implications for attention effects are further elaborated in the SI B.

2 Attention Effect

Individuals have limited attention spans. Economists have observed that when people make decisions, such as purchasing a computer, they do not consider all possible options (Hausman, 2008). Within the literature, the subset of alternatives that a DM considers when making a decision is referred to as the "consideration set" (Hausman, 2008; Wright and Barbour, 1977). It has been well-argued and demonstrated that due to cognitive limitations, humans cannot allocate sufficient attention to every potential attribute (Stigler, 1961; Jones and Baumgartner, 2005; Chetty et al., 2009). This indicates that the consideration set is typically smaller than the full set X that may influence a decision. We denote the consideration set in real-world decision-making as $X^r \subset X$ and the consideration set in an experimental context as $X^e \subset X$.

In experiments, the consideration set of DMs is significantly influenced by the specific context of the experiment. For instance, in a candidate choice experiment, if researchers provide information only on age and gender, respondents' attention is likely focused exclusively on these two attributes. Thus, the following assumption of limited attention is upheld.

Assumption 1 (Limited Attention)**.** $X^e \subset X^r$.

The consideration set utilized in experiments, denoted as X^e, is often a subset of the real-world consideration set, X^r, for several reasons. Firstly, researchers seldom possess complete knowledge of which attributes constitute the real-world consideration set X^r. Secondly, even with such knowledge, it is impractical to present all relevant information about these attributes within

the confines of a single survey experiment.

Without loss of generality, we assume that the real-world consideration set is $X^r = \{X_1, X_2, ..., X_m\}$ and the experimental consideration set is $X^e = \{X_1, X_2, ..., X_k\}$, such that the cardinality of X^r, denoted by m, is greater than that of X^e, denoted by k ($|X^r| = m > |X^e| = k$). In other words, the consideration set for the experiments contains only the first k attributes from the real-world consideration set.

2.1 Inconsistency of Effect Magnitude

Now, let us provide an intuition for how limited attention influences the magnitude of causal effects. Suppose that in the experiment, DMs are asked to make a comparison between two alternatives, A_1 and A_2. Let's say A_1 has attributes realized as $X^{A_1} = (X_1, X_2, ..., X_k)$ and A_2 as $X^{A_2} = (X_1', X_2', ..., X_k')$, with X_1 being the treatment variable. In the control group, the realized value of X_1 is x_1^c, and in the treatment group, it is x_1^t.

In this experimental setup, individuals in the control group exhibit a certain utility difference between A_1 and A_2, which can be represented as:

$$\begin{aligned} V_c^{\text{exp}} &= V_c^{\text{exp}}(A_1) - V_c^{\text{exp}}(A_2) \\ &= \alpha_1' u_1(x_1^c, \theta_i^1) + \alpha_2' u_2(x_2, \theta_i^2) + \ldots + \alpha_k' u_k(x_k, \theta_i^k) - V_c^{\text{exp}}(A_2), \end{aligned}$$

where $\alpha_1' + ... + \alpha_k' = 1$. The superscript exp denotes the experimental condition, and the subscript c denotes the control group. Similarly, for the same individual in the treatment group, the utility evaluation can be represented as:

$$V_t^{\text{exp}} = \alpha_1' u_1(x_1^t) + \alpha_2' u_2(x_2) + \ldots + \alpha_k' u_k(x_k) - V_t^{\text{exp}}(A_2),$$

where $\sum_{j=1}^{k} \alpha_j' = 1$ (We omit θ because it will not affect the analysis.). In this discussion, we neglect the salience effect; thus, the salience α' remain constant, implying $V_c^{\text{exp}}(A_2) = V_t^{\text{exp}}(A_2)$.

Therefore, the individual-level causal effect in the experiment is defined as:

$$ITE^{\text{exp}} = V^{\text{exp}}(x_1^t) - V^{\text{exp}}(x_1^c) = \alpha_1'[u_1(x_1^t) - u_1(x_1^c)].$$

Given the assumption of no salience effect, $V(A_2)$ cancels out, illustrating that the ITE remains the same regardless of whether the DM is comparing two alternatives in conjoint experiments or focusing on a single A_1 in other survey experiments.

In the real-world context, where the consideration set includes additional attributes $\{X_{k+1}, \ldots, X_m\}$, the utility differences in the control and treatment scenarios are expressed as:

$$V_c^{\text{real}} = \alpha_1 u_1(x_1^c) + \alpha_2 u_2(x_2) + \ldots + \alpha_m u_m(x_m) - V_c^{\text{exp}}(A_2)$$

and

$$V_t^{\text{real}} = \alpha_1 u_1(x_1^t) + \alpha_2 u_2(x_2) + \ldots + \alpha_m u_m(x_m) - V_t^{\text{exp}}(A_2),$$

where $\sum_{j=1}^m \alpha_j = 1$.

Thus, in the real-world scenario, the individual-level causal effect is:

$$ITE^{\text{real}} = \alpha_1[u_1(x_1^t) - u_1(x_1^c)].$$

To compare the two causal effects without a salience effect, we establish a benchmark assumption that the salience ratio for any given dimension remains constant.

Assumption 2 (Stable Salience)**.** *Consider two decision situations in which an individual has a salience vector* $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_m)$ *and* $\alpha' = (\alpha_1, \alpha_2, \ldots, \alpha_n)$*. We say salience is stable if for any non-zero* α_j*,* α_k *and* α_j'*,* α_k' *where* $j, k \leq \min(m, n)$*, the ratio* $\frac{\alpha_j}{\alpha_k} = \frac{\alpha_j'}{\alpha_k'}$ *holds.*

This assumption reflects the natural idea that if, in one circumstance, a DM considers attribute j (e.g., gender) to be more salient than attribute k (e.g., age), that is, $\alpha_j > \alpha_k$, then in another similar circumstance, the same DM should still perceive gender as more salient than age, $\alpha_j' > \alpha_k'$,

even if the exact values of salience could differ ($\alpha_j \neq \alpha'_j$ and $\alpha_k \neq \alpha'_k$). The assumption implies that their relative magnitude should remain the same. Under this assumption, we can derive the relationship between α'_1 and α, leading to the following result.

Proposition 1 (Amplification Bias)**.** *Suppose the limited attention and stable salience assumptions 1 and 2 hold. If $u_1(x_1^t) \neq u_1(x_1^c)$, then the Individual (and Average) Treatment Effect in the experiment, compared to the real-world effect, is amplified by $\delta = \frac{1}{\alpha_1+\alpha_2+...+\alpha_k} > 1$.*

Proof. Proofs are provided in the SI. □

Generally, the experimental effect is larger than the real-world effect. The amplification factor δ depends on the total salience of the attributes (measured by the real-world salience α) included in the experiment. If DMs actually care more about attributes not included, then the experimental causal effect is significantly amplified. We summarize this in the following corollary.

Corollary 1. *The Amplification Bias in the experiment, δ, increases with the salience of attributes (measured by the real-world salience) excluded from the experiment.*

This result is intuitive. In the experiment, DMs disproportionately focus on the attributes provided by researchers, amplifying the causal effect of each attribute compared to the real-world causal effect, where the effect is diluted by other attributes. Because the real-world consideration set is not directly observable, we cannot test this theory directly. However, if our theory holds, we can derive the following testable hypothesis from the above corollary.

Hypothesis 1. *The experimental causal effect decreases as more attributes are included.*

We will test this hypothesis in the following subsection. Specifically, following common practice, the experimental causal effect is measured by the AMCE. As noted in the SI A, since AMCE is essentially a weighted sum of ITE, the amplification bias of ITE is also reflected in AMCE.

2.1.1 Test Bias Amplification Hypothesis

2.1.2 Meta Analysis

To empirically test our hypotheses, we utilize a dataset comprising 67 candidate choice conjoint and vignette experiments aggregated by Schwarz and Coppock (2022). These experiments share two core characteristics: (1) a gender attribute is randomly assigned across all experiments, and (2) the outcome variable is binary, or can be transformed into a binary choice for the candidate. This design allows us to isolate and analyze the average causal effect of the gender attribute. Additionally, we gathered data on the number of non-gender attributes presented in each experiment, noting that the minimum number of attributes provided in some experiments was as low as two.

To investigate the hypothesized relationship between the number of attributes and the AMCE of gender across these experiments, we conducted a meta-analysis. Figure 2 presents the meta-regression results, employing a random effects model, for all countries and specifically for the USA—given that a majority of the experiments were conducted in the United States. The vertical line in the figure denotes the absolute value of the estimated effect, acknowledging that some experimental causal effects of gender may manifest as negative. [6] In both analytical scenarios, there is a significant negative correlation between the number of attributes and the experimental causal effect of gender, corroborating our theoretical predictions. Excluding experiments with a small number of negative estimates and outliers does not alter the results.

2.1.3 Candidate-Choice Experiment

In order to address the inherent variability across different experiments and to directly test our hypothesis regarding the influence of attribute number on experimental effects, we designed and conducted a separate candidate-choice experiment maintaining a consistent set of attributes. Further details about the experimental design are provided in SI D. This experiment was structured around a total of 10 attributes, which were incrementally introduced across five distinct groups to

[6] Consistent with our theoretical framework, a true positive effect is expected to diminish with an increasing number of attributes, whereas a true negative effect would exhibit an increasing magnitude (because its absolute value decreases.

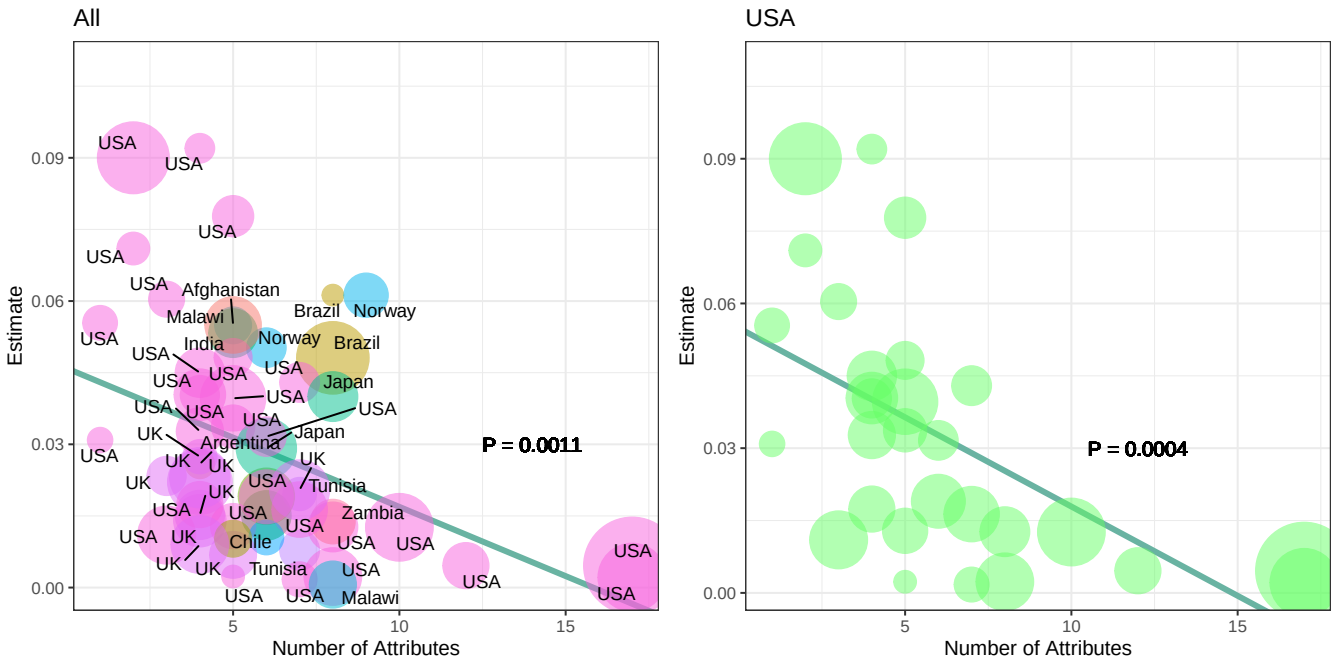

Figure 2: TEST HYPOTHESIS OF AMPLIFIED EFFECT: META-REGRESSION PLOT

systematically increase the number of attributes participants considered.

- Group 1 was presented with only two attributes: gender and age.
- Group 2 included the previous attributes plus education and tax policy, totaling four attributes.
- Group 3 added race and income to the attributes in Group 2, resulting in six attributes.
- Group 4 further included military service and religious beliefs, bringing the total to eight attributes.
- Finally, Group 5 encompassed all ten attributes by adding children and marital status to those in Group 4.

Aligning with our meta-analytical focus, we specifically examined the gender effect across these groups. The results, depicted in Figure 3, despite of 5 observations, reveal a statistically significant negative trend between the number of attributes and the magnitude of experimental

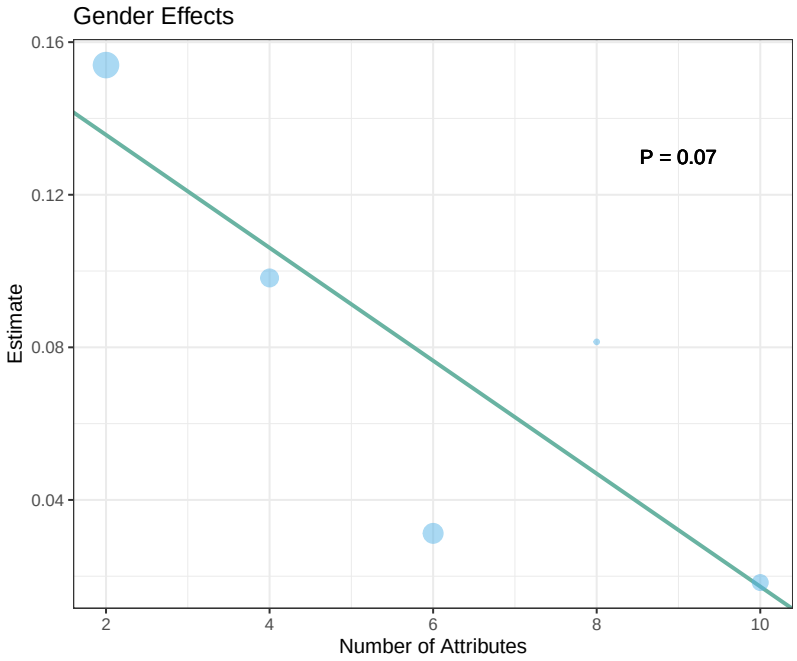

Figure 3: TEST HYPOTHESIS OF AMPLIFIED EFFECT

effects on gender. This finding supports Hypothesis 1. Interestingly, we observed that the gender effect in Group 4 (with eight attributes) was anomalously larger than anticipated. While this could be attributed to random variation, an alternative explanation may relate to the salience effect, which we will explore in subsequent sections.

3 Salience Effect

Exploring further, we investigate the distortion in the linkage between experimental causal effects and real-world implications, moving beyond the assumption of fixed salience. We adopt a psychological framework as developed by Bordalo et al. (2012) to elucidate how variations in attribute salience can influence this relationship. This approach posits that individuals disproportionately weight more salient attributes when evaluating alternatives.

We introduce the concept of salience through a function that captures the degree of emphasis an individual places on a particular attribute relative to a reference point. Consider candidates A and B, with respective attribute profiles $X^A = (x_1, x_2)$ and $X^B = (x_1', x_2')$, where we might imagine

X_1 represents gender and X_2 denotes education level. The salience of an attribute hinges on its contrast with a reference, which we define as the average attribute levels, $(\overline{X}_1, \overline{X}_2) = \left(\frac{x_1+x_1'}{2}, \frac{x_2+x_2'}{2}\right)$. Accordingly, the salience function for attribute j, denoted $\sigma(x_j, \overline{X}_j)$, is a measure of the 'distance' between attribute j and its reference point.

This function aptly captures psychological perceptions of distance. A more formal discussion is available in Bordalo et al. (2012). The function follows that for any two intervals $[x, y]$ and $[x', y']$, with $[x, y]$ fully contained within $[x', y']$, it holds that $\sigma(x, y) < \sigma(x', y')$. This aligns with the psychological principle that higher contrasts are more perceptually salient. We also assume Homogeneity of Degree Zero: The function satisfies $\sigma(\beta x, \beta y) = \sigma(x, y)$ for any positive scalar β. A viable form for σ meeting these criteria is $\sigma(x, y) = \frac{|x-y|}{y}$.

An attribute is deemed salient for a candidate if it markedly stands out compared to other aspects. For instance, $\sigma(x_1, \overline{X}_1)$ quantifies the salience of the gender attribute for candidate A, and $\sigma(x_2, \overline{X}_2)$ does so for the education attribute. The attribute with higher salience, indicated by a greater σ value, is considered more influential in the decision-making process.

Subsequently, individuals adjust their initial attribute weights α_i in the utility function $V_i = \sum_{k=1}^{K} \alpha_k u_k(X_k, \theta_i)$ based on the salience rankings, where r_i represents the rank of attribute k's salience (with lower r_k indicating higher salience). For instance, should the education attribute be more salient than gender for candidate A, i.e., $\sigma(x_1, \overline{X}_1) < \sigma(x_2, \overline{X}_2)$, we would have $r_1 = 2$ and $r_2 = 1$. The salience of each attribute then is adjusted by $\delta^{r_k - 1}$, where $\delta \in (0, 1]$ is a parameter denotes the degree of the severity of salient thinking. Suppose there are only two attributes, the utility function with adjusted salience has the following form:

$$V = \begin{cases} \frac{\alpha_1}{\alpha_1+\delta\alpha_2} u_1(x_{1j}, \theta_1) + \frac{\delta\alpha_2}{\alpha_1+\delta\alpha_2} u_2(x_{2j}, \theta_2) & \text{if } \sigma(x_{1j}, \overline{x}_1) > \sigma(x_{2j}, \overline{x}_2) \\ \frac{\delta\alpha_1}{\delta\alpha_1+\alpha_2} u_1(x_{1j}, \theta_1) + \frac{\alpha_2}{\delta\alpha_1+\alpha_2} u_2(x_{2j}, \theta_2) & \text{if } \sigma(x_{1j}, \overline{x}_1) < \sigma(x_{2j}, \overline{x}_2) \\ \alpha_1 u_1(x_{1j}, \theta_1) + \alpha_2 u_2(x_{2j}, \theta_2) & \text{if } \sigma(x_{1j}, \overline{x}_1) = \sigma(x_{2j}, \overline{x}_2) \end{cases} \tag{2}$$

Accordingly, when an attribute, such as x_1, is deemed salient, its weight in the utility function, α_1, is effectively increased to $\frac{\alpha_1}{\alpha_1+\delta\alpha_2}$, while the weight of a less salient attribute, α_2, diminishes

to $\frac{\delta\alpha_2}{\alpha_1+\delta\alpha_2}$ (The denominator makes sure that the total salience is one.). The parameter δ serves as a measure of this effect, where a value of $\delta = 1$ indicates no salience effect, implying that all attributes are weighted equally irrespective of their standout features. To encapsulate the potential for dynamic changes in attribute salience as perceived by the DM, we introduce the following assumption:

Assumption 3 (Salience Effect)**.** $\delta < 1$.

The salience model underscores the idea that the salience attributed to each feature by an individual hinges on the extent to which the realized attribute level deviates from expectations or norms, rendering it surprising or noteworthy. With the acknowledgment of salience effects, we delve into its implications, exploring how salience effect, combined with attention effect, can significantly distort the consistency between experimental and real-world effect.

3.1 Inconsistency of Effect Direction

Experimental findings can be significantly distorted due to limited attention, leading to disproportionately amplified causal effects. However, such amplification can be advantageous for researchers aiming to confirm the directionality of theoretical effects since it can reduce the required sample size for achieving a certain statistical power. One critical requirement for experimental results to meaningfully validate theoretical predictions is the consistency of effect sign: the experimental causal effect should have the same effect direction as the real-world one. However, the presence of salience effects introduces a risk of effect sign reversal, undermining the reliability of experimental outcomes as indicators of theoretical truth.

Let us still consider the ITE in an experimental context with k attributes and the treatment attribute is X_1. It is evident from equation 2 that salience is influenced by the levels of attributes. Therefore, the salience rank for attribute k in the j^{th} alternative, r_k^j, is treated as a function of the treatment, $r_k^j(X_1)$. The following proposition specifies the sufficient condition for effect sign reversal.

Proposition 2 (Effect Sign Reversal)**.** *Suppose stable salience assumptions 2 and 3 hold. Individual and average treatment effect directions are reversible if* $\exists j$ *and* k*, such that* $r_k^j(x_1) \neq r_k^j(\tilde{x}_1)$*.*

The critical condition within this proposition, "$\exists j$ and k, such that $r_k^j(x_1) \neq r_k^j(\tilde{x}_1)$," elucidates a scenario where altering the treatment level affects the salience of at least one attribute within one profile. In the accompanying proof (detailed further in the SI), it is demonstrated that such a reversal can occur irrespective of the utility magnitudes associated with excluded attributes, or the inherent salience of the treatment attribute (X_1). This finding has profound implications for experimental design and interpretation, particularly in contexts where the primary interest lies in understanding the direction of causal effects. It warns researchers of the possibility that, under the influence of salience effects, the observed direction may actually invert.

To gain intuition regarding the phenomenon of Effect Sign Reversal, consider two hypothetical worlds, as depicted in the left panel of Table 2. In Experiment 1, we compare B_1 and B_2, each with three attributes. In Experiment 2, the comparison involves two-attribute profiles A_1 and A_2. We are particularly interested in the treatment effect of attribute X_1 as it changes from x_1 to $\tilde{x}_1$. Consequently, B_2 (and A_2), serving as the controlled profiles, are fixed at $X^c = (x_1', x_2', x_3')$ (and (x_1', x_2')) respectively.

We use β_j (α_j) to denote the salience of each attribute x_j when the DM evaluates these alternatives. The corresponding sets of salience are illustrated in the right panel of Table 2. Assume the existence of prior salience for each attribute, denoted by $\beta = (\beta_1^0, \beta_2^0, \beta_3^0)$ and $\alpha = (\alpha_1^0, \alpha_2^0)$. Consider how salience is formed and evolves as the DM observes the realized attributes during comparison. When an individual compares two profiles, differing levels of each attribute will "distort" the original salience based on the rule previously mentioned. Specifically, in World 1, when comparing $B_1 = (x_1, x_2, x_3)$ to $B_2 = (x_1', x_2', x_3')$, for each attribute j, we compute the salience function $\sigma_j(\cdot, \frac{x_j + x_j'}{2})$. The original salience β^0 is then discounted by $\delta^{r_k - 1}$, where r_k represents the relative salience rank introduced earlier. We assume that the updated salience attached to B_1 follows $\beta_1 > \beta_2 > \beta_3$ and the salience attached to B_2 follows $\beta_1' > \beta_2' > \beta_3'$.

In the hypothetical World 2, only the level of attribute 1 in the treatment group is altered, from

Experiment 1			
World 1		World 2	
B_1	B_2	$\tilde{B}_1$	B_2
$\boxed{x_1}$	x_1'	$\boxed{\tilde{x}_1}$	x_1'
x_2	x_2'	x_2^t	x_2'
x_3	x_3'	x_3^t	x_3'

Experiment 1			
World 1		World 2	
β_1	β_2	$\tilde{\beta}_1$	β_2
β_1	β_1'	$\tilde{\beta}_1$	β_1'
β_2	β_2'	$\tilde{\beta}_2$	β_2'
β_3	β_3'	$\tilde{\beta}_3$	β_3'

Experiment 2			
World 1		World 2	
A_1	A_2	$\tilde{A}_1$	A_2
$\boxed{x_1}$	x_1'	$\boxed{\tilde{x}_1}$	x_1'
x_2	x_2'	x_2^t	x_2'

Experiment 2			
World 1		World 2	
α_1	α_2	$\tilde{\alpha_1}$	α_2
α_1	α_1'	$\tilde{\alpha}_1$	α_1'
α_2	α_2'	$\tilde{\alpha}_2$	α_2'

Table 2: Illustration of Salience Effect

x_1 to $\tilde{x}_1$. This modification affects the reference level for attribute X_1 and, consequently, the value of the salience function σ_1 for attribute 1. For instance, the updated salience attached to $\tilde{B}_1$ is now $\tilde{\beta}_2 > \tilde{\beta}_1 > \tilde{\beta}_3$; in other words, the salience of attribute 1 is now less than that of attribute 2. For simplicity, we assume that the salience values for the controlled profile remain the same as in the first comparison.

For Experiment 1, the utility differences in hypothetical World 1 are given by $\sum_j \beta_j u_j(x_j) - \sum_j \beta_j' u_j(x'j)$, and in hypothetical World 2, they are $\sum_j \tilde{\beta}_j u_j(\tilde{x}_j) - \sum_j \beta_j' u_j(x_j')$, where $\tilde{x}_2 = x_2$ and $\tilde{x}_3 = x_3$. Consequently, the individual treatment effect of attribute X_1 can be expressed as:

$$\begin{aligned} ITE_3 &= \sum_{j=1}^{3} [\beta_j u_j(x_j) - \tilde{\beta}_j u_j(\tilde{x}_j)] \\ &= \sum_{j=1}^{2} [\beta_j u_j(x_j) - \tilde{\beta}_j u_j(\tilde{x}_j)] + [\beta_3 u_3(x_3) - \tilde{\beta}_3 u_3(\tilde{x}_3)] \end{aligned}$$

where the subscript 3 indicates that a total of three attributes are considered in Experiment 1. Similarly, the individual treatment effect for Experiment 2 is

$$ITE_2 = \sum_{j=1}^{2} [\alpha_j u_j(x_j) - \tilde{\alpha}_j u_j(\tilde{x}_j)]$$

It is evident that ITE_3 includes an additional term $[\beta_3 u_3(x_3) - \tilde{\beta}_3 u_3(\tilde{x}_3)]$ compared to ITE_2.

Furthermore, the salience values denoted by β in ITE_3 differ from those denoted by α in ITE_2. These two factors could lead to ITE_3 and ITE_2 having different signs.

Based on Proposition 2, we formulate the following testable hypothesis:

Hypothesis 2. *The direction of attribute effects reverses with the number of attributes in a conjoint experiment.*

We do not assert that an effect sign reversal will inevitably occur. For instance, if one attribute—or its salience—predominantly overshadows others in the evaluation process, the reversal of this attribute's effect is unlikely to be observed. The reasoning behind this is straightforward: if a particular attribute (consider X_1) is significantly more important than any other attributes, it will be assigned a very large weight α_1, and, typically, it will also possess a large utility value $u(X_1)$. Therefore, in the evaluation V, even if the order of salience changes, the product $\alpha_1 u(X_1)$ remains considerably large, overshadowing other terms in the evaluation. The direct consequence is that its sign will predominantly determine the sign of V, and thus the ITE.

Another scenario that could preclude the observation of an effect sign reversal is the non-fulfillment of the rank change assumption. As highlighted by Proposition 2, the condition of salience rank change is critical for the observed result. The proposition below clarifies that without this condition, effect sign reversal cannot be observed under any circumstances.

Proposition 3. *Assuming stable salience conditions 2 and 3 are satisfied, if the relative salience rank $r_k^j(x_1) = r_k^j(\tilde{x}_1)$ for all j and k, then the effect sign reversal will not occur.*

When can we expect the salience order to remain unchanged, as presumed in the aforementioned proposition? The mechanism behind salience effects suggests that the rank of salience is contingent upon the value of the salience function, $\sigma(x_j, \overline{X}_j)$. Should the treatment attribute X_1 undergo a level change that is not substantial enough to significantly alter $\overline{X}_j$, then $\sigma(x_j, \overline{X}_j)$ will approximate $\sigma(\tilde{x}_j, \overline{X}_j)$. Consequently, the variation in salience will be insufficient to modify its relative rank. Drawing on this reasoning, the following hypothesis is proposed:

Hypothesis 3. *Attribute effect direction reversal is less likely if the level of attributes changes only marginally.*

3.1.1 Test Effect Reversal Hypotheses

We utilize data from a conjoint experiment on hotel rooms conducted by Bansak et al. (2021). They identified four core attributes of hotel rooms as follows: "view from the room (ocean or mountain view), floor (top, club lounge, or gym and spa floor), bedroom furniture (1 king bed and 1 small couch or 1 queen bed and 1 large couch), and the type of in-room wireless internet (free standard or paid high-bandwidth wireless)." In addition to these core attributes, 18 supplementary attributes, unrelated to the core four, were included in the study. Respondents were required to select their preferred hotel room from 15 pairs of profiles that varied in attributes, each encompassing the four core attributes plus a randomly selected set of other attributes. Consequently, respondents were randomly assigned to one of 11 different experimental groups, which included profiles with 4, 5, 6, 7, 8, 9, 10, 12, 14, 18, or up to 22 attributes.

Figure 4 presents a heatmap of the results. The horizontal axis represents the number of attributes, and the vertical axis denotes the level of each attribute. Dark colors indicate negative AMCEs, while light colors signify positive AMCEs. The results corroborate our hypothesis on effect sign reversal (Hypothesis 2). For instance, the effects of certain variables, such as the menu, bar, closet, or pillow, are notably unstable. However, others, including View, Towels, and Internet, demonstrate considerable stability. This observation aligns with our theoretical framework, which posits that these variables are among the most critical factors in decision-making evaluations of a hotel, possessing sufficiently high prior salience and utility to dominate over other attributes. This evidence further solidifies the notion that View and Internet are indeed core attributes, as described by Bansak et al. (2021).

To test Hypothesis 3, we adapted our experimental design for candidate choice. To control for salience, instead of randomly assigning levels to two profiles, we ensured that the level differences for each attribute were minimal. For instance, with age comprising five levels (40, 52, 60, 68, 75),

only adjacent values could appear in a comparison. If 52 was assigned to one profile, only 40 or 60 could be assigned to the other.

Table 3 displays the AMCE for gender in both the original and modified experiments. When attribute levels were assigned randomly—without controlling for salience effects—the corresponding AMCEs, as shown in the second column were noted. The third column illustrates the effects when we controlled for salience as described. A key observation is the absence of a clear effect reversal in the reduced-salience setting. This result does not suggest that researchers should assign adjacent levels for all conjoint experiments. It only used to test the hypotheses derived from the theory. The implication on research design is discussed in the section 4.

Table 3: Gender Effect

Num	Salience Effect	Reduced Salience
2	0.15	0.45
4	0.10	0.10
6	0.03	0.15
8	-0.08	0.10
10	-0.02	0.27

3.2 Inconsistency of Attribute Importance

One of the principal advantages of conjoint experiments is their ability to examine the causal effects of multiple treatments simultaneously. This methodology enables researchers to assess the importance of each attribute comprehensively. The determination of relative importance is not only crucial for theoretical development but also has practical implications for policy formulation. For instance, understanding the relative importance of attributes in a candidate choice experiment can provide real-world candidates with insights into effective strategies for election campaigns. However, this application necessitates that the significance of attributes within the experimental context aligns closely with their effects in real-world scenarios.

A key insight from the discussion on salience effects is their potential to create correlations between attributes that are theoretically independent. Such correlations can distort the causal effect

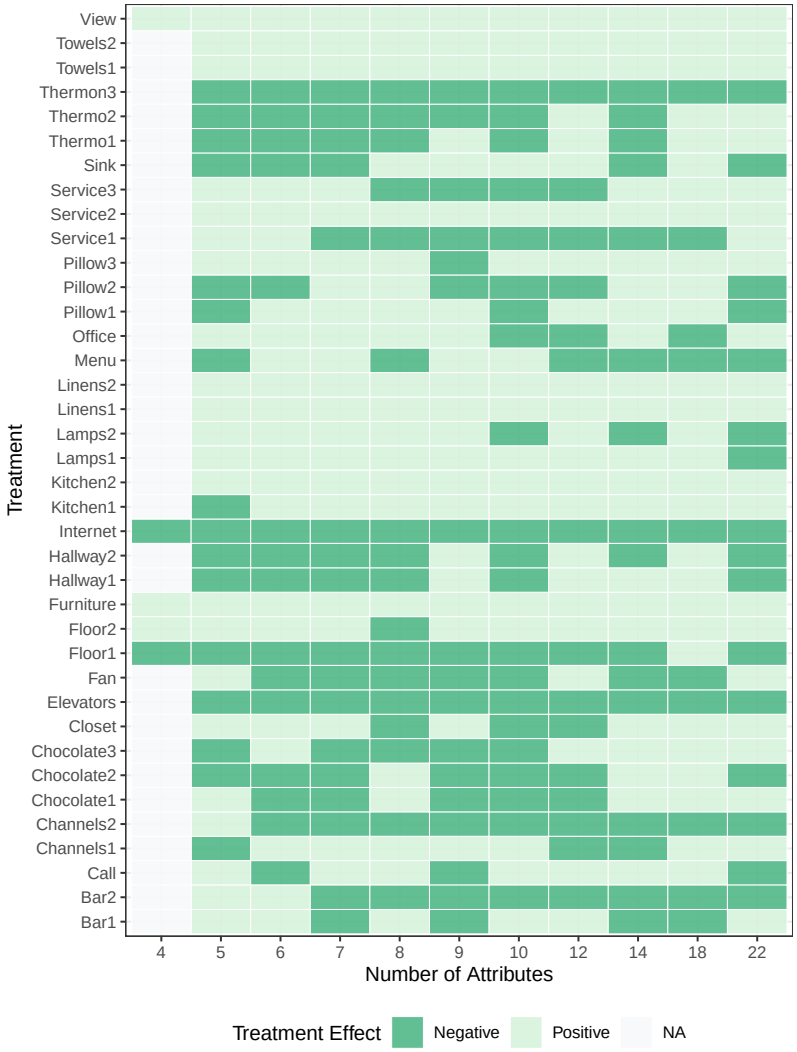

Figure 4: TESTING HYPOTHESIS OF EFFECT REVERSAL

of an attribute, leading to a reversal of the effect sign. In a similar vein, it is reasonable to anticipate that the correlation induced by salience could also alter the perceived importance of attributes.

Proposition 4 (Attribute importance Reversal)**.** *Suppose stable salience assumptions 2 and 3 hold. Attribute importance is reversible if* $\exists j$ *and* k*, such that* $r_k^j(x_1) \neq r_k^j(\tilde{x}_1)$.

In applications, researchers assess relative importance by comparing average marginal component effects, for example, in the studies conducted by Hainmueller et al. (2015) and Lehrer et al. (2024). Additionally, the use of R^2 in linear regression models offers an alternative measure of importance that circumvents the need to address the sign of the effect directly. Based on this approach, we propose the following hypothesis:

Hypothesis 4. *The perceived importance of attributes changes with the number of attributes included in a conjoint experiment.*

It is important to note, however, that not all attributes are susceptible to this phenomenon, just as not all attributes will experience a reversal in effect sign. Attributes that significantly overshadow others in terms of influence are likely to maintain their relative importance consistently.

3.2.1 Test hypothesis

To test the hypothesis, we utilize the previously mentioned data from the hotel rooms conjoint experiments. Recall that there are 11 experiment groups, featuring 4, 5, 6, 7, 8, 9, 10, 12, 14, 18, or 22 attributes. For each group, we calculate the R^2 for the attribute. To facilitate a clearer visualization of the results, we assign each attribute a rank based on its R^2 value within each group. In the heatmap 5, each column presents the relative importance rank of attributes in a specific experiment group, as listed on the x-axis. The colors denote the importance rank.

Consistent with the hypothesis, most rows (attributes) exhibit varying colors, indicating that the relative importance of most attributes changes when the number of attributes in the experiments varies. Additionally, from Section 3.1.1, we identified View, Towels, and Internet as core variables.

Our findings similarly indicate that the relative importance of these variables remains relatively stable as well. Notably, the attribute of Internet maintains its position as the most important attribute throughout, which is understandable given its perceived necessity by most individuals.

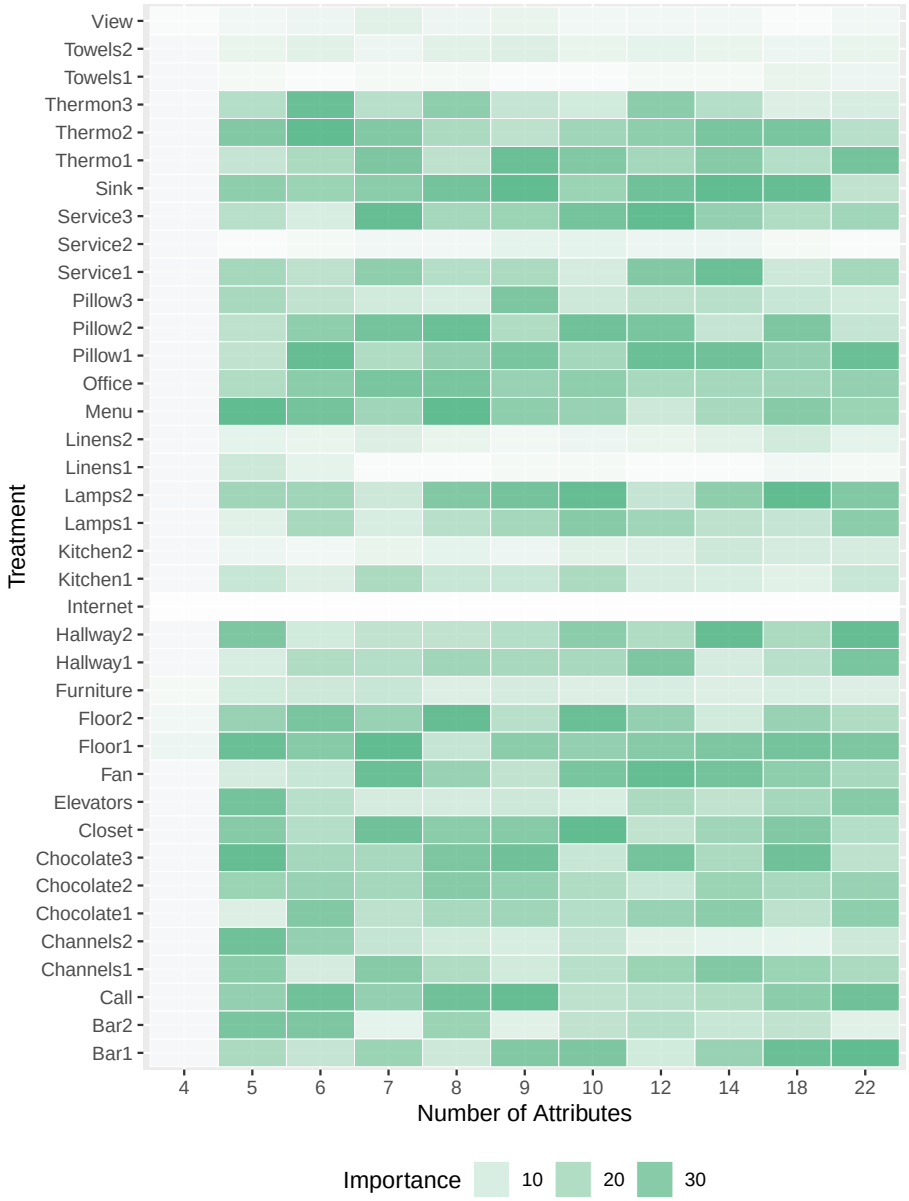

Figure 5: TESTING HYPOTHESIS OF IMPORTANCE REVERSAL

4 Discussion

We have both theoretically and empirically demonstrated that limited attention and salience effects contribute to inconsistencies between experimental effects and real-world effects, including effect magnitude, effect sign, and relative importance. In this section, we delve deeper into the correlation and differences between these two types of effects and offer recommendations for experimental design.

Firstly, it is crucial to recognize that salience effects are not inherently detrimental to experimental design. Whether individuals are making decisions in the real world or within an experimental framework, salience effects invariably play a role in multi-dimensional decision-making processes. As such, the objective should not be to eliminate salience effects from experiments. Instead, our primary concern should be to ensure that the salience effects observed in the experiment closely approximate those in the real world. It is only the artificially induced salience effects that are problematic. Note that in the absence of attention effects—implying that the experiment perfectly replicates the real-world decision-making environment—experimental ITEs would align with real-world effects due to the accurate matching of salience effects. However, the main challenge is that achieving this ideal scenario is often impractical. Furthermore, our focus on Average Effects (AMCE) rather than Individual Effects (ITE) necessitates a particular consideration. As outlined in the SI D, AMCEs are derived by marginalizing over attribute components and individuals, typically requiring participants to make several rounds of decisions among multiple profiles in a conjoint experiment. Each round features randomly generated profiles. If the randomly assigned levels are uniquely distinct and lack counterparts in the real world, this can lead to distorted salience. This observation leads to our first suggestion for experimental design.

Suggestion 1: Ensure that the attribute levels and their combinations within profiles are reflective of real-world scenarios.

Secondly, attention effects are a primary cause of the inconsistencies observed in experimental results. Our propositions are fundamentally based on the assumption of limited attention, which posits that the attributes included in an experiment does not match those attributes considered

by DMs in real-world settings. An immediate implication for research design is that researchers should endeavor to identify which attributes DMs truly value in real-life situations, i.e. match the the consideration sets in experiment and real world. As shown by Hainmueller et al. (2014a), if the consideration sets are matched, survey experiments have ecological validity. This can be achieved through a meticulous preliminary survey prior to the experimental design. It is important to highlight that this does not advocate for the inclusion of an excessive number of attributes, especially those that are unnecessary. If an unnecessary attribute exerts no influence on the evaluation function, its presence is benign. However, even a minimal effect from such an attribute, that is utility function u is not zero for this attribute, when combined with the salience effect, can significantly distort outcomes. Consider a scenario where a DM encounters an unexpected attribute in a conjoint experiment. This surprise can lead the DM to assign disproportionately high salience to this attribute, even though it may have negligible importance in a real-world context. This observation leads us to our second suggestion for experimental design:

Suggestion 2: Prior to designing an experiment, researchers should identify and include only the necessary set of attributes. Including irrelevant attributes can be detrimental.

The presence of attention and salience effects raises concerns about the reliability of average treatment effects in experiments that involve multidimensional decision processes. However, it is possible to derive valuable insights by adjusting our expectations to accommodate somewhat weaker conclusions. While a minor deviation from assumption 1 may inevitably lead to inconsistencies in effect magnitude, our primary interest often lies in understanding relative information, such as effect sign or relative importance. The discussions surrounding salience effects and the testing of hypotheses have highlighted the resilience of certain "core" attributes. These core attributes play a pivotal role in decision-making, exerting a dominant influence over others. Consequently, the effect sign and relative importance associated with these attributes tend to remain stable.

Suggestion 3: Conjoint experiments should prioritize the investigation of core attributes. This approach ensures the consistency of crucial metrics like effect sign and relative importance, even amidst the complexities introduced by attention and salience effects.

5 Concluding Remarks

An increasing number of studies employ experiments to establish causal effects within the social sciences (Druckman et al. 2006). Beyond the assurance of internal validity, external validity remains a longstanding concern. This article delves into a more fundamental aspect of generalizability: the consistency of experimental effects with the real-life effects experienced by the same set of population. We theoretically and empirically demonstrate that experiments engaging in multi-dimensional decision-making processes, especially conjoint survey experiments, may lack ecological validity. In particular, we find that the magnitude of experimental effects can be inflated, and the direction of effects and their relative importance may diverge from those observed in real-world scenarios.

Through the lens of formal theory, this study uncovers how attention and salience effects could undermine the generalizability of experimental findings. Drawing on these theoretical insights, we highlight how a careful experimental design can mitigate the influence of attention and salience, thereby enhancing the ecological validity and overall generalizability of experimental results.

References

Abramson, S. F., Koçak, K., and Magazinnik, A. (2019). What do we learn about voter preferences from conjoint experiments. *Unpublished Working Paper*.

Abramson, S. F., Kocak, K., Magazinnik, A., and Strezhnev, A. (2023). Detecting preference cycles in forced-choice conjoint experiments.

Allais, M. (1953). Le comportement de l'homme rationnel devant le risque: critique des postulats et axiomes de l'école américaine. *Econometrica: Journal of the Econometric Society*, pages 503–546.

Ascencio, S. and Gibilisco, M. B. (2015). Endogenous issue salience in an ownership model of elections.

Bansak, K., Hainmueller, J., Hopkins, D. J., and Yamamoto, T. (2021). Beyond the breaking point? survey satisficing in conjoint experiments. *Political Science Research and Methods*, 9(1):53–71.

Barabas, J. and Jerit, J. (2010). Are survey experiments externally valid? *American Political Science Review*, 104(2):226–242.

Bordalo, P., Gennaioli, N., and Shleifer, A. (2012). Salience theory of choice under risk. *The Quarterly journal of economics*, 127(3):1243–1285.

Bordalo, P., Gennaioli, N., and Shleifer, A. (2013). Salience and consumer choice. *Journal of Political Economy*, 121(5):803–843.

Bordalo, P., Gennaioli, N., and Shleifer, A. (2016). Competition for attention. *The Review of Economic Studies*, 83(2):481–513.

Carnes, N. and Lupu, N. (2016). Do voters dislike working-class candidates? voter biases and the descriptive underrepresentation of the working class. *American Political Science Review*, 110(4):832–844.

Chetty, R., Looney, A., and Kroft, K. (2009). Salience and taxation: Theory and evidence. *American economic review*, 99(4):1145–77.

Clayton, K., Horiuchi, Y., Kaufman, A. R., King, G., and Komisarchik, M. (2023). Correcting measurement error bias in conjoint survey experiments. Technical report, Working paper, URL: https://tinyurl. com/24btw3dq.

Cook, T. D., Campbell, D. T., and Shadish, W. (2002). *Experimental and quasi-experimental designs for generalized causal inference*, volume 1195. Houghton Mifflin Boston, MA.

De la Cuesta, B., Egami, N., and Imai, K. (2022). Improving the external validity of conjoint analysis: The essential role of profile distribution. *Political Analysis*, 30(1):19–45.

De Mesquita, E. B. and Tyson, S. A. (2020). The commensurability problem: Conceptual difficulties in estimating the effect of behavior on behavior. *American Political Science Review*, 114(2):375–391.

Dragu, T. and Fan, X. (2016). An agenda-setting theory of electoral competition. *The Journal of Politics*, 78(4):1170–1183.

Druckman, J. N., Green, D. P., Kuklinski, J. H., and Lupia, A. (2006). The growth and development of experimental research in political science. *American Political Science Review*, 100(4):627–635.

Egami, N. and Hartman, E. (2023). Elements of external validity: Framework, design, and analysis. *American Political Science Review*, 117(3):1070–1088.

Eggers, A. C., Vivyan, N., and Wagner, M. (2018). Corruption, accountability, and gender: Do female politicians face higher standards in public life? *The Journal of Politics*, 80(1):321–326.

Ellsberg, D. (1961). Risk, ambiguity, and the savage axioms. *The quarterly journal of economics*, 75(4):643–669.

Fu, J. and Slough, T. (2023). Heterogeneous treatment effects and causal mechanisms. Technical report, Working Paper.

Gaines, B. J., Kuklinski, J. H., and Quirk, P. J. (2007). The logic of the survey experiment reexamined. *Political Analysis*, 15(1):1–20.

Guthrie, C., Rachlinski, J. J., and Wistrich, A. J. (2000). Inside the judicial mind. *Cornell L. Rev.*, 86:777.

Hainmueller, J., Hangartner, D., and Yamamoto, T. (2014a). Do survey experiments capture real-world behavior? external validation of conjoint and vignette analyses with a natural experiment. *Proceedings of the National Academy of Sciences*, 112(8):2395–2400.

Hainmueller, J., Hangartner, D., and Yamamoto, T. (2015). Validating vignette and conjoint survey experiments against real-world behavior. *Proceedings of the National Academy of Sciences*, 112(8):2395–2400.

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

Ham, D. W., Imai, K., and Janson, L. (2022). Using machine learning to test causal hypotheses in conjoint analysis. *arXiv preprint arXiv:2201.08343*.

Hausman, D. (2008). Mindless or mindful economics: A methodological evaluation. *The foundations of positive and normative economics: A handbook*, pages 125–55.

Horiuchi, Y., Markovich, Z., and Yamamoto, T. (2022). Does conjoint analysis mitigate social desirability bias? *Political Analysis*, 30(4):535–549.

Huang, M. (2022). Sensitivity analysis in the generalization of experimental results. *arXiv preprint arXiv:2202.03408*.

Jones, B. D. and Baumgartner, F. R. (2005). *The politics of attention: How government prioritizes problems*. University of Chicago Press.

Kihlstrom, J. F. (2021). Ecological validity and "ecological validity". *Perspectives on Psychological Science*, 16(2):466–471.

Lehrer, R., Stöckle, P., and Juhl, S. (2024). Assessing the relative influence of party unity on vote choice: evidence from a conjoint experiment. *Political Science Research and Methods*, 12(1):220–228.

List, J. A. and Levitt, S. D. (2005). What do laboratory experiments tell us about the real world. *NBER working paper*, pages 14–20.

Masatlioglu, Y., Nakajima, D., and Ozbay, E. Y. (2016). Revealed attention. In *Behavioral Economics of Preferences, Choices, and Happiness*, pages 495–522. Springer.

Mullinix, K. J., Leeper, T. J., Druckman, J. N., and Freese, J. (2015). The generalizability of survey experiments. *Journal of Experimental Political Science*, 2(2):109–138.

Riker, W. H., Riker, W. H., and Riker, W. H. (1986). *The art of political manipulation*, volume 587. Yale University Press.

Sanbonmatsu, K. (2002). Gender stereotypes and vote choice. *american Journal of political Science*, pages 20–34.

Schwarz, S. and Coppock, A. (2022). What have we learned about gender from candidate choice experiments? a meta-analysis of sixty-seven factorial survey experiments. *The Journal of Politics*, 84(2):655–668.

Slough, T. (2023). Phantom counterfactuals. *American Journal of Political Science*, 67(1):137–153.

Slough, T. and Tyson, S. A. (2023). External validity and meta-analysis. *American Journal of Political Science*, 67(2):440–455.

Stefanelli, A. and Lukac, M. (2020). Subjects, trials, and levels: Statistical power in conjoint experiments.

Stigler, G. J. (1961). The economics of information. *Journal of political economy*, 69(3):213–225.

Viscusi, W. K. (2001). Jurors, judges, and the mistreatment of risk by the courts. *The Journal of Legal Studies*, 30(1):107–142.

Wright, P. and Barbour, F. (1977). *Phased decision strategies: Sequels to an initial screening*. Graduate School of Business, Stanford University.

Supplementary Information

A Formal Definition of Causal Targets

We follow Abramson et al. (2023) to develop the Average Marginal Component Effect (AMCE) and the corresponding assumptions.

In practice, decision-makers (DMs) complete R rounds of choice tasks. In each round, they evaluate J profiles, each with K attributes. The treatment assigned to DM i for the jth profile in the rth task is the K-dimensional vector T_{ijr}. For example, in round $r = 1$, $T_{ijr} = T_{i11} = (x_1, x_2, x_3)$ and $T_{i21} = (x_1', x_2', x_3')$. T_i denotes the set of treatments across all profiles and tasks. Thus, in a two-round ($R = 2$) candidate choice experiment with $J = 2$, $T_i = \{T_{i11}, T_{i21}, T_{i12}, T_{i22}\}$.

The potential outcome in round r for profile j, denoted by $Y_{ijr}(T_i)$, is a function of T_i rather than just T_{ijr}, without further assumptions.

The first assumption is analogous to STUVA. Let $T_{i(j)r}$ be the set that contains profiles in that round, $T_{i(j)r} = \{T_{i1r}, ..., T_{iJr}\}$.

Assumption 4 (Stability and No Carryover)**.**

$$Y_{ijk}(T_i) = Y_{ijk'}(T_i')$$

if $T_{i(j)r} = T'_{i(j)r'}$ *for any* $j, r,$ *and* r'.

This assumption implies that potential outcomes in round r depend only on profiles in that round. Thus, we can express potential outcomes in terms of only the treatment profiles assigned in the relevant task, denoted by $Y_{ijr}(T_{i(j)r})$.

The next assumption allows us to further simplify the relationship between potential outcomes and the set of profiles.

Assumption 5 (No Profile Order Effects)**.**

$$Y_{ijr}(T_{i(j)r}) = Y_{ij'r}(T'_{i(j')r})$$

if $T_{ijr} = T'_{ij'r}$, $T_{ij'r} = T'_{ijr}$, $T_{ijr} \neq T_{ij'r}$, $T'_{ijr} \neq T'_{ij'r}$ *for any* i, j, j', r.

This means that simply shuffling the order of profiles in round r does not affect potential outcomes as long as the profiles remain the same. Under this assumption, we can write the potential outcome as $Y_{ijr}(T_{ijr}, T_{i[-j]r})$, where $T_{i[-j]r}$ is the unordered set of non-jth profiles.

To focus on a specific attribute, it is useful to further decompose T_{ijr}. We define T_{ijrk} as the k-th attribute for profile j in round r, and $T_{ijr[-k]}$ is the set of remaining attributes. Hence, the potential outcome can be expressed as $Y_{ijr}(T_{ijrk}, T_{ijr[-k]}, T_{i[-j]r})$. The two assumptions imply that the round r and profile j are irrelevant in most tasks, so we omit them in the potential outcome notation.

We can now define the Individual Component Effect (ICE) as follows:

Definition 2 (Individual Component Effect)**.** Suppose $T_{ijrk} = t_k$, $T_{ijrk'} = t_{k'}$, $T_{ijr[-k]} = t_{-k}$, $T_{i[-j]r} = \mathbf{t}$,

$$ICE_i = Y_i(t_k, t_{-k}, \mathbf{t}) - Y_i(t_{k'}, t_{-k}, \mathbf{t})$$

This effect reflects the change in an individual's response when the k-th attribute changes while holding the other attributes constant.

In practice, most potential outcomes Y are binary forced choices, represented by 0 or 1. Theoretically, an individual chooses profile j over j' in task r because the utility for profile j is higher than that for the latter. Hence, Y is a function of the utility differences, as defined in the main text. Formally, given two profiles j and j', the binary potential outcome Y can be represented by $Y_{ij} = \mathbb{1}[V_{ij}(t_k, t_{-k}, \mathbf{t}) - V_{ij'}(t_{k'}, t_{-k}, \mathbf{t}) \geq 0]$, where $\mathbb{1}$ is the indicator function.

Researchers typically summarize the component effect across different profile combinations in experiments. Generally, we can marginalize over the distribution of t_{-k} and $\mathbf{t}$.

Definition 3 (Marginal (Individual) Component Effect)**.**

$$MCE_i = \sum_{(t_{-k}, \mathbf{t}) \in T} [Y_i(t_k, t_{-k}, \mathbf{t}) - Y_i(t_{k'}, t_{-k}, \mathbf{t})] \times \mathbb{P}[T_{ijr[-k]} = t_{-k}, T_{i[-j]r} = \mathbf{t} | T_{ijr[-j]}, T_{i[-j]r} \in \tilde{T}]$$

where $\tilde{T}$ is the intersection of the supports of $\mathbb{P}[T_{ijr[-k]} = t_{-k}, T_{i[-j]r} = \mathbf{t} | T_{ijrk} = t_k]$ and

$\mathbb{P}[T_{ijr[-k]} = t_{-k}, T_{i[-j]r} = \mathbf{t} | T_{ijrk} = t_{k'}]$.

Finally, the Average Marginal Component Effect (AMCE) is the expectation of MCE taken over the distribution of DMs.

Definition 4 (AMCE)**.**

$$
\begin{aligned}
AMCE &= \mathbb{E}[MCE_i] \\
&= \sum_{(t_{-k}, \mathbf{t}) \in T} \mathbb{E}[Y_i(t_k, t_{-k}, \mathbf{t}) - Y_i(t_{k'}, t_{-k}, \mathbf{t})] \times \mathbb{P}[T_{ijr[-k]} = t_{-k}, T_{i[-j]r} = \mathbf{t} | T_{ijr[-j]}, T_{i[-j]r} \in \tilde{T}]
\end{aligned}
$$

B Diminish effect due to attention

In the context of unbalanced experimental designs, where the control group lacks exposure to treatment attributes, a nuanced issue arises in the estimation of causal effects. Gaines et al. (2007) critically evaluates the logic underpinning survey experiments, suggesting the potential pitfalls of a control setup devoid of treatment attribute information.

Based on the discussion in section 3, the utility function for the control group is

$$V_i^{exp}(x_1^c) = \alpha_2' u_2(x_2, \theta_2) + ... + \alpha_k' u_k(x_2, \theta_2).$$

For individuals in the treatment group, their utility is represented by:

$$V_i^{exp}(x_1^t) = \alpha_1'' u_1(x_1, \theta_1) + \alpha_2'' u_2(x_2, \theta_2) + ... + \alpha_k'' u_k(x_2, \theta_2)$$

The individual causal effect of X_1 in the experiment is

$$V_i^{exp}(x_1^t) - V_i^{exp}(x_1^c) = \alpha_1'' u_2(x_1, \theta_1) + \sum_j (\alpha_j'' - \alpha_j') u_j(x_2, \theta_2).$$

This formulation uncovers at least two significant challenges. First, The objective is to isolate the causal effect attributed solely to the first dimension X_1 only. However, the term $\sum_j (\alpha_j'' -$

$\alpha'_j)u_j(x_2, \theta_2)$ introduces a bias that undermines this objective by capturing the differential impact of other dimensions, thereby diluting the intended causal effect of X_1. Second, the term $\alpha''_1 u_2(x_1, \theta_1)$ is meant to capture the causal effect of X_1, considering the salience α'' in the treatment group. Yet, the absence of a counterfactual in the control group complicates the interpretation of how the effect X_1, modulated by salience, contrasts with untreated scenarios.

C Proof of Proposition 1

Proof. Recall, $ITE^{exp} = V^{exp}(x_1^t) - V^{exp}(x_1^c) = \alpha_1'[u_1(x_1^t) - u_1(x_1^c)]$ and $ITE^{real} = \alpha_1[u_1(x_1^t) - u_1(x_1^c)]$.

The key in the proof is to re-write salience in the smaller set (exp group) α' in terms of salience in the larger set α (real world.) Under stable salience assumption 2, we can write

$$\alpha_i = (1 - \sum_{j=k+1}^{m} \alpha_j)\alpha_i' \quad \forall i \in [1, 2, ..., k].$$

Then, because of $u_1(x_1^t) \neq u_1(x_1^c)$, $\delta = \frac{ITE^{exp}}{ITE^{real}} = \frac{\alpha_1'}{\alpha_1} = \frac{1}{\sum_{j=1}^{k} \alpha_j}$. The ATE follows because of the random assignment of the treatment and $ATE = \mathbb{E}[ITE]$.

□

D Design of Candidate-Choice Experiment

We conduct two candidate-choice experiment to test hypotheses outlined in the main text.

Part 1:

Respondents see a conjoint table with a hypothetical candidate that is described by K (=2,4,6,8,10) attributes as shown in the example table below:

We randomly assign participants into 5 groups. The total number of participants is decided by budget. For each attribute, levels are also randomly assigned.

- Group 1: each hypothetical candidate has K=2 attributes including Gender and Age; there are total of 6 rounds for each participant
- Group 2: each hypothetical candidate has K=4 attributes including Gender, Age, Education, and Tax; there are total of 6 rounds for each participant
- Group 3: each hypothetical candidate has K=6 attributes including Gender, Age, Education, Tax, Race, and Income; there are total of 6 rounds for each participant
- Group 4: each hypothetical candidate has K=8 attributes including Gender, Age, Education, Tax, Race, Income, Religion, and Military service; there are total of 6 rounds for each participant
- Group 5: each hypothetical candidate has K=10 attributes including Gender, Age, Education, Tax, Race, Income, Religion, Military service, Gay Marriage, and Children; there are total of 6 rounds for each participant

Attribute level:

Part 2:

Respondents are required to complete another 5 rounds of candidate choice experiment. However, we control attribute salience in each pair. In particular, we let some attributes be randomly chosen so that their levels are close to each other. For example, for attribute age = [40,52,60,68,75],

Please carefully review the two candidates for President detailed below. Then please answer the questions about these two candidates below.

	Candidate 1	**Candidate 2**
Gender	Female	Female
Age	68	75
Education	BA from College	BA from College
Increase tax to Rich	Indifference	Opposite
Religion	Mormon	Catholic
Race	Black	White
Military	Served	Not served
Income	65,000	54,000

Which of these two candidates would you prefer to see as President of the United States?

- ◯ Candidiate 1
- ◯ Candidate 2

On a scale from 1 to 10, where 1 indicates that you would never support this candidate, and 10 indicates that you would always support this candidate, where would you place for

Attribute	Levels
Gender	Male, Female
Age	40,52,60,68,75
Education	BA from College; BA from State University; BA from Ivy League University
Increase tax to the rich	Support, Indifference, Opposite
Religion	Mormon, Catholic, Protestant
Race	White, Black, Asian American
Military Service	Served, Not served
Income	32000, 54000, 65000, 92000, 210000, 5 .1 million
Same-sex Marriage	Support, Indifference, Opposite
Number of Children	0 ,1, 2, 3, 4

Figure A.2: ATTRIBUTE TABLE

only levels (40,52) , (52,60), (60,68), and (68,75) will be randomly chosen. Those levels are age, income, same sex marriage, tax, education, and children.

- If K=2, attributes are Gender and Age
- If K=4, attributes are Gender, Age, Tax, and Education.
- If K=6, attributes are Gender, Age, Tax, Education, Race, and Income.
- If K=8, attributes are Gender, Age, Tax, Education, Race, Income, Military, and Religion.
- If K=10, attributes are Gender, Age, Tax, Education, Race, Income, Military, Religion, Same Sex Marriage, and Children.

E Proof of Proposition 2

Proof. Without loss of generality, we consider the same setting as depicted in Table 2. This entails comparing the change in effect sign between scenarios with two and three attributes. For instances involving more than two attributes, we can aggregate these attributes into a single composite attribute with a composite salience (thanks to linearity), and the proof follows analogously.

In the proposition, the condition is '$\exists k$ such that $r_k^j(x_1) \neq r_k^j(\tilde{x}_1)$'. According to the salience mechanism, varying levels of X_1 influence its salience function and relative rank r_1. Hence, we can confidently assume $r_1^1(x_1) \neq r_1^1(\tilde{x}_1)$. Under this premise, as detailed in the main text, the Individual Treatment Effect (ITE) for two attributes is

$$ITE_2 = [\alpha_1 u_1(x_1) - \tilde{\alpha}_1 u_1(\tilde{x}_1)] + (\alpha_2 - \tilde{\alpha}_2) u_2(x_2)$$

and the ITE for three attributes is

$$ITE_3 = [\beta_1 u_1(x_1) - \tilde{\beta}_1 u_1(\tilde{x}_1)] + (\beta_2 - \tilde{\beta}_2) u_2(x_2) + (\beta_3 - \tilde{\beta}_3) u_3(x_3)$$

See the differences between the first term:

$$(\alpha_1 - \beta_1) u_1(x_1) + (\tilde{\beta}_1 - \tilde{\alpha}_1) u_1(\tilde{x}_1) \tag{3}$$

Without loss of generality, we normalize utility to be positive. By stable salience assumption, $\beta_1^0 = (1 - \beta_3^0)\alpha_1^0$ and $\beta_2^0 = (1 - \beta_3^0)\alpha_2^0$.

Note that $\alpha_1 = \frac{\alpha_1^0}{\alpha_1^0 + \delta\alpha_2^0} > \beta_1 = \frac{\beta_1^0}{\beta_1^0 + \delta\beta_2^0 + \delta^2\beta_3^0} = \frac{\alpha_1^0}{\alpha_1^0 + \delta\alpha_2^0 + \frac{\beta_3^0}{1-\beta_3^0}\delta^2}$ and thus $\alpha_1 - \beta_1 > 0$; similarly, $\tilde{\beta}_1 - \tilde{\alpha}_1 < 0$.

Because the proposition concerns the existence of a particular scenario, it suffices to identify a single instance in which the effect sign reverses.

Case 1: Suppose $(\alpha_1 - \beta_1) u_1(x_1) + (\tilde{\beta}_1 - \tilde{\alpha}_1) u_1(\tilde{x}_1) = 0$.

The second part in ITE_2 is negative because $\alpha_2 = \frac{\delta\alpha_2^0}{\alpha_1^0 + \delta\alpha_2^0} < \frac{\alpha_2^0}{\delta\alpha_1^0 + \alpha_2^0} = \tilde{\alpha}_2$, recall $\delta \in (0, 1)$ by

assumption. Now, WLOG, let us assume $ITE_2 < 0$. Therefore, we need to find conditions such that $ITE_3 \geq 0$.

We know the second part in ITE_3, $(\beta_2 - \tilde{\beta}_2)u_2(x_2)$ must be negative because $\beta_2 = \frac{\delta\beta_2^0}{\beta_1^0+\delta\beta_2^0+\delta^2\beta_3^0} < \tilde{\beta}_2 = \frac{\beta_2^0}{\delta\beta_1^0+\beta_2^0+\delta^2\beta_3^0}$. Thus, in order to $ITE_3 \geq 0$, we can let $\beta_3 > \tilde{\beta}_3$. Because $\beta_3 = \frac{\delta^2\beta_3^0}{\beta_1^0+\delta\beta_2^0+\delta^2\beta_3^0}$ and $\tilde{\beta}_3 = \frac{\delta^2\beta_3^0}{\delta\beta_1^0+\beta_2^0+\delta^2\beta_3^0}$, simple algebra shows that

$$\begin{cases} \beta_3 > \tilde{\beta}_3 \quad if \quad \beta_1^0 < \beta_2^0 \\ \beta_3 < \tilde{\beta}_3 \quad if \quad \beta_1^0 > \beta_2^0 \end{cases}$$

This means that, $ITE_3 \geq 0$ if and only if $u_3(x_3) \geq \frac{\tilde{\beta}_2 - \beta_2}{\beta_3 - \tilde{\beta}_3} u_2(x_2)$ and $\beta_1^0 > \beta_2^0$.

Here is another possible situation.

Case 2: Suppose $(\alpha_1 - \beta_1)u_1(x_1) + (\tilde{\beta}_1 - \tilde{\alpha}_1)u_1(\tilde{x}_1) < 0$ and $ITE_2 > 0$.

We need find conditions such that $ITE_3 \leq 0$. We notice the first term in ITE_3 is positive and the second term is negative.

Now suppose $\beta_1^0 < \beta_2^0$ so that $\tilde{\beta}_3 > \beta_3$. Therefore, in order to $ITE_3 \leq 0$, $u_3(x_3)$ must be close to 0.

As demonstrated in the proof, there exists considerable flexibility to achieve the desired outcome. Identifying just one scenario that aligns with our hypothesis is sufficient.

□

F Proof of Proposition 3

Proof. Following the notational conventions established in the Proof of Proposition 2, let us consider comparing the effect sign in scenarios with m and n attributes, where $m < n$.

The differential in evaluation for the world with treatment x_1 is expressed as: $\sum_{j=1}^{n} \beta_j u_j(x_j) - \beta'_j u_j(x'_j)$, and for the world with treatment $\tilde{x}_1$, it is: $\sum_{j=1}^{n} \beta_j u_j(\tilde{x}_j) - \beta'_j u_j(x'_j)$, with the assumption that $\tilde{x}_{-1} = x_{-1}$.

Given that there is no change in the salience ranking, in accordance with the salience mecha-

nism, the salience vectors β and β' remain consistent across the two worlds despite varying levels of treatment for X_1.

Hence, the calculation of ITE becomes straightforward,

$$ITE_m = \beta_1[u_1(x_1) - u_1(\tilde{x}_1)]$$

and

$$ITE_n = \alpha_1[u_1(x_1) - u_1(\tilde{x}_1)]$$

Given salience are positive, we conclude that the sign of ITE_m and ITE_m must be the same.

□

G Proof of Proposition 4

Proof. In the scenario described, where we observe a shift in treatment value from x_1^t to $\tilde{x}_1^t$, resulting in a decreased updated salience for the treatment profile, specifically $\tilde{\beta}_1^t < \tilde{\beta}_2^t$, while the salience remains unchanged when x_2^t transitions to $\tilde{x}_2^t$, and salience for the control profile remains constant, our analysis can be simplified.

For the two-attribute case, the Individual Treatment Effect (ITE) for the first attribute can be articulated as follows:

$$ITE_2^1 = [\alpha_1 u_1(x_1) - \tilde{\alpha}_1 u_1(\tilde{x}_1)] + (\alpha_2 - \tilde{\alpha}_2)u_2(x_2)$$

and the ITE for the second attribute is

$$ITE_2^2 = \alpha_2[u_2(x_2) - u_2(\tilde{x}_2)]$$

Similarly, we get two ITEs in the three-attributes case,

$$ITE_3^1 = [\beta_1 u_1(x_1) - \tilde{\beta}_1 u_1(\tilde{x}_1)] + (\beta_2 - \tilde{\beta}_2) u_2(x_2) + (\beta_3 - \tilde{\beta}_3) u_3(x_3)$$

and

$$ITE_3^2 = \beta_2 [u_2(x_2) - u_2(\tilde{x}_2)]$$

Again, we assume there is $u_1(\tilde{x_1})$ such that $(\alpha_1 - \beta_1)u_1(x_1) + (\tilde{\beta}_1 - \tilde{\alpha}_1)u_1(\tilde{x}_1) = 0$.

Now, we consider the case that $|ITE_2^1| > |ITE_2^2|$. We hope to find conditions such that $|ITE_3^1| \leq |ITE_3^2|$.

Without loss of generality, we further assume $u_2(x_2) > u_2(\tilde{x}_2)$ so that both ITE_2^2 and ITE_3^2 are positive.

Recall $\beta_2 < \tilde{\beta}_2$, therefore, we can see if $\beta_3 > \tilde{\beta}_3$, which means $\beta_1^0 < \beta_2^0$, we can find $u_3(x_3) \in [\frac{(\tilde{\beta}_2 - \beta_2)u_2(x_2) - ITE_3^2}{\beta_3 - \tilde{\beta}_3}, \frac{ITE_3^2 + (\tilde{\beta}_2 - \beta_2)u_2(x_2)}{\beta_3 - \tilde{\beta}_3}]$ satisfies the condition.

□