9-24-2022

Best practices for implementing experimental research methods

Jennifer L. Stoner
University of North Dakota, jennifer.stoner@UND.edu

1 Best practices for implementing experimental research methods in consumer studies

Abstract

Experimental research methods have a long history across a number of different disciplines—including consumer research. Although experiments are just one of many alternative research methods, experiments are notable because they are the best way establish causation. This makes experiments a powerful tool when researchers need to show cause and effect relationships. In this paper, we provide best practices for implementing experimental research methods in consumer studies. Specifically, we discuss several important topics researchers need to consider when designing experiments, including developing hypotheses, operationalizing the variables (manipulated or measured), deciding on the research design (between-subjects, within-subjects, or mixed), selecting the research setting (laboratory, field, or online), understanding the main effect (via moderation, mediation, or moderated mediation), including manipulation and attention checks, determining the sample size, and choosing participants. We provide recommendations that researchers can use to conduct high-quality experiments in a consumer context.

Keywords

Experiments, design, moderation, mediation, manipulation checks, attention checks, sample

2 INTRODUCTION

Experiments are becoming increasingly important in academic research (Koschate-Fischer & Schandelmeier, 2014; Viglia et al., 2021) and top consumer research journals often require a series of experiments for publication (Pechman, 2019). This may be because many researchers deem experiments "the gold standard in scientific research" (Kardes & Herr, 2019, p. 4), given that experiments are the best way to establish causation (Kirk, 2013). Specifically, researchers test hypotheses by manipulating the independent variable (IV) and measuring its impact on the dependent variable (DV) (Koschate-Fischer & Schandelmeier, 2014). This allows researchers to draw conclusions about cause (IV) and effect (DV) relationships. In other words, changes in the IV result in changes in the DV. Thus, experiments are appropriate when researchers need to show cause and effect relationships.

Importantly, researchers should randomly assign participants to experimental conditions in order to make claims about causation (Kardes & Herr, 2019; Viglia et al., 2021). With random assignment, all participants have the same chance of being in each of the experimental conditions. The goal of random assignment is to create equivalent groups of participants that do not differ in terms of extraneous variables (e.g., individual difference variables). In this way, random assignment controls for extraneous variables (or systematic error) without the need to measure these variables (Kardes & Herr, 2019). Because the only thing that differs across experimental conditions is the level of the IV, results between experimental conditions should be equivalent in the absence of the IV (Kardes & Herr, 2019). Thus, random assignment enables researchers to show that only the manipulation of the IV leads to changes in the DV.

In the following sections, we provide guidance on the decisions that researchers need to make in order to design a well thought out experiment (see Figure 1 for an overview of topics). In the next section, we discuss the importance of developing testable hypotheses. Then, we examine operationalizing the variables as either manipulated or measured variables in Section 3. In Section 4, we compare various research designs, such as between-subjects, within-subjects, and mixed designs. Similarly, we compare various research settings, such as laboratory, field, and online settings in Section 5. In Section 6, we explain how researchers can enhance their understanding of the main effect via moderation, mediation, and moderated mediation. Next, we discuss the merits of including manipulation and attention checks in Section 7. Finally, in Section 8, we provide guidance on determining sample size and selecting participants. We provide recommendations based on best practices—as well as corresponding exemplars from top tier marketing journals that illustrate these best practices—throughout the remainder of the paper, which consumer researchers can use to conduct high-quality experiments.

Insert Figure 1 about here

----

3 DEVELOPING HYPOTHESES

Because experiments test hypotheses that predict relationships between variables (Koschate-Fischer & Schandelmeier, 2014), researchers should develop testable hypotheses based on strong theory. We agree with Van de Ven (1989, p. 486) and others who stress that "nothing is quite so practical as a good theory," as good theory enables researchers to contribute to both theory and practice in a meaningful way. Thus, experiments should start with theoretically sound constructs and hypotheses. For consumer studies, Pechman (2019) recommends developing two to five interesting and important hypotheses, which researchers can introduce either informally or formally in the text. Hypotheses can test main effects (the impact of the IV on the DV), moderation ("when" the IV causes the DV), or mediation ("how" the IV causes the DV) (Janiszewski et al., 2016). (We discuss moderation and mediation further in Section 6.) As an example, Han et al. (2021) provide a strong theoretical foundation for their work on heritage branding and test several hypotheses that predict a main effect, moderation, and mediation. Specifically, the authors show how heritage branding decreases evaluations of enhanced products (main effect). Importantly, this effect only emerges when heritage branding uses longevity cues (moderator). Finally, a decrease in perceived continuity causes this effect (mediator).

Although most researchers report hypotheses using a deductive approach, where they develop hypotheses based on existing theory as we describe above, Janiszewski and van Osselaer (2021) argue that research is often inductive or abductive, where researchers develop or revise hypotheses based on data. As a result, the authors propose a more flexible reporting template that encourages exploratory studies to develop hypotheses, constructs, and new theory. Thus, when researchers run exploratory studies, which can be especially helpful in the early stages of experimental research, they should consider Janiszewski and van Osselaer's (2021) recommendation to report exploratory studies and the role they played in developing hypotheses.

4 OPERATIONALIZING VARIABLES: MANIPULATED OR MEASURED

Before we discuss operationalizing variables, it is important to distinguish between defining constructs and operationalizing variables (Calder et al., 1981; 2021). After researchers establish a strong theoretical foundation to define their constructs and develop their hypotheses, they need to think about how to operationalize the variables required to test their hypotheses. Because one of the decisive characteristics of experiments is the manipulation of the IV, researchers need to determine the number of levels of the IV and manipulate the IV accordingly. Moderators (which we discuss in Section 6) can be either manipulated or measured. Mediators (which we also discuss in Section 6) are typically measured using continuous scales. That said, some scholars advocate for manipulating mediators in a series of experiments to establish a causal chain, an approach that helps researchers demonstrate a causal (vs. correlational) relationship between the IV and DV (see Pirlott & MacKinnon, 2016 and Spencer et al., 2005 for further discussion and guidance). Finally, DVs (which we discuss in more detail in the next paragraph) are typically measured using either categorical or continuous scales. As an example of a study that includes an IV, moderator, mediator, and DV, Blank and Bolton (2019) manipulated psychological distance (IV) and measured self-efficacy (moderator), price fairness (mediator), and willingness to pay (DV) (see Study 1).

Before we continue with manipulating versus measuring variables, the DV merits some additional discussion. Researchers must decide whether to measure DVs using categorical or continuous scales. The specification of the DV as categorical or continuous is important because it affects the data and analysis by determining the statistical tests researchers can perform as well as the interpretation of results. As an example of categorical DVs, researchers could ask participants to indicate which snack product (e.g., apple, brownie, or pretzels) they would purchase or observe which product they purchased. As an example of continuous DVs, researchers could instead ask participants how likely they are to purchase each snack product (e.g., on a Likert scale from 1 = extremely unlikely to 7 = extremely likely) or observe the number of ounces of each product participants consume. There are benefits and drawbacks of both categorical and continuous measures. One major benefit of categorical DVs is ease of interpretability but one major drawback is loss of information, which limits the types of statistical tests researchers can perform and the ability to detect subtle effects of an IV on a DV (Chu & Anderson, 1992). The reverse is true for continuous DVs. In practice, researchers often include both categorical and continuous DVs in an overall empirical package. For example, Torelli et al. (2017) manipulated cultural distinctiveness and asked participants to indicate their preferences for products from different states on 9-point Likert scales in Study 2A (continuous DV) and which product from different countries they preferred in Study 2B (categorical DV). Next, we return to our discussion on and share best practices for manipulating versus measuring variables.

When manipulating variables (IVs and moderators), it is imperative that researchers only change one construct across conditions, holding everything else constant. As an example, To and Patrick (2021) manipulated eye gaze of the model in an ad as averted versus direct (see Study 2 and Appendix C for stimuli). If the authors had manipulated eye gaze by changing the head position of the model in the stimuli, they would not have been able to ascertain whether eye gaze or head position caused the effect. In other words, eye gaze would be confounded with head position. Researchers can manipulate variables in a variety of ways. We share a few common approaches here and acknowledge that there are several other valid ways to manipulate variables. First, scenario-based approaches are common. As an example, Garcia-Rada et al. (2022) explored how investing effort in caregiving tasks affected caregivers' self-perceptions. The authors asked participants to imagine taking care of a sick friend and manipulated whether participants invested time, effort, or money in taking care of their friend (see Study 2). Instead of using an imagined scenario, experimental researchers may also design stimuli such as an advertisement or packaging that has the manipulation embedded directly in it. For example, Lim et al. (2020) created ads with green versus gray color schemes to test how color affected consumer responses to an eco-friendly product (see Study 1). Similarly, to understand how framing the discount depth effects consumer purchases, Guha et al. (2018) manipulated promotional frames, where half of the participants read that the new price was “now 31% lower” and the other half read that the old price “was 44% higher” while holding the sale price constant (see Study 1). Although the aforementioned examples include a number of different ways to manipulate IVs, researchers can also manipulate moderators. As an example, we return to Torelli et al. (2017), where the authors manipulated cultural distinctiveness (IV) as high (vs. low) using videos that featured the sights and sounds of countries that are culturally distinct from the U.S. and rivalry salience (moderator) as high (vs. low) via an article that described a local college football rivalry (see Study 4).

When measuring variables (moderators, mediators, and DVs), we recommend that researchers use established scales from the literature that are both reliable and valid. When it comes to measured variables, there is some debate over the use of single versus multi-item scales. Single-item scales tend to be less prevalent in marketing than other social sciences (Bergkvist, 2015; see Batra et al., 2000; Diener et al., 2013; and Nichols & Webster, 2013 for representative examples). Although some past research shows no differences in the predictive validity between single and multi-item scales (Bergkvist & Rossiter, 2007), other researchers strongly refute this assertion. For example, Diamantopoulos et al. (2012) compared single and multi-item scales and found that predictive validity varied widely using single-item scales. Furthermore, the authors found that multi-item scales performed better in terms of predictive validity across a variety of empirical applications. Results from Kamakura (2015) are consistent in that multi-item scales performed better than single-item scales. Kamakura (2015) also points out that multiple items enable researchers to assess reliability (which is not possible with a single item) as well as select items to create a more reliable measurement instrument if needed. As a result, although there may be instances where multi-item scales are not appropriate (e.g., when multi-item scales lead to participant fatigue), in general, we recommend that researchers use established, multi-item scales when possible. In reporting multi-item scales, researchers should include information on reliability (e.g., Cronbach alpha, composite reliability) at a minimum. When researchers include many measured variables in the same study (e.g., moderators, mediators, and DVs), it may also be beneficial to provide information on convergent validity (e.g., loading factor above .70, average variance extracted above .5) and discriminant validity (e.g., cross loading, square root of average variance extracted above correlation between variables).

Importantly, regardless of whether variables are manipulated or measured, researchers should try to increase the degree of experimental realism in the IV and measure actual behavior in the DV as much as possible—a topic we return to and expand on in Section 5 (Morales et al., 2017).

5 DECIDING ON THE RESEARCH DESIGN: BETWEEN-SUBJECTS, WITHIN-SUBJECTS, OR MIXED

After researchers define and operationalize the IV, they must decide how to assign participants to experimental conditions. In other words, they need to decide on the research design. There are three common research designs for experiments: between-subjects, within-subjects, and mixed. Importantly, we agree with past research that maintains each of these "designs have their merits, and the choice of designs should be carefully considered in the context of the question being studied and in terms of the practical implementation of the research study" (Charness et al., 2012, p. 1). We describe and summarize major advantages and disadvantages of each of the three designs below.

First, in between-subjects designs, researchers randomly assign participants to one of the experimental conditions (Koschate-Fischer & Schandelmeier, 2014) and test the impact of the IV on the DV between participants in each condition (Viglia & Dolnicar, 2020). As an example, Felix et al. (2022) examined the impact of packaging color on purchase intentions and the robustness of this effect via an experiment in which they randomly assigned participants to a 2 (packaging color: red vs. green) x 2 (saturation: low vs. high) x 2 (product positioning: strength-related vs. gentleness-related message) between-subjects design (see Study 2). There was a total of 2 x 2 x 2 = 8 experimental conditions and each participant saw one of eight possible combinations.

Benefits of between-subjects designs include easier experimental setup and analysis as well as shorter experimental sessions (Viglia et al., 2021). In addition, because participants only respond to one of the experimental conditions, learning and order effects (like lower engagement due to participant wear-out or survey fatigue; Savage & Waldman, 2008) are not an issue (Greenwald, 1976; Viglia et al., 2021). Therefore, between-subjects designs are more appropriate when carryover effects are likely for all IVs and moderate to large in size (Greenwald, 1976; Koschate-Fischer & Schandelmeier, 2014). Furthermore, some scholars recommend using between-subjects designs whenever possible because they offer a more conservative test (Charness et al., 2012) and provide greater confidence in the results (Viglia & Dolnicar, 2020). Drawbacks of between-subjects designs include less statistical power

(Charness et al., 2012; Greenwald, 1976; Thompson & Campbell, 2004), which means that they also require larger sample sizes, increasing the cost of the study (Charness et al., 2012; Viglia & Dolnicar, 2020). Moreover, because between-subjects designs contain more noise (i.e., random or unexplained variability), it can be more difficult for researchers to uncover true differences between conditions (Charness et al., 2012; Viglia & Dolnicar, 2020)—although researchers can include control variables to reduce this noise (Koschate-Fischer & Schandelmeier, 2014).

Second, in within-subjects (also known as repeated measures) designs, participants respond to all of the experimental conditions (Koschate-Fischer & Schandelmeier, 2014) and researchers test the impact of the IV on the DV within each participant (Viglia & Dolnicar, 2020). As an example, Raghubir and Srivastava (2002) examined consumer spending using different currencies and ran an experiment with a 6 (exchange rates: $1.00 = 9.5 Norwegian krone, 48 Luxembourg francs, 110 Japanese yen, 1,100 Korean wons, 24,500 Romanian leu, and 685,000 Turkish lira) group within-subjects design (see Study 1). Participants rated all six exchange rates presented in random order.

Benefits of within-subjects designs include more statistical power, which means that they also require smaller sample sizes (Alferes, 2012; Thompson & Campbell, 2004; Wedel & Dong, 2020). Furthermore, because each participant serves as their own control (i.e., the same participant responds more than once and the only thing that differs is the manipulation), researchers are more likely to uncover true differences across conditions and can rule out individual differences as potential alternative explanations, which reduces the need to include control variables (Koschate-Fischer & Schandelmeier, 2014; Viglia & Dolnicar, 2020). In terms of drawbacks, because participants respond to all the experimental conditions, learning and order effects can emerge in within-subjects designs (Greenwald, 1976; Judd et al., 2001). Therefore, this type of design is more appropriate when carryover effects are either unlikely for all IVs or small in size (Koschate-Fischer & Schandelmeier, 2014). Researchers can minimize learning and order effects by randomizing (i.e., counterbalancing) the order in which participants see each condition (Viglia & Dolnicar, 2020; Viglia et al., 2021) or increasing the amount of time between each condition (Greenwald, 1976). That said, in some studies, carryover effects are the topic of the investigation or also emerge in the real world (e.g., consumers view multiple products), in which case researchers may not want to minimize learning and order effects (Greenwald, 1976).

Finally, mixed designs include at least one between-subjects and one within-subjects factor. Researchers randomly assign participants to one of the between-subjects experimental conditions, and participants respond to all of the within-subjects experimental conditions (Koschate-Fischer & Schandelmeier, 2014; Wedel & Kopyakova, 2022). Researchers test the impact of the IV on the DV between participants in each condition for the between-subjects factor and within each participant for the within-subjects factor. As an example, Stoner et al. (2018) examined how naming products affected consumer evaluations and ran an experiment in which they randomly assigned participants to a 2 (product: mug vs. stapler) between-subjects x 3 (name: self-name, descriptive name, and nondescriptive name) within-subjects design (see Study 3). Participants reported their attitudes and purchase intentions for all three name conditions (presented in random order) for one product (either the mug or the stapler).

Mixed designs take advantage of the benefits of both between-subjects and within-subjects designs: more statistical power via the within-subjects factor as well as fewer learning and order effects via the between-subjects factor (Viglia & Dolnicar, 2020; Viglia et al., 2021). Past research recommends using mixed designs when carryover effects are likely for some but not all of the IVs, and the effects are moderate to large in size (Koschate-Fischer & Schandelmeier, 2014). However, as with between-subjects designs, researchers should include control variables to reduce the noise in mixed designs (Koschate-Fischer & Schandelmeier, 2014).

We refer interested readers to Wedel and Dong (2020; 2021) and Wedel and Kopyakova (2022) for a Bayesian extension of ANOVA (BANOVA) to conduct analyses (including moderation, mediation, and moderated mediation analyses, which we discuss in Section 6) for between-subjects, within-subjects, and mixed designs.

6 DECIDING ON THE RESEARCH SETTING: LABORATORY, FIELD, OR ONLINE

After researchers decide on the research design, they must decide where they want to conduct their experiment—that is, the research setting. There are three common research settings for experiments: in the laboratory, in the field, or online. Laboratory experiments take place in more artificial settings. This approach offers researchers more control over what the participants experience during the experiment. Therefore, lab studies typically provide more internal validity (Koschate-Fischer & Schandelmeier, 2014; Viglia et al., 2021), that is, "the extent to which the design, methods, and procedures of a study allow one to conclude that the independent variable (and no extraneous variable) influenced the dependent variable" (Kardes & Herr, 2019, p. 7). However, due to the artificial nature of the setting, lab studies typically provide less external validity (Koschate-Fischer & Schandelmeier, 2014; Viglia et al., 2021), that is, "the extent to which the results of a study generalize to other people and to other situations" (Kardes & Herr, 2019, p. 7). As an example, Torelli et al. (2017) ran a lab experiment in which they randomly assigned participants to a 2 (cultural distinctiveness vs. control group) group between-subjects design (see Study 1A). To manipulate cultural distinctiveness, the participants in the cultural distinctiveness condition watched two videos that showed the sights and sounds of a street in an Indian town and a cafeteria in an Asian college. Participants in the control group watched two videos that showed facts about insects. By conducting this study in the lab, the researchers were able to ensure that all participants focused on the videos (vs. multitasking) and viewed the videos under the same conditions in the same environment.

Unlike lab experiments, field experiments take place in more natural settings, where consumers are unaware that they are participating in an experiment (Baldassarri & Abascal, 2017; Charness et al., 2013; Gneezy, 2017). More specifically, participants are unaware of the experimental manipulation and that researchers are unobtrusively measuring their actual behavior (Baldassarri & Abascal, 2017; Gneezy, 2017; Morales et al., 2017). This approach gives researchers less control. Therefore, field studies typically provide less internal validity. However, due to the natural setting, field studies typically provide more external validity (Koschate-Fischer & Schandelmeier, 2014; Viglia et al., 2021). As an example, Robitaille et al. (2021) ran a field experiment in which participants were visitors to an actual, operational service center that provided public services, such as issuing driver's licenses, health cards, and photo identification. All visitors on a given day were exposed to a different type of promotional messaging regarding organ donation. Researchers assessed the impact of the promotional message intervention (IV) on organ donation registration (DV), an actual behavior with real world consequences.

Finally, online experiments take place over the internet with participants typically using their own electronic devices and participating at their convenience. Online experiments are particularly useful when researchers are interested in either a broader group of participants or conversely a very specific group of participants (Koschate-Fischer & Schandelmeier, 2014; Reips, 2002). These types of experiments provide a convenient way to recruit larger samples and tend to be relatively quick and inexpensive (Birnbaum, 2004). However, like field experiments, online experiments give researchers less control over what the participants experience during the experiment (e.g., the equipment participants use or the environment they are in) (Dandurand et al., 2008). As an example of the benefits of an online experimental setting, we return to Torelli et al. (2017). The authors ran an online experiment in which they randomly assigned MTurk participants (i.e., online panel participants recruited through Amazon's Mechanical Turk platform) in the U.S. to a 2 (cultural distinctiveness vs. control group) group between-subjects design (see Study 3). The authors examined how rivalry salience influenced the effect of cultural distinctiveness. Because Canada was a rival of the U.S. during the 2014 Winter Olympics, participants read about the final for women's curling between Canada and Sweden the day before the match. Given that Olympic finals are not predetermined, the authors had to wait for a match that included a team with high rivalry salience. Conducting an online experiment allowed the authors to collect data quickly after the teams were announced but before the match was played.

We note that when it comes to internal versus external validity, lab experiments do not necessarily offer high internal and low external validity, nor do field experiments necessarily offer low internal and high external validity (Koschate-Fischer & Schandelmeier, 2014). Researchers can design lab experiments in a realistic research setting to increase external validity or field experiments in a controlled environment to increase internal validity (see

Table 2 in Koschate-Fischer & Schandelmeier, 2014 for examples). Interestingly, Anderson et al. (1999) compared effect sizes between lab and field experiments with similar IVs and DVs and found a correlation of $r = .73$, indicating that lab experiments probably have higher external validity and field experiments probably have higher internal validity than we give them credit for (Kardes & Herr, 2019). However, some scholars stress that validity is more about theory and less about the decision to use a lab versus field experiment (Lynch, 1982; 1999). In addition, many scholars advocate the use of both lab and field experiments to strengthen the overall empirical package. For example, Gneezy (2017) recommends researchers provide converging evidence across different research settings. Researchers could run a lab experiment first to establish an effect and a field experiment second to show that the effect replicates in a more natural setting with actual behavior (Viglia et al., 2021). Alternatively, researchers could run a field experiment first and a lab experiment second to identify boundary conditions (moderation) or the underlying process (mediation) (Gneezy, 2017; Viglia et al., 2021). As an example, Chen et al. (2022) include both lab and field experiments to understand how customer rating affects tipping. In the lab study, participants read a scenario and indicated how much they would tip in that situation (see Study 2). In the field study, participants were diners in a Chinese restaurant, and the authors recorded actual tips (see Study 5).

In the end, we agree that researchers should use the setting that allows them to answer their research question best (Lieberman et al., 2019). That said, whichever setting researchers choose, Morales et al. (2017) encourage researchers to increase the degree of experimental realism in the IV and measure actual behavior in the DV to the greatest extent possible. For example, researchers can provide participants with real, physical products (IV) to assess to increase experimental realism and have them actually select or purchase products or services (DV) to measure actual behavior in a lab experiment. For example, Biswas et al. (2014) examined the impact of oral haptics via soft versus hard brownies (IV) on the number of grams of brownies participants ate (DV) in a lab setting. Researchers can also include "incentive-compatible" DVs to measure actual behavior. The goal of incentive-compatible DVs is to reveal consumers' true preferences. Some examples include measuring willingness to pay (WTP) using a BDM auction where participants submit binding bids for the focal product (Franke et al., 2010) or measuring donation behavior by asking participants to donate either the money they received for their participation or their time to a charitable organization (Macdonnell & White, 2015). Ultimately, measuring actual behavior in the DV allows researchers to gain better insights into real consumer behavior, enhancing the validity, generalizability, and applicability of the research. Importantly, not every experiment in a series of studies has to achieve this degree of realism or measure actual behavior (Morales et al., 2017). Rather, researchers can conduct different studies that vary in terms of experimental realism and behavioral measures to strengthen the overall empirical package. In fact, using different measures across different studies can enhance the robustness of the results.

7 UNDERSTANDING THE EFFECT VIA MODERATION, MEDIATION, OR MODERATED MEDIATION

7.1 Moderation

Once researchers identify their main effect of interest as well as how to test it, they can turn their attention to potential boundary conditions (moderation) and the underlying process or mechanism (mediation) to enhance the contribution of their work. We begin with moderation. Moderation occurs when there is an interaction between the IV and a third variable (i.e., the moderator) (Baron & Kenny, 1986; Edwards & Lambert, 2007; Hayes, 2022). When an interaction is significant, it means that the effect of the IV on the DV changes based on the level of the moderator (Holland et al., 2017) (see Figure 2 for a conceptual model). Therefore, moderation helps researchers explain "when" or under what circumstances the effect of the IV on the DV is significant (Hayes, 2022; Viglia & Dolnicar, 2020). In this way, moderator variables help researchers establish "boundary conditions" of the effect by identifying when the effect of the IV on the DV is significant and when it is not (Hayes, 2022). Given a significant interaction, researchers should assess the nature of the interaction with "simple effects" tests to explore how the effect of the IV on DV changes at different levels of the moderator (Kirk, 2013). A moderator can change either the magnitude (by strengthening or weakening) or the direction of the effect of the IV on the DV (Kardes & Herr, 2019; Viglia et al., 2021) (see Figure 3).

Insert Figures 2 and 3 about here

Researchers can either manipulate or measure moderators. Some researchers advocate for using both manipulated and measured moderators in an overall empirical package when possible to enhance confidence in the results and provide converging evidence of moderation (Kardes & Herr, 2019). While manipulating moderators provides a stronger test of causality, we acknowledge that some moderators (e.g., individual difference variables) may be difficult or even impossible to manipulate.

For manipulated moderators, researchers can run planned contrasts to examine the simple effects. As an example, Wang et al. (2022) ran an experiment in which they randomly assigned participants to a 2 (ambivalent attitudes: low vs. high) x 2 (anxious mood: low vs. high) between-subjects design (see Study 2). The 2-way interaction was significant. Specifically, low (vs. high) ambivalent attitudes increased green purchase intentions, and the magnitude of this effect was stronger when anxious mood was high (vs. low) (see Figure 1 in Wang et al., 2022). Thus, anxious mood moderated the effect of ambivalent attitudes (IV) on green purchase intentions (DV).

For measured moderators, we summarize three ways to assess the nature of the interaction. First, researchers can run spotlight analysis. A spotlight analysis tests the simple effects of the IV at specific values of the moderator (Krishna, 2016). In other words, a spotlight analysis shines a light on a single value of the moderator. For an example of a spotlight analysis, see Study 1 in Pundak et al. (2021), where wrongdoer identifiability moderated the effect of nonmaleficence (IV) on sharing likelihood (DV). While most researchers assess the effect of the IV on the DV at one standard deviation below the mean, the mean, and one standard deviation above the mean, Krishna (2016) and Spiller et al. (2013) recommend using meaningful focal values. When meaningful focal values do not exist, Spiller et al. (2013) recommend running a floodlight analysis instead.

Second, researchers can run a floodlight analysis (Johnson-Neyman technique), which provides more information than a spotlight analysis. A floodlight analysis tests the simple effects of the IV at all values of the moderator (Spiller et al., 2013). That is, a floodlight analysis shines a light on the entire range of the moderator and identifies "regions of significance" (Hayes, 2022). In a floodlight analysis, Johnson-Neyman (J-N) point(s) represent the specific value(s) of the moderator where the effect of the IV on the DV becomes significant (Wedel & Dong, 2020). One J-N point indicates that values are significant above or below that point, two J-N points indicate that values are significant either between those points or above and below those points, and no J-N point could indicate that either all or no values are significant (Hayes, 2022). For an example of a floodlight analysis, see Study 1 in Blank and Bolton (2019), where self-efficacy moderated the effect of psychological distance (IV) on price fairness (DV). The floodlight analysis revealed two J-N points: .69 SD below the mean and 1.55 SD above the mean. Importantly, a spotlight analysis at one standard deviation below the mean, the mean, and one standard deviation above the mean would have missed some of the nuances of this moderating effect, given that one of the J-N points occurs above +1SD.

Finally, researchers can perform a median split. Here, researchers covert a continuous moderator into a categorical moderator by splitting the data at the median. Then, researchers can run planned contrasts to examine the simple effects (as outlined above for manipulated moderators). For an example of a median split, see Study 4 in Wang et al. (2022). The authors measured depressed mood and performed a median split, which resulted in two groups: low versus high depressed mood. Depressed mood moderated the effect of ambivalent attitudes (IV) on green purchase intentions (DV). We note that although some scholars discourage the use of median splits (Krishna, 2016; Spiller et al., 2013) others encourage their use. For example, Iacobucci et al. (2015) argue that median splits are acceptable when the IVs are uncorrelated. Furthermore, Kardes and Herr (2019) point out that a median split is not only a simpler analysis but typically leads to the same substantive conclusions. When deciding whether to perform a median split, we suggest researchers investigate and follow the recommendations in their respective fields.

From an analysis standpoint, researchers can use the use the PROCESS macro to test moderation in between-subjects designs (this includes spotlight and floodlight analyses as well as the identification of J-N points in floodlight analyses) (Hayes, 2022) or the MEMORE (mediation and moderation analysis for repeated measures designs) macro to test moderation in two-condition within-subjects designs (Montoya, 2019).

We would like to share two final comments on moderation. First, researchers should carefully consider where they place moderators in the experiment, because they could affect the IV and/or DV. In general, we recommend placing moderators after the DV. In the case of moderators that are stable over time (e.g., individual difference variables), researchers could consider measuring them before the main study (see Studies 2 and 4 in Kirk et al., 2022) for an example, where the authors measured moderators over two weeks in advance). Second, although moderation occurs when the interaction between the IV and the moderator is significant, it is still possible that the effect of the IV on the DV is significant for some but not all values of the moderator when the interaction is not significant (Hayes, 2022). In these cases, it may be worthwhile to explore the pattern of results. However, researchers should be very careful in their interpretation of results. For example, if the effect of the IV on the DV is significant at one value of the moderator but not another, we cannot say that the two values are significantly different from each other (see Hayes, 2022 for more discussion).

7.2 Mediation

In contrast to moderation, mediation occurs when the IV leads to variation in a third variable (i.e., the mediator) which in turn leads to variation in the DV (Baron & Kenny, 1986; Edwards & Lambert, 2007; Hayes, 2022). In mediation, there is a direct effect from the IV to the DV and an indirect effect from the IV to the mediator to the DV. When mediation is significant, it means that the indirect effect from the IV to the mediator to the DV is significant. Therefore, mediation helps researchers explain "how" the IV causes the DV (Hayes, 2022). In other words, mediator variables help researchers establish the "underlying process" or "mechanism" of an effect by identifying the variables that intervene between the IV and DV (Fairchild & McKinnon, 2009; Hayes, 2022). Importantly, the mediator should be conceptually distinct from both the IV and the DV.

We highlight three common types of mediation in consumer research: simple, parallel, and serial. First, simple mediation includes one mediator between the IV and the DV (see Figure 4 for a conceptual model). As an example, Stoner et al. (2018) found that psychological ownership mediated the relationship between product name (IV) and purchase intentions (DV) (see Study 2). We note that this is the simplest form of mediation and that the underlying process is often more complex.

Insert figure 4 about here

Next, parallel mediation includes more than one mediator between the IV and the DV (see Figure 5 for a conceptual model). Parallel mediation can help researchers test the relative importance of multiple mediators or rule out alternative explanations. Importantly, when mediators operate "in parallel," they do not affect one another (Hayes, 2022). When more than one mediator is significant, mediation can be either complementary (same signs) or competitive (opposite signs) (Zhao et al., 2010). As an example, Li et al. (2019) found that both warmth (+) and competence (-) mediated the relationship between emoticons (IV) and willingness to interact (DV) in opposite directions (see Study 1). Such competitive mediation illustrates how an indirect effect can still be present even when there is no total (i.e., main) effect, as is the case in this study.

Insert figure 5 about here

Finally, serial mediation includes more than one mediator between the IV and the DV and each mediator affects the variable directly after it (see Figure 6 for a conceptual model). When two or more mediators operate "in serial" they are part of a causal chain between the IV and the DV (Hayes, 2022). As an example, Newman et al. (2014) tested the following serial mediation model: green enhancement intentions (IV) → resource allocation (M_1) → product quality (M_2) → purchase intent (DV) (see Experiment 1). The authors found that green enhancement intentions (IV) affected resource allocation (M_1) which in turn affected product quality (M_2) which in turn affected purchase intent (DV).

Insert figure 6 about here

Researchers can use the use the PROCESS macro to test simple, parallel, and serial mediation in between-subjects designs (Hayes, 2022) or the MEMORE macro to test simple, parallel, and serial mediation in two-condition within-subjects designs (Montoya & Hayes, 2017).

Next, we discuss two somewhat outdated guidelines for mediation. We are addressing these recommendations as they still persist in some circles as a topic of discussion even though the majority of experimental researchers in consumer behavior have moved away from them. First, early research on mediation specifies that the direct effect from the IV to the DV needs to be significant (Baron & Kenny, 1986). However, the only requirement to establish mediation is that the indirect effect from the IV to the mediator to the DV is significant (Hayes, 2022; Zhao et al., 2010). As a general rule, researchers should report the direct effect from the IV to the DV, the indirect effect from the IV to the mediator to the DV, and the total effect (i.e., the sum of the direct and indirect paths) when assessing mediation (Pieters, 2017). Second, past research often labels mediation as partial or full. Partial mediation occurs when the direct effect of the IV on the DV remains significant after adding the mediator, whereas full mediation occurs when the direct effect of the IV on the DV becomes insignificant after adding the mediator. A growing body of research recommends avoiding this terminology for two reasons. First, the distinction between partial and full mediation largely depends on sample size (Hayes & Rockwood, 2017; Hayes, 2022; Montoya & Hayes, 2017; Zhao et al., 2010). Second, using the term "full mediation" may signal that a single mediator totally explains why the IV causes the DV (i.e., there are no other potential mediators), which is rarely the case in practice (Hayes & Rockwood, 2017; Hayes, 2022).

Finally, we note that researchers often measure the mediator after the IV but before the DV to mimic the causal chain. However, in some cases, researchers may measure the mediator after the DV to minimize potential interference from the mediator. This may be more common when researchers use actual behavior as the DV or when demand effects might be likely. For example, Wu et al. (2017) tested the aesthetic design of food on consumption. In Study 2, participants ate actual cupcakes while watching video clips with the amount of the cupcake consumed as the DV. Participants responded to questions regarding perceptions of expensiveness of the cupcakes (a potential mediator) after they ate the cupcake. This makes sense for several reasons. First, the cupcake itself likely affected perceptions of expensiveness, and therefore, participants could not have rated expensiveness prior to seeing or eating the cupcake. Second, having participants pause their consumption to answer questions may have directly impacted how much they consumed either by making them more mindful of their consumption (decreasing consumption) or causing them to reflect on how expensive what they were eating was (increasing consumption because the cupcake was expensive).

7.3 Moderated mediation

Sometimes, researchers test both moderation and mediation in a single model. Models that combine both moderation and mediation analysis are called conditional process models (Hayes, 2022; Holland et al., 2017). We focus on moderated mediation, which occurs when mediation changes based on the level of the moderator (Edwards & Lambert, 2007; Hayes, 2015; Muller et al., 2005). The two most common forms of conditional process models are moderated mediation models where 1) moderation takes place between the IV and mediator (see Figure 7 for a conceptual model) and 2) moderation takes place between the mediator and DV (see Figure 8 for a conceptual model) (Hayes & Rockwood, 2020). As an example, Hasford et al. (2022) found an interaction between selfishness (IV) and emotional intelligence (moderator) such that high selfishness and emotional intelligence decreased physiological arousal (mediator) and increased subsequent fraud behavior (DV) (see Study 3).

Insert figures 7 and 8 about here

For conditional process models with more than one moderator and/or mediator, we refer readers to Hayes (2015, 2022) and Hayes and Rockwood (2020). Regardless of whether researchers are testing moderation, mediation, moderated mediation, or any other conditional process model, presenting conceptual models (such as Figures 2 and 4-8) can help readers understand the relationships researchers test and interpret the results (Viglia et al., 2021).

8 INCLUDING MANIPULATION AND ATTENTION CHECKS

One of the hallmarks of experiments is their ability to test cause and effect relationships. Researchers manipulate the IV to test its impact on the DV. Therefore, it is essential that the manipulation of the IV reflects the construct of interest and affects participants as expected. Researchers can implement manipulation checks to test the effectiveness of the manipulation. A manipulation check is a direct measure of the state that a manipulation should induce (Hauser et al., 2018). Some past research argues that manipulation checks are essential to ensure the validity of results (Fiedler et al., 2021; Pechman, 2019; Perdue & Summers, 1986). However, other researchers question whether manipulation checks are necessary (Hauser et al., 2018; Gruijters, 2022; Sawyer et al., 1995). For example, manipulation checks are probably not necessary when no plausible alternative explanation for the effect exists, researchers use well-established manipulations, the IV is isomorphic with its operationalization (e.g., the impact of color vs. black and white images where researchers manipulate the color of the image), or the manipulation is common sense or generally accepted knowledge (Kardes & Herr, 2019; Sawyer et al., 1995; Sigall & Mills, 1998). As an example, Stoner et al. (2018) manipulated whether participants named a stress ball or not (Study 1). In this case, the authors did not need to include a manipulation check because the name was objectively present or absent. Asking participants whether they named a stress ball would actually serve as an attention check (which we discuss later in this section) not a manipulation check.

In the event that a manipulation check is necessary, we concur with past research that strongly advocates for conducting a pretest with a separate sample from the same population to test the manipulation (Perdue & Summers, 1986, Kardes & Herr, 2019). This approach allows researchers to identify any problems before the main study and minimizes concerns related to including manipulation checks in the main study (Kardes & Herr, 2019). In some cases, a pretest may not be feasible when collecting data is prohibitively expensive or time consuming or when the participant pool is small, and a pretest might contaminate the main study using the same participants. In these instances, researchers often embed a manipulation check in the main study. However, placing a manipulation check before the DV can lead to an interaction with the IV or demand effects, where participants respond in a way that they otherwise would not. Potential problems with placing the manipulation check after the DV include the possibility that the manipulation may wear off by the end of the survey or that the intervening variables may contaminate the manipulation check. However, this is generally preferable to placing the manipulation check before the DV and contaminating the DV. We therefore recommend that if a manipulation check is in the main study, it should be located after the DV, preferably with a filler task between the DV and the manipulation check. Regardless of whether the manipulation check is run separately or embedded in the study, it needs to ensure face validity (i.e., the stimuli is effective in manipulating what it is supposed to). As an example of a manipulation check, we return to Torelli et al. (2017). The authors implemented manipulation checks in a series of pilot studies to test the effectiveness of videos in manipulating cultural distinctiveness. Participants watched the videos that the authors wanted to use in the main studies and responded to items that assessed cultural distinctiveness. In addition, the authors used the pilot studies to rule out alternative explanations for the effect of the IV on the DV. Specifically, the authors found that the cultural distinctiveness manipulation did not lead to differences in need for closure, positive mood, or negative mood. In this way, researchers can also use pretests with manipulation checks to rule out alternative explanations.

In contrast to manipulation checks, attention checks (also known as instructional manipulation checks), are questions that researchers can ask to confirm that participants are paying attention (Oppenheimer et al., 2009). Inattentive participants can increase the amount of noise and decrease the validity of the data, making removing participants who fail attention checks attractive to researchers (Berinsky et al., 2014, 2021; Oppenheimer et al., 2009). As an example of an attention check, researchers could ask participants to select a specific response option (e.g., "select 'neither disagree nor agree'" or "select the number four"). Researchers can design attention checks by either embedding them into a matrix table with other questions or having them stand alone (Berinsky et al., 2021). Another form of attention check (Abbey & Meloy, 2017) is a comprehension check (also known as factual manipulation check; Hauser et al., 2019), where participants summarize or answer basic questions about what they just read or viewed. In addition to traditional attention checks, researchers could also identify inattentive participants by examining response speed or checking for patterns in the data (e.g., straight lining, extreme responses, Christmas tree or other patterns, etc.) (Abbey & Meloy, 2017; Hauser et al., 2019).

Although Berinsky et al. (2014) argue that attention checks do not affect participant responses on other items, an increasing body of evidence outlines several drawbacks of attention checks. For example, attention checks can lead to participant attrition (Abbey & Meloy, 2017), participant backlash that can reduce subsequent data quality (Oppenheimer et al., 2009; Vannette, 2017), changes in attention (Hauser & Schwarz, 2015), and an increase in systematic thinking (Hauser & Schwarz, 2015). In addition, removing participants who fail attention checks can lead to “cherry picking” participants (Abbey & Meloy, 2017) or introduce bias if certain demographics or psychographics are more likely to fail the attention checks (Vannette, 2017). For example, participants with low trait attentiveness are more likely to fail attention checks, which could be problematic if trait attentiveness is correlated with other traits integral to the investigation at hand (Berinsky et al., 2014; Hauser et al., 2019; Thomas & Clifford, 2017). Finally, past research contends that because attention changes over time and attention checks vary in terms of difficulty, are weakly correlated, and are sensitive to learning effects, attention is actually quite difficult to measure (Berinsky et al., 2014, 2021; Curran, 2016; Hauser et al., 2019; Huang et al., 2015; Meade & Craig, 2012; Niessen et al., 2016; Thomas & Clifford, 2017).

Because attention checks can reduce data quality, Qualtrics no longer recommends screening out participants based on attention checks (Vannette, 2017). An alternative strategy is to collect more data (i.e., a larger sample size) to offset any noise created by inattentive participants. If researchers do opt to implement attention checks, however, we agree with past research that recommends embedding them in an unobtrusive manner (Hauser et al., 2019; Thomas & Clifford, 2017) and placing them after the most important variables to minimize the potential for contamination (Hauser & Schwarz, 2015; Hauser et al., 2019). Overall, we recommend that researchers exercise caution in using attention checks and do not recommend removing participants simply for failing attention checks.

9 DETERMINING SAMPLE SIZE AND SELECTING PARTICIPANTS

In an experiment, it is important to ensure that there are enough participants in each experimental condition—that is "per cell." Past research argues that a number of published studies are woefully underpowered, which means that the sample size is too small to detect the effect of interest (Baroudi & Orlikowski, 1989; Maxwell, 2004). For example, in an investigation of 57 articles, Baroudi and Orlikowski (1989) found that researchers had a 40% chance of not finding the effect due to low power. Therefore, the authors recommended that the sample size should be large enough to detect medium effects but small enough that trivial associations do not reach significance.

Researchers typically determine the sample size in one of two ways. First, researchers could use a rule of thumb. For example, a common rule of thumb is 30 participants per cell (Koschate-Fischer & Schandelmeier, 2014). However, other existing rules of thumb are both lower and higher. For example, Simmons et al. (2011) recommend at least 20 participants per cell, whereas Pechman (2019) recommends at least 50 participants per cell for online samples like MTurk. Top journals may require even more participants per cell, especially for online samples (Pechman, 2019).

Another way to determine sample size is to calculate it based on the significance level, statistical power, and effect size (Koschate-Fischer & Schandelmeier, 2014). Researchers can estimate the effect size based on previous research with similar studies in the same domain or via a pretest (Baroudi & Orlikowski, 1989; Koschate-Fischer & Schandelmeier, 2014; Viglia et al., 2021). There are tools to help researchers determine the sample size using this approach. For example, Cohen (1992) offers a useful chart for determining the appropriate sample size for different statistical tests. In addition, there are a variety of programs and procedures to help researchers conduct power and sample size analysis, such as G*power 3 (Faul et al., 2007), GLMPOWER and POWER procedures in SAS, the Power Analysis procedure in SPSS, etc. However, even if researchers calculate the sample size using this statistical approach, they should still be aware of the sample size expectations in their field in case they need to collect even larger samples to publish in the journals in their field.

In addition to increasing sample size to increase power, we note that Meyvis and Van Osselaer (2018) recommend increasing the effect size to increase power. Researchers can achieve this through the careful planning and analysis of experiments. This approach may be especially helpful when very large sample sizes are just not feasible.

Once researchers determine the sample size, they must select participants. Unfortunately, researchers are often constrained by the cost of recruitment and the availability of participants. Therefore, convenience samples are useful in experimental research (Schwarz, 2019). The use of convenience samples, however, often leads to trade-offs in that inexpensive and readily available participants may be either more homogeneous (such as with student participants) or cause concerns about inattention (such as with online participants). In general, using participants from a variety of sources in a series of studies can often offset the limitations inherent in a single source (although we acknowledge that narrow research questions may require narrower samples, e.g., the impact of anti-smoking advertising on cigarette smokers; Schwarz, 2019). Next, we discuss two of the most common participant pools in consumer research: student and online samples.

When it comes to student samples, two of the biggest criticisms are that students are not representative of “real” consumers (Kardes & Herr, 2019) and that student samples are rather homogenous (e.g., students tend to be similar in age, education, etc.) (Koschate-Fischer & Schandelmeier, 2014). As a result, according to the critics of student samples, researchers cannot extend their findings to “real” consumers, limiting the generalizability (i.e., external validity) of the results (Espinosa & Oritnau, 2016; Lynch, 1982). These concerns have led some journals to discourage student samples (Bello et al., 2009). In contrast, some past research argues that student samples are often appropriate and sometimes even more appropriate than non-student samples. Koschate-Fischer and Schandelmeier (2014) summarize several reasons for this. Regarding representativeness, students do not pose a problem if the differences between students and non-students do not affect the IV. That is, although differences may exist on irrelevant attributes to the research at hand, students may be similar to non-students on the relevant attributes (Kardes & Herr, 2019). Regarding homogeneity, this can reduce error variance and the impact of individual difference variables on the DV, making it easier to find an effect. The exception to this rule is when researchers cannot study the variable of interest (e.g., age) due to the homogeneity. Finally, students are especially qualified when they are familiar with the research context or when researchers are not sure if individual difference variables will interact with the IV. Given the controversy over the use of student samples, we recommend that if researchers use student samples, they should also use non-student samples to show that the effect replicates (Espinosa & Oritnau, 2016). For example, Rodas and John (2020) study the impact of women's secret consumption on product evaluations and choice using three undergraduate student samples and five samples where the authors recruited women at a state fair. Ultimately, while student samples are generally acceptable, researchers should think about the relevance of using student participants given their research topic (Ferber, 1977).

Like student samples, online samples facilitated by Amazon Mechanical Turk (MTurk), CloudResearch, Prolific, Qualtrics, and other companies offer researchers a relatively quick and inexpensive way to recruit a large number of participants. Some criticisms of online samples include inattentiveness, poor language comprehension, non-naiveté, deception, attrition, and self-selection. We expand on some of the most common concerns here and refer interested readers to Hauser et al. (2019) and Wright and Goodman (2019) for additional information. First, although past research shows that results from student participants replicate with MTurk participants, and that in some cases, MTurk participants are more attentive and provide higher quality data than students (Goodman & Paolacci, 2017; Hauser & Schwarz, 2016; Hauser et al., 2019; Wright & Goodman, 2019), the quality of online participants varies widely (Schwarz, 2019). Some platforms like CloudResearch (formerly TurkPrime) try to help researchers save time and resources as well as improve data quality (see Litman et al., 2017). However, there are still concerns with attention (Goodman & Paolacci, 2017; Hauser & Schwarz, 2016; Hauser et al., 2019; Peer et al., 2017; Wright & Goodman, 2019) and bots (e.g., computer programs that complete HITs on MTurk) (Chmielewski & Kucker, 2020) in online samples. Thus, researchers who use online samples may want to restrict participation to participants with high approval rates or implement attention checks (see Section 7) to ensure that participants are paying attention (Goodman et al., 2013; Hauser et al., 2019). Another potential issue with online samples is that participants are more experienced with completing surveys and thus "non-naïve" (Chandler et al., 2014; Goodman & Paolacci, 2017; Goodman et al., 2013; Hauser et al., 2019; Peer et al., 2017; Wright & Goodman, 2019). To minimize this concern, researchers can exclude online participants who participated in their previous similar surveys and/or use less common manipulations or measures (Goodman & Paolacci, 2017; Hauser et al., 2019). Next, similar to student samples, online samples may not be representative of the general population. For example, past research shows that online participants differ from the general population in terms of their age, education, religiosity, political ideology, and other characteristics (Goodman & Paolacci, 2017; Paolacci & Chandler, 2014; Wright & Goodman, 2019). Finally, there is some concern that because online samples are so quick and inexpensive, they may lead to questionable research practices, where researchers run multiple studies and only report those that are consistent with hypotheses or that offer the most novel contributions (Wright & Goodman, 2019). Because of these criticisms, we recommend that if researchers use online samples, they should try to replicate the effect using other samples as well.

As a final note, although we advocate using participants from multiple sources, researchers should exercise caution in interpreting results from cross-cultural samples. For example, different languages (or even dialects) can alter participant interpretation and responses (Malhotra et al., 1996). Likewise, past research shows that cultural differences can generate question order effects as well as socially desirable responding and affect responses related to frequency estimates or items with reverse scoring (Haberstroh et al., 2002; Ji et al., 2000; Lalwani et al., 2006; Schwarz, 2003; Wong et al., 2003). While we encourage replication using diverse samples, researchers should keep in mind that failure to replicate an effect could be due to the complexities of cross-cultural samples rather than the absence of the effect. (See Malhotra et al., 1996 for an overview of considerations when performing cross-cultural research.)

10 CONCLUSION

In this article, we provide best practices for implementing experimental research methods in consumer studies. Experiments are important in consumer research because they can help researchers establish cause and effect relationships through the manipulation of an IV and random assignment of participants to experimental conditions. When designing experiments, researchers must make several decisions to design a well thought out experiment, which we outline in Figure 1.

First, researchers need to define their constructs and develop testable hypotheses based on strong theory. Next, researchers need to operationalize the IV, moderator, mediator, and DV. Typically, the IV is a manipulated variable (where researchers need to determine the number of levels), the moderator is either manipulated or measured, and the mediator and DV are measured variables. The mediator is usually continuous, whereas the DV can be either categorical or continuous. Regardless of these specific decisions, researchers should try to enhance the degree of experimental realism in the IV and measure actual behavior in the DV as much as they can.

Another issue for researchers to consider is the research design. We discussed the benefits and drawbacks of between-subjects, within-subjects, and mixed designs. When carryover effects are a major concern, researchers should use a between-subjects or mixed design. Closely related to the research design is the research setting. We compared internal and external validity across laboratory, field, and online experiments. Typically, lab experiments offer more internal validity and less external validity, whereas the reverse is true for field experiments. However, it is possible (and we encourage) researchers to try to enhance external validity in lab experiments and internal validity in field experiments. Although some past research recommends using between-subjects designs and field experiments whenever possible, researchers should use the research design and setting that helps them answer their research question best.

Once researchers establish a main effect, they can assess potential boundary conditions (via moderation) as well as the underlying process (via mediation) to enhance their contribution. We discussed manipulated and measured moderators as well as how to assess the nature of the interaction. We also described three common types of mediation in consumer research: simple, parallel, and serial mediation.

Next, researchers can consider including manipulation or attention checks in their experiments. However, we note that there is some controversy surrounding manipulation and attention checks. If manipulation checks are necessary, we recommend implementing them in a pretest when possible. Although using attention checks can improve data quality (especially among online samples), we outline several potential drawbacks of attention checks. Perhaps a better course of action is to collect a large enough sample to account for any noise created due to inattentive responding instead.

Finally, we discussed the number and selection of participants. First, we shared some rules of thumb for cell sizes as well as tools to calculate sample size. Second, we recommended that although conducting experiments with students is generally acceptable, researchers should target non-students in studies where differences between students and non-students could affect the variables in the experiment. Ideally, researchers should use a mix of participants from multiple sources in a series of studies to show that the effect replicates across different samples. Related to this recommendation, one major theme that emerged is that researchers should consider making different decisions across studies (e.g., in operationalizing variables, research design, research setting, participants, etc.) to provide converging evidence and strengthen the overall empirical package. Although we focus on experiments in this paper, where researchers manipulate the IV to show cause and effect relationships, we note that some researchers may also measure the IV to demonstrate the robustness of the effect across a series of studies.

Importantly, we acknowledge that we only scratched the surface of experimental research methods in this paper. Therefore, we refer readers to the articles we cited to understand the nuances of each topic and for additional guidance. We also recognize that best practices for implementing experiments are constantly evolving. For example, an emerging (and controversial) trend in top tier marketing journals is pre-registration—the act of documenting your research question or hypothesis, IV, DV, control variables, analyses, rules to exclude participants, and sample size before you collect data (e.g., on AsPredicted.org) (Simmons et al., 2021). Some scholars think pre-registration is required to believe the results, while others think it just adds more work (Krishna, 2021, p. 146). For instance, Simmons et al. (2021) argue pre-registration can increase transparency and decrease p-hacking (the selective reporting of results). However, Pham and Oh (2021b) argue that pre-registration is neither necessary nor sufficient for good science and that better tools with lower opportunity costs exist (e.g., sharing study materials and data) (Pham & Oh, 2021a; 2021b). Given this and other ongoing debates in various fields, we encourage researchers to familiarize themselves with the author guidelines of target journals and to seek out new publications to stay up to date on best practices for experiments.

REFERENCES

Abbey, J. D., & Meloy, M. G. (2017). Attention by design: Using attention checks to detect inattentive respondents and improve data quality. *Journal of Operations Management*, *53-56*, 63-70. https://doi.org/10.1016/j.jom.2017.06.001

Alferes, V. R. (2012). *Methods of randomization in experimental designs*. SAGE.

Anderson, C. A., Lindsay, J. J., & Bushman, B. J. (1999). Research in the psychological laboratory: Truth or triviality? *Current Directions in Psychological Science*, *8*(1), 3-9. https://doi.org/10.1111%2F1467-8721.00002

Baldassarri, D., & Abascal, M. (2017). Field experiments across the social sciences. *Annual Review of Sociology*, *43*, 41-73. https://doi.org/10.1146/annurev-soc-073014-112445

Baron, R. M., & Kenny, D. A. (1986). The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, *51*(6), 1173-1182. https://doi.org/10.1037/0022-3514.51.6.1173

Baroudi, J. J., & Orlikowski, W. J. (1989). The problem of statistical power in MIS research. *MIS Quarterly*, *13*(1), 87-106. https://doi.org/10.2307/248704

Batra, R., Ramaswamy, V., Alden, D. L., Steenkamp, J.-B. E. M., & Ramachander, S. (2000). Effects of brand local and nonlocal origin on consumer attitudes in developing countries. *Journal of Consumer Psychology*, *9*(2), 83-95. https://doi.org/10.1207/S15327663JCP0902_3

Bello, D., Leung, K., Radebaugh, L., Tung, R. L., & van Witteloostuijn, A. (2009). From the editors: Student samples in international business research. *Journal of International Business Studies*, *40*, 361-364. https://doi.org/10.1057/jibs.2008.101

Bergkvist, L. (2015). Appropriate use of single-item measures is here to stay. *Marketing Letters*, *26*, 245-255. https://doi.org/10.1007/s11002-014-9325-y

Bergkvist, L., & Rossiter, J. R. (2007). The predictive validity of multiple-item versus single-item measures of the same construct. *Journal of Marketing Research*, *44*(2), 175-184. https://doi.org/10.1509%2Fjmkr.44.2.175

Berinsky, A. J., Margolis, M. F., & Sances, M. W. (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. https://doi.org/10.1111/ajps.12081

Berinsky, A. J., Margolis, M. F., Sances, M. W., & Warshaw, C. (2021). Using screeners to measure respondent attention on self-administered surveys: Which items and how many? *Political Science Research and Methods*, *9*(2), 430-437. https://doi.org/10.1017/psrm.2019.53

Birnbaum, M. H. (2004). Human research and data collection via the Internet. *Annual Review of Psychology*, *55*, 803-832. https://doi.org/10.1146/annurev.psych.55.090902.141601

Biswas, D., Szocs, C., Krishna, A., & Lehmann, D. R. (2014). Something to chew on: The effects of oral haptics on mastication, orosensory perception, and calorie estimation. *Journal of Consumer Research*, *41*(2), 261-273. https://doi.org/10.1086/675739

Blank, A. S., & Bolton, L. E. (2019). Putting a price on user innovation: How consumer participation can decrease perceived price fairness. *Journal of the Association for Consumer Research*, *4*(3), 256-268. https://doi.org/10.1086/703566

Calder, B. J., Brendl, C. M., Tybout, A. M., & Sternthal, B. (2021). Distinguishing constructs from variables in designing research. *Journal of Consumer Psychology*, *31*(1), 188-

208. https://doi.org/10.1002/jcpy.1204

Calder, B. J., Phillips, L. W., & Tybout, A. M. (1981). Designing research for application. *Journal of Consumer Research*, *8*(2), 197-207. https://doi.org/10.1086/208856

Chandler, J., Mueller, P., & Paolacci, G. (2014). Nonnaïveté among Amazon Mechanical Turk workers: Consequences and solutions for behavioral researchers. *Behavior Research Methods*, *46*, 112-130. https://doi.org/10.3758/s13428-013-0365-7

Charness, G., Gneezy, U., & Kuhn, M. A. (2012). Experimental methods: Between-subject and within-subject design. *Journal of Economic Behavior & Organization*, *81*(1), 1-8. https://doi.org/10.1016/j.jebo.2011.08.009

Charness, G., Gneezy, U., & Kuhn, M. A. (2013). Experimental methods: Extra-laboratory experiments-extending the reach of experimental economics. *Journal of Economic Behavior & Organization*, *91*, 93-100. https://doi.org/10.1016/j.jebo.2013.04.002

Chen, J., Xu, A. J., Rodas, M. A., & Liu, X. (2022). Order matters: Rating service professionals first reduces tipping amount. *Journal of Marketing*. Advance online publication. https://doi.org/10.1177%2F00222429221098698

Chmielewski, M., & Kucker, S. C. (2020). An MTurk crisis? Shifts in data quality and the impact on study results. *Social Psychological and Personality Science, 11*(4), 464-473. https://doi.org/10.1177%2F1948550619875149

Chu, W., & Anderson, E. M. (1992). Capturing ordinal properties of categorical dependent variables: A review with application to modes of foreign entry. *International Journal of Research in Marketing*, *9*(2), 149-160. http://dx.doi.org/10.1016/0167-8116(92)90035-J

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*(1), 155-159.

https://doi.org/10.1037/0033-2909.112.1.155

Curran, P. G. (2016). Methods for the detection of carelessly invalid responses in survey data. *Journal of Experimental Social Psychology*, *66*, 4-19. https://doi.org/10.1016/j.jesp.2015.07.006

Dandurand, F., Shultz, T. R., & Onishi, K. H. (2008). Comparing online and lab methods in a problem-solving experiment. *Behavior Research Methods*, *40*(2), 428-434. https://doi.org/10.3758/BRM.40.2.428

Diamantopoulos, A., Sarstedt, M., Fuchs, C., Wilczynski, P., & Kaiser, S. (2012). Guidelines for choosing between multi-item and single-item scales for construct measurement: A predictive validity perspective. *Journal of the Academy of Marketing Science*, *40*, 434-449. https://doi.org/10.1007/s11747-011-0300-3

Diener, E., Inglehart, R., & Tay, L. (2013). Theory and validity of life satisfaction scales. *Social Indicators Research*, *112*, 497-527. https://doi.org/10.1007/s11205-012-0076-y

Edwards, J. R., & Lambert, L. S. (2007). Methods for integrating moderation and mediation: A general analytical framework using moderated path analysis. *Psychological Methods*, *12*(1), 1-22. https://psycnet.apa.org/doi/10.1037/1082-989X.12.1.1

Espinosa, J. A., & Ortinau, D. J. (2016). Debunking legendary beliefs about student samples in marketing research. *Journal of Business Research, 69*(8), 3149-3158. https://doi.org/10.1016/j.jbusres.2015.12.003

Fairchild, A. J., & MacKinnon, D. P. (2009). A general model for testing mediation and moderation effects. *Prevention Science*, *10*, 87-99. https://doi.org/10.1007/s11121-008-0109-6

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*, 175-191. https://doi.org/10.3758/BF03193146

Felix, R., González, E. M., Castaño, R., Carrete, L., & Gretz, R. T. (2022). When the green in green packaging backfires: Gender effects and perceived masculinity of environmentally friendly products. *International Journal of Consumer Studies*, *46*(3), 925-943. https://doi.org/10.1111/ijcs.12738

Ferber, R. (1977). Research by convenience. *Journal of Consumer Research, 4*(1), 57-58. https://doi.org/10.1086/208679

Fiedler, K., McCaughey, L., & Prager, J. (2021). Quo vadis, methodology? The key role of manipulation checks for validity control and quality of science. *Perspectives on Psychological Research*, *16*(4), 816-826. https://doi.org/10.1177%2F1745691620970602

Fitzsimons, G. J. (2008). Death to dichotomizing. *Journal of Consumer Research*, *35*(1), 5-8. https://doi.org/10.1086/589561

Franke, N., Schreier, M., & Kaiser, U. (2010). The "I designed it myself" effect in mass customization. *Management Science*, *56*(1), 125-140. https://doi.org/10.1287/mnsc.1090.1077

Garcia-Rada, X., Steffel, M., Williams. E. F., & Norton, M. I. (2022). Consumers value effort over ease when caring for close others. *Journal of Consumer Research*, *48*(6), 970-990. https://doi.org/10.1093/jcr/ucab039

Gneezy, A. (2017). Field experimentation in marketing research. *Journal of Marketing Research*, *54*(1), 140-143. https://doi.org/10.1509%2Fjmr.16.0225

Goodman, J. K., & Paolacci, G. (2017). Crowdsourcing consumer research. *Journal of*

Consumer Research, *44*(1), 196-210. https://doi.org/10.1093/jcr/ucx047

Goodman, J. K., Cryder, C. E., & Cheema, A. (2013). Data collection in a flat world: The strengths and weaknesses of Mechanical Turk samples. *Journal of Behavioral Decision Making*, *26*(3), 213-224. https://doi.org/10.1002/bdm.1753

Greenwald, A. G. (1976). Within-subjects designs: To use or not to use? *Psychological Bulletin*, *83*(2), 314-320. https://psycnet.apa.org/doi/10.1037/0033-2909.83.2.314

Gruijters, S. L. K. (2022). Making inferential leaps: Manipulation checks and the road towards strong inference. *Journal of Experimental Social Psychology*, *98*, 104251. https://doi.org/10.1016/j.jesp.2021.104251

Guha, A., Biswas, A., Grewal, D., Verma, S., Banerjee, S., & Nordfält, J. (2018). Reframing the discount as a comparison against the sales price: Does it make the discount more attractive? *Journal of Marketing Research*, *55*(3), 339-351. https://doi.org/10.1509%2Fjmr.16.0599

Haberstroh, S., Oyserman, D., Schwarz, N., Kühnen, U., & Ji, L.-J. (2002). Is the interdependent self more sensitive to question context than the independent self? Self-construal and the observation of conversational norms. *Journal of Experimental Social Psychology, 38*(3), 323-329. https://doi.org/10.1006/jesp.2001.1513

Han, M., Newman, G. E., Smith, R. K., & Dhar, R. (2021). The curse of the original: How and when heritage branding reduces consumer evaluations of enhanced products. *Journal of Consumer Research, 48*(4), 709-730. https://doi.org/10.1093/jcr/ucab028

Hasford, J., Kidwell, B., Hardesty, D. M., & Farmer, A. (2022). Your cheatin' heart: How emotional intelligence and selfishness impact the incidence of consumer fraud. *Journal of Consumer Research*, 49(1), 112-131. https://doi.org/10.1093/jcr/ucab047

Hauser, D. J., & Schwarz, N. (2016). Attentive Turkers: Mturk participants perform better on online attention checks than do subject pool participants. *Behavior Research Methods*, *48*, 400-407. https://doi.org/10.3758/s13428-015-0578-z

Hauser, D. J., Ellsworth, P. C., & Gonzalez, R. (2018). Are manipulation checks necessary? *Frontiers in Psychology*, *9*, 1-10. https://doi.org/10.3389/fpsyg.2018.00998

Hauser, D., Paolacci, G., & Chandler, J. (2019). Common concerns with MTurk as a participant pool. In F. R. Kardes, P. M. Herr, & N. Schwarz (Eds.), *Handbook of research methods in consumer psychology* (pp. 319-337). Routledge. https://doi.org/10.4324/9781351137713

Hayes, A. F. (2015). An index and test of linear moderated mediation. *Multivariate Behavioral Research*, *50*(1), 1-22. https://doi.org/10.1080/00273171.2014.962683

Hayes, A. F. (2022). *Introduction to mediation, moderation, and conditional process analysis: A regression-based approach* (3rd ed.). Guilford.

Hayes, A. F., & Rockwood, N. J. (2017). Regression-based statistical mediation and moderation analysis in clinical research: Observations, recommendations, and implementation. *Behaviour Research and Therapy*, *98*, 39-57. https://doi.org/10.1016/j.brat.2016.11.001

Hayes, A. F., & Rockwood, N. J. (2020). Conditional process analysis: Concepts, computation, and advances in the modeling of the contingencies of mechanisms. *American Behavioral Scientist*, *64*(1), 19-54. https://doi.org/10.1177%2F0002764219859633

Holland, S. J., Shore, D. B., & Cortina, J. M. (2017). Review and recommendations for integrating mediation and moderation. *Organizational Research Methods*, *20*(4), 686-

720. https://doi.org/10.1177%2F1094428116658958

Huang, J. L., Bowling, N. A., Liu, M., & Li, Y. (2015). Detecting insufficient effort responding with an infrequency scale: Evaluating validity and participant reactions. *Journal of Business and Psychology*, *30*, 299-311. https://doi.org/10.1007/s10869-014-9357-6

Iacobucci, D., Posavac, S. S., Kardes, F. R., Schneider, M. J., Popovich, D. L. (2015). The median split: Robust, refined, and revived. *Journal of Consumer Psychology*, *25*(4), 690-704. https://doi.org/10.1016/j.jcps.2015.06.014

Janiszewski, C., & van Osselaer, S. M. J. (2021). The benefits of candidly reporting consumer research. *Journal of Consumer Psychology*, *31*(4), 633-646. https://doi.org/10.1002/jcpy.1263

Janiszewski, C., Labroo, A., & Rucker, D. D. (2016). A tutorial in consumer research: Knowledge creation and knowledge appreciation in deductive-conceptual consumer research. *Journal of Consumer Research*, *43*(2), 200-209. https://doi.org/10.1093/jcr/ucw023

Ji, L.-J., Schwarz, N., & Nisbett, R. E. (2000). Culture, autobiographical memory, and behavioral frequency reports: Measurement issues in cross-cultural studies. *Personality and Social Psychology Bulletin*, *26*(5), 585-593. https://doi.org/10.1177%2F0146167200267006

Judd, C. M., Kenny, D. A., & McClelland, G. H. (2001). Estimating and testing mediation and moderation in within-subjects designs. *Psychological Methods*, *6*(2), 115-134. https://psycnet.apa.org/doi/10.1037/1082-989X.6.2.115

Kamakura, W. A. (2015). Measure twice and cut once: The carpenter's rule still applies.

Marketing Letters, *26*, 237-243. https://doi.org/10.1007/s11002-014-9298-x

Kardes, F. R., & Herr, P. M. (2019). Experimental research methods in consumer psychology. In F. R. Kardes, P. M. Herr, & N. Schwarz (Eds.), *Handbook of research methods in consumer psychology* (pp. 3-16). Routledge. https://doi.org/10.4324/9781351137713

Kirk, C. P., Peck, J., Hart, C. M., & Sedikides, C. (2022). Just my luck: Narcissistic admiration and rivalry differentially predict word of mouth about promotional games. *Journal of Business Research*, *150*, 374-388. https://doi.org/10.1016/j.jbusres.2022.06.004

Kirk, R. E. (2013). *Experimental design: Procedures for the behavioral sciences* (4th ed.). SAGE.

Koschate-Fischer, N., & Schandelmeier, S. (2014). A guideline for designing experimental studies in marketing research and a critical discussion of selected problem areas. *Journal of Business Economics*, *84*, 793-826. https://doi.org/10.1007/s11573-014-0708-6

Krishna, A. (2016). A clearer spotlight on spotlight: Understanding, conducting and reporting. *Journal of Consumer Psychology*, 26(3), 315-324. https://doi.org/10.1016/j.jcps.2016.04.001

Krishna, A. (2021). The need for synergy in academic policies: An introduction to the dialogue on pre-registration. *Journal of Consumer Psychology*, *31*(1), 146-150. https://doi.org/10.1002/jcpy.1211

Lalwani, A. K., Shavitt, S., & Johnson, T. (2006). What is the relation between cultural orientation and socially desirable responding? *Journal of Personality and Social Psychology*, *90*(1) 165-178. https://psycnet.apa.org/doi/10.1037/0022-3514.90.1.165

Li, X., Chan, K. W., & Kim, S. (2019). Service with emoticons: How customers interpret employee use of emoticons in online service encounters. *Journal of Consumer Research*, *45*(5), 973-987. https://doi.org/10.1093/jcr/ucy016

Lieberman, A., Morales, A. C., & Amir, O. (2019). Beyond the lab: Using data from the field to increase research validity. In F. R. Kardes, P. M. Herr, & N. Schwarz (Eds.), *Handbook of research methods in consumer psychology* (pp. 41-60). Routledge. https://doi.org/10.4324/9781351137713

Lim, D., Baek, T. H., Yoon, S., & Kim, Y. (2020). Colour effects in green advertising. *International Journal of Consumer Studies*, *44*(6), 552-562. https://doi.org/10.1111/ijcs.12589

Litman, L., Robinson, J., & Abberbock, T. (2017). TurkPrime.com: A versatile crowdsourcing data acquisition platform for the behavioral sciences. *Behavior Research Methods*, *49*, 433-442. https://doi.org/10.3758/s13428-016-0727-z

Lynch, Jr., J. G. (1999). Theory and external validity. *Journal of the Academy of Marketing Science*, *27*(3), 367-376. https://doi.org/10.1177/0092070399273007

Lynch, Jr., J. G., (1982). On the external validity of experiments in consumer research. *Journal of Consumer Research*, *9*(3), 225-239. https://doi.org/10.1086/208919

Macdonnell, R., & White, K. (2015). How construals of money versus time impact consumer charitable giving. *Journal of Consumer Research*, *42*(4), 551-563. https://doi.org/10.1093/jcr/ucv042

Malhotra, N. K., Agarwal, J., & Peterson, M. (1996). Methodological issues in cross-cultural marketing research: A state-of-the-art review. *International marketing review*, *13*(5), 7-43. https://doi.org/10.1108/02651339610131379

Maxwell, S. E. (2004). The persistence of underpowered studies in psychological research: Causes, consequences, and remedies. *Psychological Methods*, *9*(2), 147-163. https://psycnet.apa.org/doi/10.1037/1082-989X.9.2.147

McClelland, G. H., Lynch, Jr., J. G., Irwin, J. R., Spiller, S. A., Fitzsimons, G. J. (2015). Median splits, Type II errors, and false positive consumer psychology: Don't fight the power. *Journal of Consumer Psychology*, *25*(4), 679-689. https://doi.org/10.1016/j.jcps.2015.05.006

Meade, A. W., & Craig, S. B. (2012). Identifying careless responses in survey data. *Psychological Methods*, *17*(3), 437-455. https://doi.org/10.1037/a0028085

Meyvis, T., & Van Osselaer, S. M. J. (2018). Increasing the power of your study by increasing the effect size. *Journal of Consumer Research*, *44*(5), 1157-1173. https://doi.org/10.1093/jcr/ucx110

Montoya, A. K. (2019). Moderation analysis in two-instance repeated measures designs: Probing methods and multiple moderator models. *Behavior Research Methods*, *51*, 61-82. https://doi.org/10.3758/s13428-018-1088-6

Montoya, A. K., & Hayes, A. F. (2017). Two-condition within-participant statistical mediation analysis: A path-analytic framework. *Psychological Methods*, *22*(1), 6-27. https://psycnet.apa.org/doi/10.1037/met0000086

Morales, A. C., Amir, O., & Lee, L. (2017). Keeping it real in experimental research—Understanding when, where, and how to enhance realism and measure consumer behavior. *Journal of Consumer Research*, *44*(2), 465-476. https://doi.org/10.1093/jcr/ucx048

Muller, D., Judd, C. M., & Yzerbyt, V. Y. (2005). When moderation is mediated and mediation is moderated. *Journal of Personality and Social Psychology*, *89*(6), 852-863. https://doi.org/10.1037/0022-3514.89.6.852

Newman, G. E., Gorlin, M., & Dhar, R. (2014). When going green backfires: How firm intentions shape the evaluation of socially beneficial product enhancements. *Journal of Consumer Research*, *41*(3), 823-839. https://doi.org/10.1086/677841

Nichols, A. L., & Webster, G. D. (2013). The single-item need to belong scale. *Personality and Individual Differences*, *55*(2), 189-192. https://doi.org/10.1016/j.paid.2013.02.018

Niessen, A. S. M., Meijer, R. R., & Tendeiro, J. N. (2016). Detecting careless respondents in web-based questionnaires: Which method to use? *Journal of Research in Personality, 63*, 1-11. https://doi.org/10.1016/j.jrp.2016.04.010

Oppenheimer, D. M., Meyvis, T., & Davidenko, N. (2009). Instructional manipulation checks: Detecting satisficing to increase statistical power. *Journal of Experimental Social Psychology*, *45*(4), 867-872. https://doi.org/10.1016/j.jesp.2009.03.009

Paolacci, G., & Chandler, J. (2014). Inside the Turk: Understanding Mechanical Turk as a participant pool. *Current Directions in Psychological Science*, *23*(3), 184-188. https://doi.org/10.1177%2F0963721414531598

Pechman, C. (2019). How to publish consumer research based on experiments in the top marketing journals. In D. W. Stewart & D. Ladik (Eds.), *How to Publish in the Best Marketing Journals* (pp. 53-69). Edward Elgar. https://doi.org/10.4337/9781788113700.00016

Peer, E., Brandimarte, L., Samat, S., & Acquisti, A. (2017). Beyond the Turk: Alternative platforms for crowdsourcing behavioral research. *Journal of Experimental Social Psychology*, *70*, 153-163. https://doi.org/10.1016/j.jesp.2017.01.006

Perdue, B. C., & Summers, J. O. (1986). Checking the success of manipulations in marketing experiments. *Journal of Marketing Research*, *23*(4), 317-326. https://doi.org/10.1177%2F002224378602300401

Pham, M. T., & Oh, T. T. (2021a). On not confusing the tree of trustworthy statistics with the greater forest of good science: A comment on Simmons et al.'s perspective on pre-registration. *Journal of Consumer Psychology*, *31*(1), 181-185. https://doi.org/10.1002/jcpy.1213

Pham, M. T., & Oh, T. T. (2021b). Preregistration is neither sufficient nor necessary for good science. *Journal of Consumer Psychology*, *31*(1), 163-176. https://doi.org/10.1002/jcpy.1209

Pieters, R. (2017). Meaningful mediation analysis: Plausible causal inference and informative communication. *Journal of Consumer Research, 44*(3), 692-716. https://doi.org/10.1093/jcr/ucx081

Pirlott, A. G., & MacKinnon, D. P. (2016). Design approaches to experimental mediation. *Journal of Experimental Social Psychology*, *66*, 29-38. https://doi.org/10.1016/j.jesp.2015.09.012

Polman, E., Ziano, I., Wu, K., & Van Kerckhove, A. (2022). Consumers believe that products work better for others. *Journal of Consumer Research*, *49*(1), 132-153. https://doi.org/10.1093/jcr/ucab048

Pundak, C., Steinhart, Y., & Goldenberg, J. (2021). Nonmaleficence in shaming: The ethical dilemma underlying participation in online public shaming. *Journal of Consumer Psychology, 31*(3), 478-500. https://doi.org/10.1002/jcpy.1227

Raghubir, P., & Srivastava, J. (2002). Effect of face value on product valuation in foreign currencies. *Journal of Consumer Research*, *29*(3), 335-347. https://doi.org/10.1086/344430

Reips, U.-D. (2002). Standards for Internet-based experimenting. *Experimental Psychology*, *49*(4), 243-256. https://psycnet.apa.org/doi/10.1026/1618-3169.49.4.243

Robitaille, N., Mazar, N., Tsai, C. I., Haviv, A. M., & Hardy, E. (2021). Increasing organ donor registrations with behavioral interventions: A field experiment. *Journal of Marketing*, *85*(3), 168-183. https://doi.org/10.1177%2F0022242921990070

Rodas, M. A., & John, D. R. (2020). The secrecy effect: Secret consumption increases women's product evaluations and choice. *Journal of Consumer Research*, *46*(6), 1093-1109. https://doi.org/10.1093/jcr/ucz041

Rucker, D. D., McShane, B. B., & Preacher, K. J. (2015). A researcher's guide to regression, discretization, and median splits of continuous variables. *Journal of Consumer Psychology*, *25*(4), 666-678. https://doi.org/10.1016/j.jcps.2015.04.004

Savage, S. J., & Waldman, D. M. (2008). Learning and fatigue during choice experiments: A comparison of online and mail survey modes. *Journal of Applied Econometrics*, *23*(3), 351-371. http://dx.doi.org/10.1002/jae.984

Sawyer, A. G., Lynch, Jr., J. G., & Brinberg, D. L. (1995). A Bayesian analysis of the information value of manipulation and confounding checks in theory tests. *Journal of Consumer Research*, *21*(4), 581-595. https://doi.org/10.1086/209420

Schwarz, N. (2003). Self-reports in consumer research: The challenge of comparing cohorts and cultures. *Journal of Consumer Research*, *29*(4), 588-594. https://doi.org/10.1086/346253

Schwarz, N. (2019). Surveys, experiments, and the psychology of self-report. In F. R. Kardes,

P. M. Herr, & N. Schwarz (Eds.), *Handbook of research methods in consumer psychology* (pp. 17-40). Routledge. https://doi.org/10.4324/9781351137713

Sigall, H., & Mills, J. (1998). Measures of independent variables and mediators are useful in social psychology experiments: But are they necessary? *Personality and Social Psychology Review*, *2*(3), 218-226. https://doi.org/10.1207%2Fs15327957pspr0203_5

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*(11), 1359-1366. https://doi.org/10.1177%2F0956797611417632

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2021). Pre-registration: Why and how. *Journal of Consumer Psychology*, *31*(1), 151-162. https://doi.org/10.1002/jcpy.1208

Spencer, S. J., Zanna, M. P., & Fong, G. T. (2005). Establishing a causal chain: Why experiments are often more effective than mediational analyses in examining psychological processes. *Journal of Personality and Social Psychology*, *89*(6), 845-851. https://doi.org/10.1037/0022-3514.89.6.845

Spiller, S. A., Fitzsimons, G. J., Lynch, Jr., J. G., & McClelland, G. H. (2013). Spotlights, floodlights, and the magic number zero: Simple effects tests in moderated regression. *Journal of Marketing Research*, 50(2), 277-288. https://doi.org/10.1509/jmr.12.0420

Stoner, J. L., Loken, B., & Blank, A. S. (2018). The name game: How naming products increases psychological ownership and subsequent consumer evaluations. *Journal of Consumer Psychology*, *28*(1), 130-137. https://doi.org/10.1002/jcpy.1005

Thomas, K. A., & Clifford, S. (2017). Validity and Mechanical Turk: An assessment of exclusion methods and interactive experiments. *Computers in Human Behavior*, *77*, 184-197. https://doi.org/10.1016/j.chb.2017.08.038

Thompson, V. A., & Campbell, J. I. D. (2004). A power struggle: Between-vs. within-subjects designs in deductive reasoning research. *Psychologia, 47*(4), 277-296. https://doi.org/10.2117/psysoc.2004.277

To, R. N., & Patrick, V. M. (2021). How the eyes connect to the heart: The influence of eye gaze direction on advertising effectiveness. *Journal of Consumer Research*, *48*(1), 123-146. https://doi.org/10.1093/jcr/ucaa063

Torelli, C. J., Ahluwalia, R., Cheng, S. Y. Y., Olson, N. J., & Stoner, J. L. (2017). Redefining home: How cultural distinctiveness affects the malleability of in-group boundaries and brand preferences. *Journal of Consumer Research, 44*(1), 44-61. https://doi.org/10.1093/jcr/ucw072

Van de Ven, A. H. (1989). Nothing is quite so practical as a good theory. *Academy of Management Review*, *14*(4), 486-489. https://doi.org/10.5465/amr.1989.4308370

Vannette, D. (2017, June 28). *Using attention checks in your surveys may harm data quality*. Qualtrics. https://www.qualtrics.com/blog/using-attention-checks-in-your-surveys-may-harm-data-quality/

Viglia, G., & Dolnicar, S. (2020). A review of experiments in tourism and hospitality. *Annals of Tourism Research*, *80*, 102858. https://doi.org/10.1016/j.annals.2020.102858

Viglia, G., Zaefarian, G., & Ulqinaku, A. (2021). How to design good experiments in marketing: Types, examples, and methods. *Industrial Marketing Management*, *98*, 193-206. https://doi.org/10.1016/j.indmarman.2021.08.007

Wang, D., Weisstein, F. L., Duan, S., & Choi, P. (2022). Impact of ambivalent attitudes on green purchase intentions: The role of negative moods. *International Journal of*

Consumer Studies, *46*(1), 182-199. https://doi.org/10.1111/ijcs.12663

Wedel, M., & Dong, C. (2020). BANOVA: Bayesian analysis of experiments in consumer psychology. *Journal of Consumer Psychology*, *30*(1), 3-23. https://doi.org/10.1002/jcpy.1111

Wedel, M., & Dong, C. (2021). A tutorial on the analysis of experiments using BANOVA. *Psychological Methods*. Advance online publication. https://doi.org/10.1037/met0000408

Wedel, M., & Kopyakova, A. (2022). A guide to the Bayesian analysis of consumer behavior experiments with BANOVA using worked examples. *Australasian Marketing Journal*, *30*(1), 3-9. https://doi.org/10.1177%2F18393349211034169

Wong, N., Rindfleisch, A., & Burroughs, J. E. (2003). Do reverse-worded items confound measures in cross-cultural consumer research? The case of the material values scale. *Journal of Consumer Research*, *30*(1), 72-91. https://doi.org/10.1086/374697

Wright, S. A., & Goodman, J. K. (2019). Mechanical Turk in consumer research: Perceptions and usage in marketing academia. In F. R. Kardes, P. M. Herr, & N. Schwarz (Eds.), *Handbook of research methods in consumer psychology* (pp. 338-357). Routledge. https://doi.org/10.4324/9781351137713

Wu, F., Samper, A., Morales, A. C., & Fitzsimons, G. J. (2017). It's too pretty to use! When and how enhanced product aesthetics discourage usage and lower consumption enjoyment. *Journal of Consumer Research*, *44*(3), 651-672. https://doi.org/10.1093/jcr/ucx057

Zhao, X., Lynch, Jr., J. G., & Chen, Q. (2010). Reconsidering Baron and Kenny: Myths and truths about mediation analysis. *Journal of Consumer Research*, *37*(2), 197-206.

https://doi.org/10.1086/651257

FIGURES

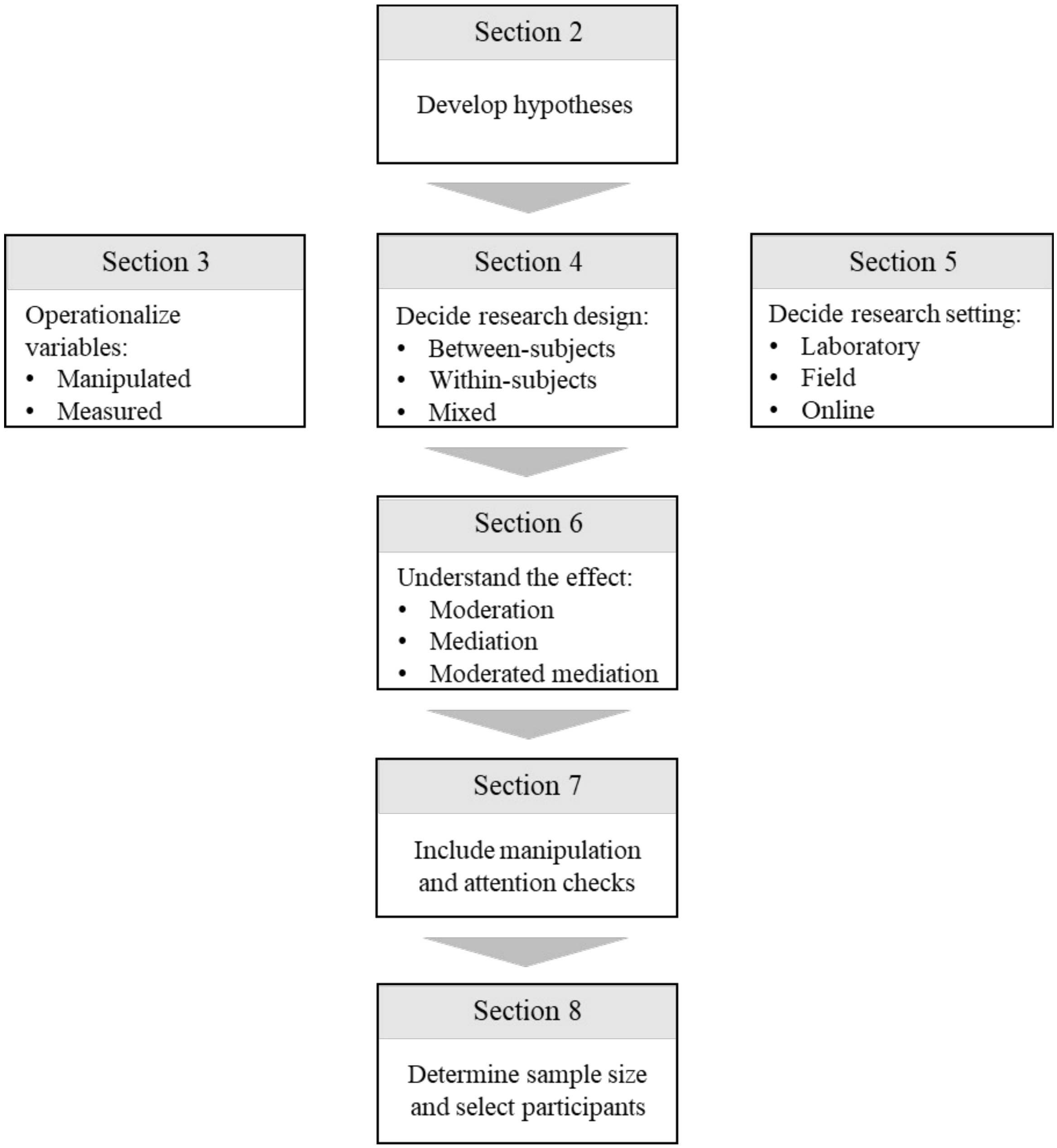

FIGURE 1 Topics researchers need to consider when designing experiments. The section number represents the corresponding section in this paper.

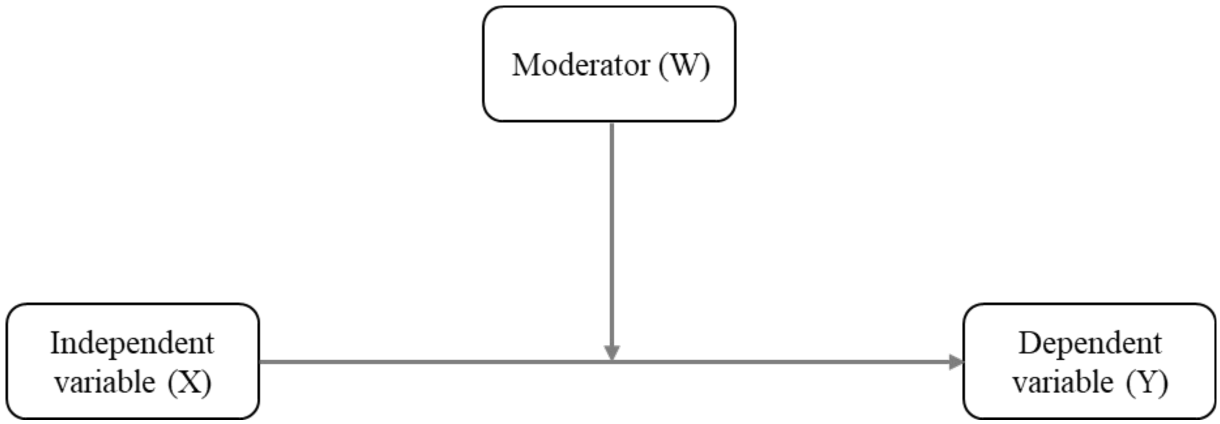

FIGURE 2 Conceptual model for moderation (PROCESS macro model 1; Hayes, 2022).

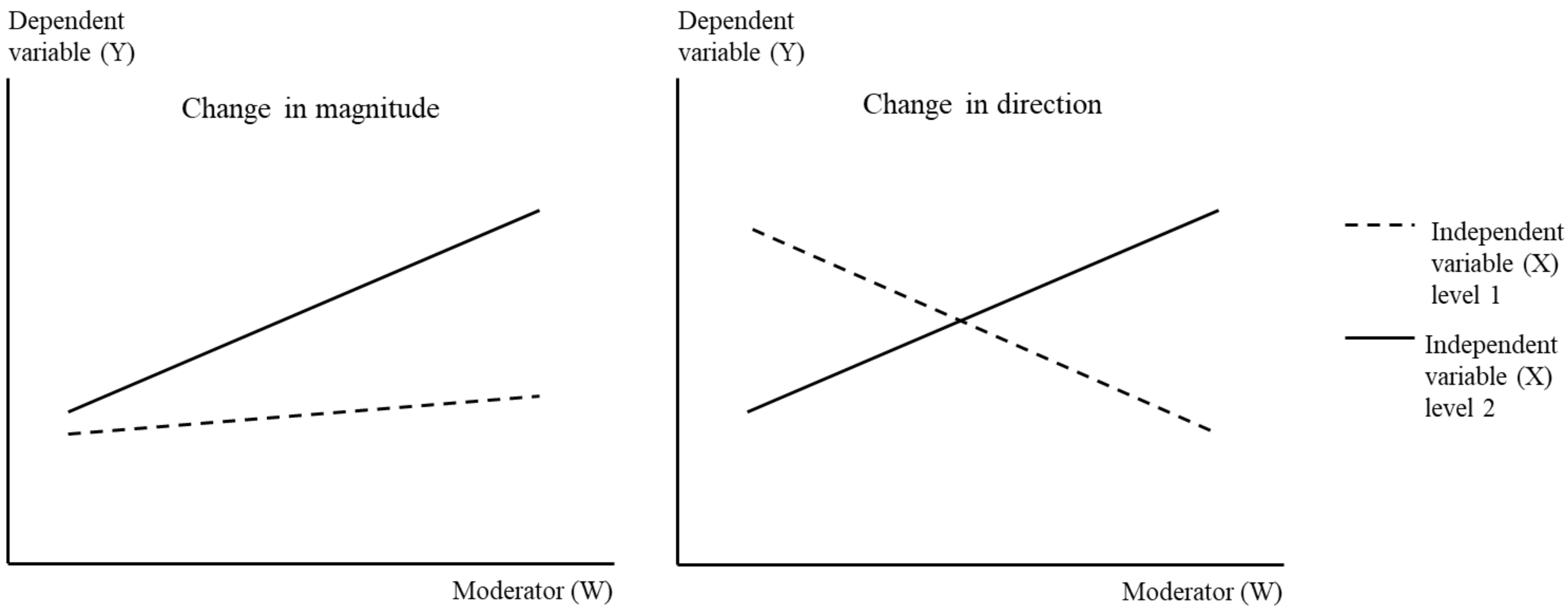

FIGURE 3A Measured moderation due to a change in magnitude (left) versus a change in direction (right).

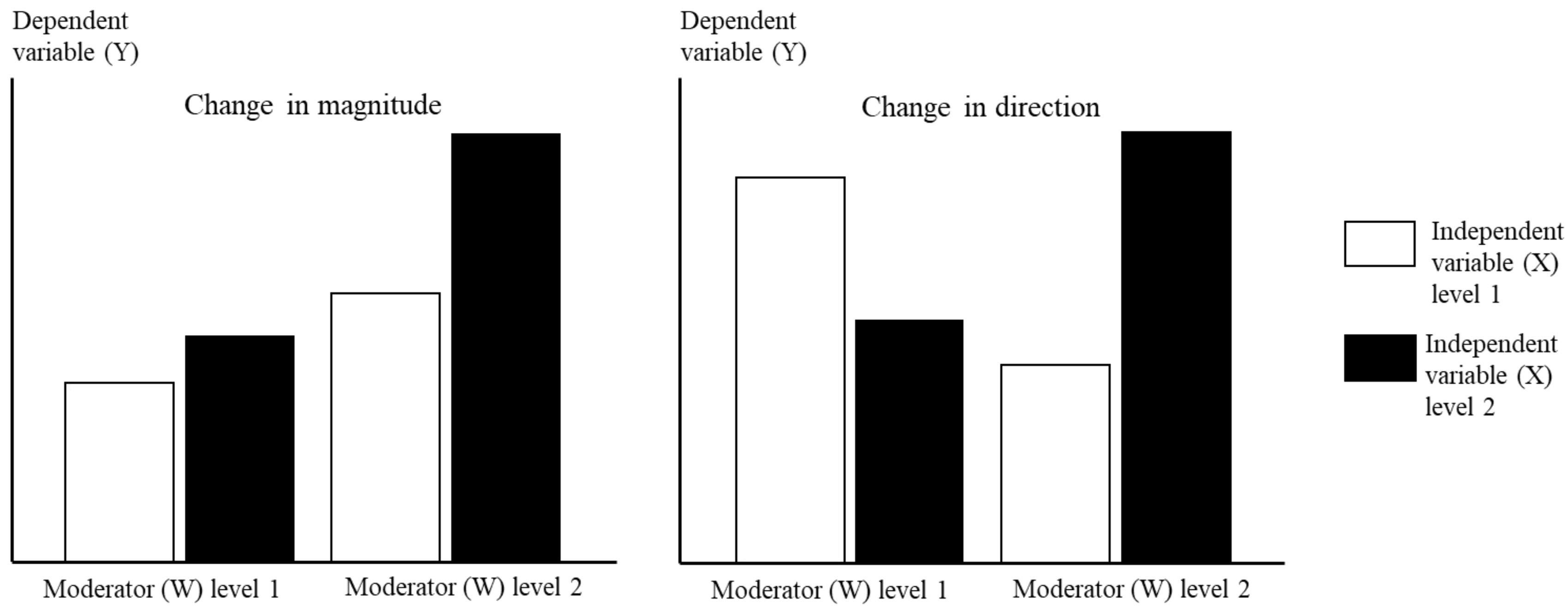

FIGURE 3B Manipulated moderation due to a change in magnitude (left) versus a change in direction (right).

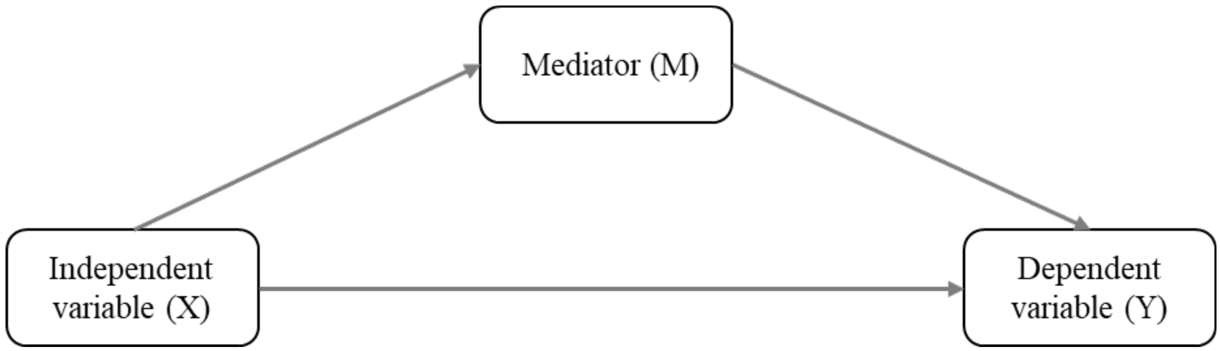

FIGURE 4 Conceptual model for simple mediation (PROCESS macro model 4; Hayes, 2022).

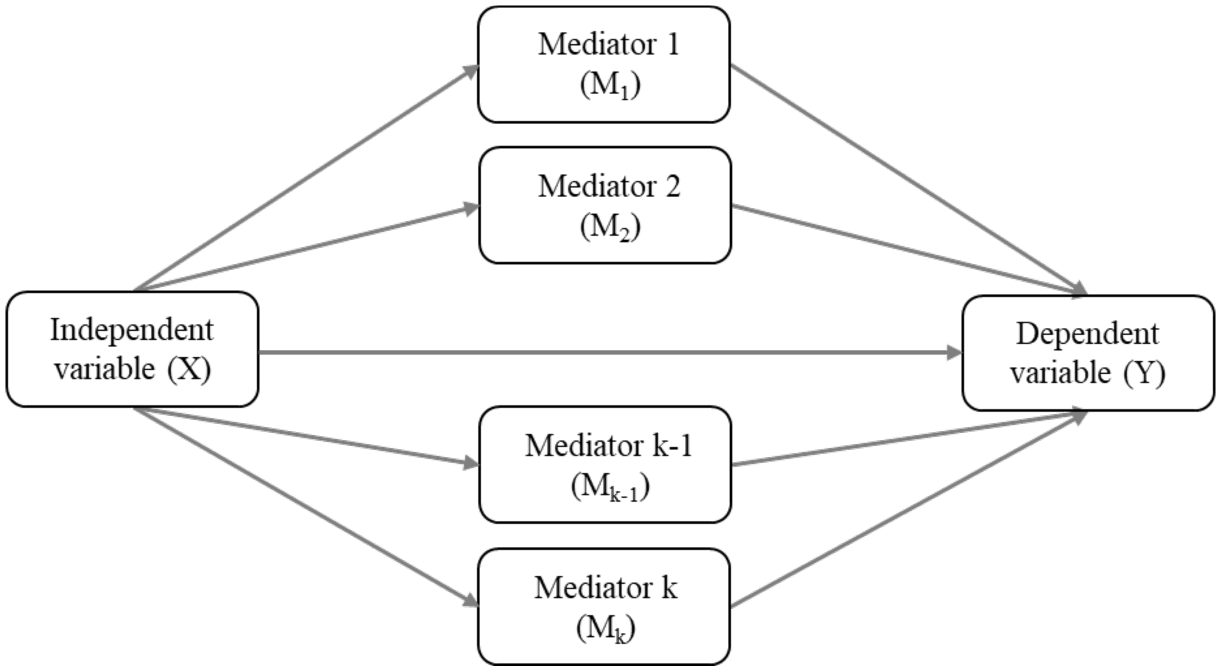

FIGURE 5 Conceptual model for parallel mediation (PROCESS macro model 4; Hayes, 2022).

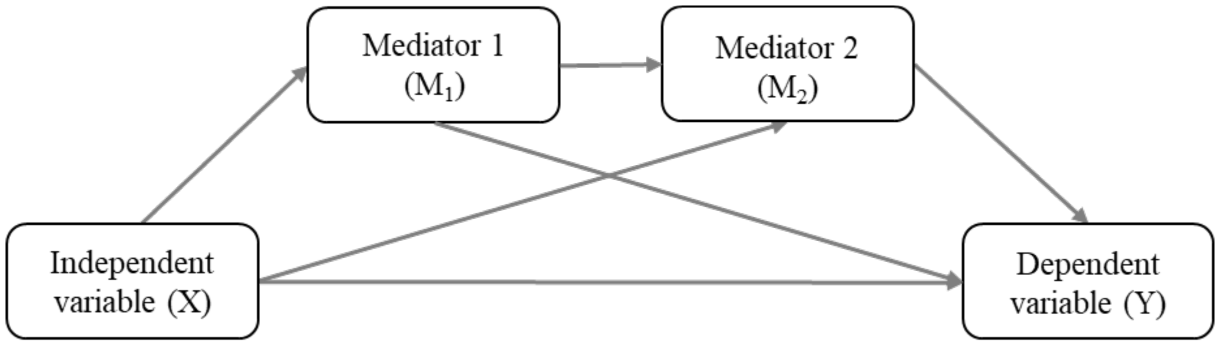

FIGURE 6 Conceptual model for serial mediation (PROCESS macro model 6; Hayes, 2022).

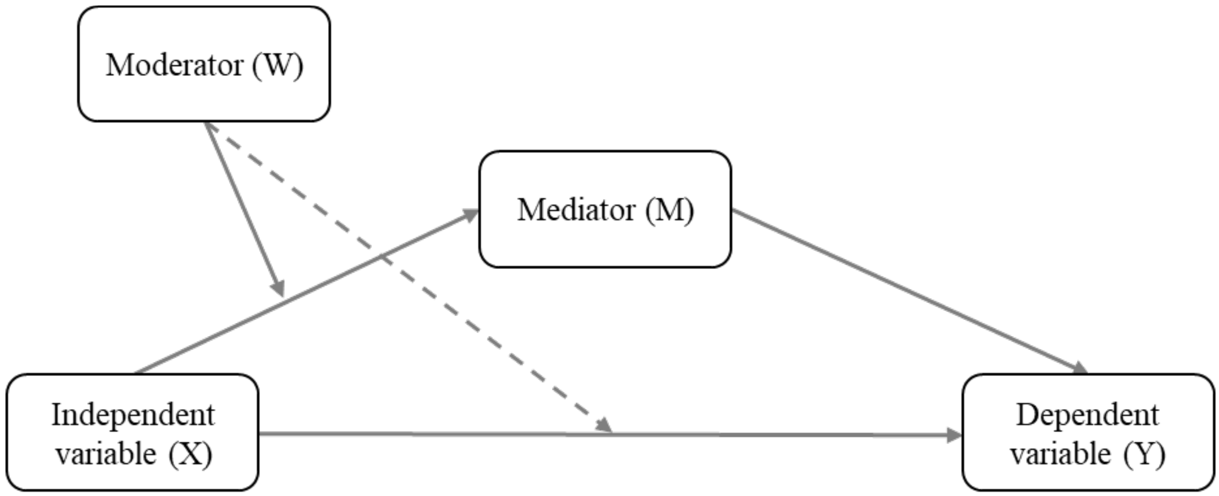

FIGURE 7 Conceptual model for moderated mediation with first stage moderation (PROCESS macro models 7 and 8; Hayes, 2022); the dashed line represents (optional) direct effect moderation (Hayes, 2015).

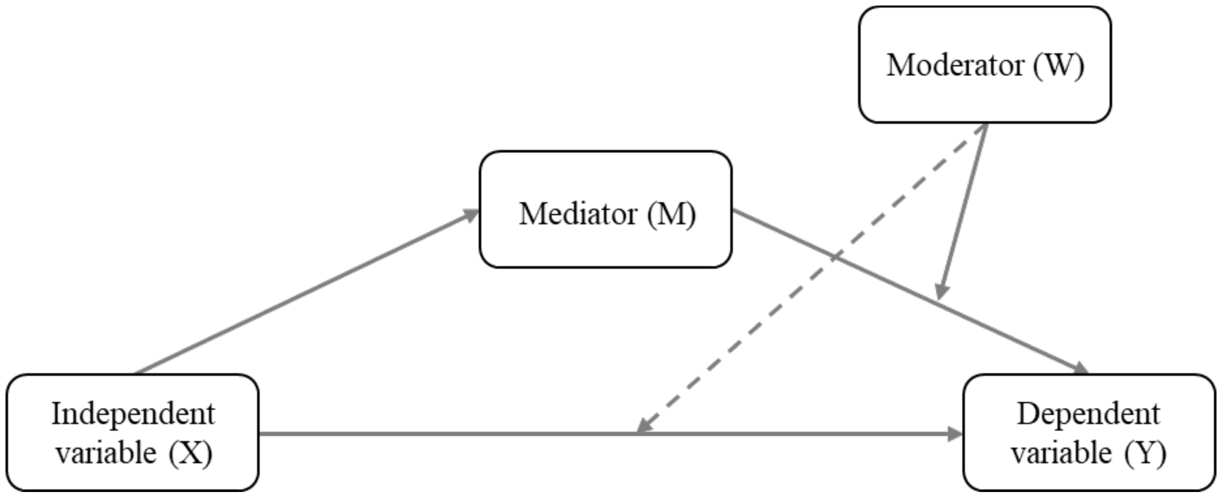

FIGURE 8 Conceptual model for moderated mediation with second stage moderation (PROCESS macro models 14 and 15; Hayes, 2022); the dashed line represents (optional) direct effect moderation (Hayes, 2015).