Overlooked or Undercooked?
Critical Review & Recommendations for Experimental Methods in Diversity Research

Jamie L. Gloor, PhD
Jamie.Gloor@Gmail.com
University of Zurich

Brook A. Gazdag, PhD
gazdag@bwl.lmu.edu
Ludwig-Maximilians-University Munich

Max Reinwald, MA
Max.Reinwald@uni-konstanz.de
University of Konstanz

Abstract

Diversity research often relies on experiments to make causal claims about the effects of various attributes (e.g., gender or race/ethnicity) on organizationally-relevant outcomes. This method allows scholars to analyze the effect of the attribute with all other information held constant. Until now, however, there remains a lack of clear, practical guidance on how to best study the effects of these attributes via experiments, including the theoretical and ethical implications that these design and method decisions might entail. Thus, we review the literature on experimental diversity research in organizations, highlighting illustrative publications and their design choices. Throughout, we highlight strengths, weaknesses, and potential pitfalls of these approaches, including a recent set of experimental studies on intersecting aspects of diversity by the first author of this chapter (Gloor, Li, & Puhl, 2018) to more practically exemplify some of the themes discussed within this chapter. We conclude with a discussion of ethical implications.

Key words: diversity, methods, stimuli, design

Citation: Gloor, J. L., Gazdag, B., & Reinwald, M. (forthcoming). Overlooked or undercooked? Critical review and recommendations for experimental methods in diversity research. A. Risberg, S. Just, & F. Villeseche (Eds.) *Routledge Companion to Organizational Diversity Research Methods*. Routledge.

Overlooked or Undercooked?
Critical Review & Recommendations for Experimental Methods in Diversity Research

"*An ounce of prevention is worth a pound of cure*."
-Benjamin Franklin, Founding Father of the United States (1706 - 1790)

Employee diversity is on the rise across the globe as more women join the workforce, more employees continue working beyond retirement, and people migrate across language, country, and continental boundaries. Thus, our need to understand these individual differences in organizational contexts and the meaningful implications of these characteristics, is also increasing. To do so, scholars can conduct experiments, which offer causal claims about particular characteristics while holding all other factors constant. For example, what is the effect of a person's gender on their selection for a leadership role, with all other factors made exactly the same (e.g., education, work experience, age, race/ethnicity, and performance)? Through controlling possible confounding factors by design, experiments offer several advantages over field studies conducted in organizations (i.e., non-experimental surveys) that may suffer threats of unmeasured variables or reverse causality, which may not only influence the accuracy of the findings, but also the interpretation of the results.

With random assignment, the foundational characteristic granting experiments their rigor and causality, field experiments that manipulate behavioral variables in a field setting (Scandura & Williams, 2000) are hailed as the "gold standard" in diversity research (King, Hebl, Morgan, & Ahmad, 2013). Field experiments are a rare method offering both internally valid *and* generalizable findings for many independent variables in organizational research, a field committed to creating "actionable knowledge" (Eden, 2017; p. 92). A few notable randomized field experiments have been recently published on diversity topics such as gender (e.g., Gloor, Morf, Paustian-Underdahl, & Backes-Gellner, in press), pregnancy (e.g., Morgan, Singletary, Hebl, & King, 2013), and obesity (e.g., Ruggs, Hebl, & Williams, 2015) in a range of organizational contexts, including leadership, hiring, and customer service (respectively). However, field experiments are not always possible.[1] For various reasons (e.g., fears about data privacy or sensitivity), organizations are often slow and unwilling to offer their practices and decision-making powers for monitoring or intervention. Similarly, field experiments in the context of hiring decisions typically involve fabricated applications that are submitted to real job advertisements. This procedure raises ethical concerns, such as wasting hiring managers' time and raising doubts about the legitimacy of future applications.

Hence, experiments based online or in the lab rather than in the field may seem more appealing due to some of the advantages they afford. Particularly web-based experiments may be increasingly convenient due to the recent growth of massive platforms for easily accessible and inexpensive samples of workers who are more representative of the population than typical lab studies such as Amazon's Mechanical Turk (MTurk). Indeed, MTurk can connect researchers with 400 participants to finish a 5-minute study within a few short hours and for only $1.50USD per person (personal experience of the first author). This creates cause for concern, as the ease with which experiments can be conducted may facilitate scholars' speed and incentivize their negligence in experimental design, which may threaten the very core of their experiments, thereby spoiling their results and the implications of these results. Thus, we aim to review the experimental literature on diversity research in organizations while highlighting the pros, cons, and potential pitfalls of various approaches

[1]See also Eden (2017) and Hauser, Linos, and Rogers (2018) for reviews and typologies of field experiments in organizational contexts, wherein the authors aim to dispel several misconceptions about field experiments and the challenges associated with such a design.

and topics, as well as ethical concerns. In doing so, we maintain a more practical focus, as we strive to provide specific, actionable strategies for those who will conduct, review, and disseminate diversity experiments.

Scholars have recently published several excellent reviews and recommendations for experiments (e.g., Lonati, Quiroga, Zehnder, & Antonakis, 2018), experimental vignette studies (e.g., Aguinis & Bradley, 2014), field experiments and interventions (e.g., Bodner & Bliese, 2018; Eden, 2017) written for audiences of organization scholars. Yet, by and large, these reviews take a broad lens, overlooking key issues and design choices that are particularly pernicious to and relevant for experimental diversity research. Although King and colleagues' (2013) review of field experiments on "sensitive organizational topics" (e.g., diversity) is a notable exception, we largely focus our review on lab- and web-based experiments, while also offering more minute, practical, and actionable strategies. In these ways, this chapter contributes to the literature on experimental methods in organizations by concentrating specifically on diversity research while also maintaining a relatively general, applied focus to touch on several topics and reach various kinds of readers (e.g., scholars, students, and practitioners).[2]

Although we have described our key aims, we also need to delineate what we do *not* cover in the scope of this short chapter. First, we focus our review on the more visible, "surface-level" forms of diversity rather than the less visible or "deep-level" forms of diversity,[3] as the former are arguably more easily manipulated and thus lend themselves more readily to experimental approaches, which is the primary focus of our methodological review and recommendations.[4] Second, we focus on studies examining perceptions of, attitudes towards, and behaviors in response to persons from various demographic groups rather than the reverse (e.g., if leadership evaluations differ when rated by a woman or a man). Given space constraints, we cannot adequately address both types of studies; the former also tends to be the more dominant approach in organizational diversity research. Similarly, we acknowledge the value of experimental studies examining diversity at the team-level (e.g., Pearsall, Ellis, & Evans, 2008), and the important implications those studies have for understanding diversity in context and its effects on organizationally-relevant outcomes. Yet, it is also beyond the scope of this chapter to review those cases in adequate detail. Third and finally, we focus on issues of design and measurement rather than on statistics and analyses. Although certain theoretical models enable endogeneity, making standard analyses inappropriate for experiments, we also outline ways to eliminate endogeneity by design.

Review of the Literature

To provide an overview of the current practices in experimental diversity research, we first reviewed the most recent decades of research published in psychology and management journals. We alphabetically highlight illustrative–but not exhaustive–examples of types of different and their respective manipulations conducted in the organizational literature.

Age

[2]For ease of comprehension, however, we subsequently refer to this group more succinctly as "scholars", as regardless of one's title or profession, conducting an experiment is a scientific, scholarly endeavor.
[3]However, there are notable examples of excellent experiments on more "deep-level" diversity in organizations, such as Grant, Gino, and Hofmann (2011), who manipulated personality (i.e., extraversion) in the context of leadership and employee proactivity.
[4]Indeed, Aguinis and Bradley (2014) would characterize our focus on explicit processes as "paper people studies" (compared with other methods assessing implicit processes and outcomes, such as policy capturing and conjoint analysis).

Age refers to the number of years someone has lived but could also include the generation to which someone belongs. This aspect of diversity has received relatively less attention. Given its continuous nature, biological age is often split into two groups: younger and older adults. Vignette-based research typically uses a specific age that they provide in the description of the interaction to manipulate the target person's age (e.g., "imagine you are interacting with someone who is 75 years old"; Prior & Sargent-Cox, 2014). Age can also be manipulated via an applicant's birthdate on a resume (i.e., Krings et al., 2011) or by providing a photograph in which an applicant appears young or old (Kaufmann, Krings & Sczesny, 2015). Future work may consider how salient age gaps between participants and targets influence perceptions.

Body Weight

Body weight is a visible characteristic, while obesity is a specific category implying greater body weights as determined through a specific cut-off of one's Body Mass Index (BMI) of 30 or more (Puhl & Peterson, 2014). King and colleagues (2006) dressed confederates with or without a body weight prosthetic to visit retail stores while another confederate nearby evaluated the subsequent customer service. Agerström and Rooth (2011) manipulated photos to represent obese and non-obese persons in a resume. Issues to consider when studying body weight include gendered perceptions of BMI (i.e., the same body weight may be perceived differently for men and women); weight perceptions may also differ depending on how the weight is distributed on a person's body. When body weight is manipulated via Curriculum Vitae (CV), it is focused on the face and therefore difficult to portray the entire body size. Finally, manipulated photos should be pre-tested to ensure they are viewed as realistic.

Gender

Of the different diversity topics, gender might be one of the most common, with the literature typically focused on binary gender (i.e., comparing male and female targets). There are a variety of different ways to manipulate gender. Typically, this is done via "paper people" or vignette studies, but it can also be implemented via simulation studies. Stimulus materials are often simple, comprising a name or an image that has been chosen or manipulated to indicate different gender categories. To illustrate the former, in a vignette study, Shaughnessy and colleagues (2015) presented participants with a negotiation transcript belonging to a man or a woman by simply changing the name (e.g., "JoAnna" or "Joseph"). To illustrate the latter, Brosi and colleagues (2016) investigated pride by presenting participants with images of men and women expressing pride or a more neutral emotion. Naturally, these techniques can also be used together or in combination with another diversity-related manipulation. For example, Latu and colleagues (2013) used Immersive Virtual Environment Technology to place their participants in a virtual reality room and give speeches to a digital audience. Within the virtual room, they included a picture of a male or a female role model.

Mental and/or Physical Ability

There are a wide variety of types and spectrums of ability to include under this heading. Some research focusing on mental disabilities compared individuals with Dyslexia to those without in a video-based studied that mentioned the disability or not (Colella et al., 1998). Physical disabilities could include functioning and mobility, but also physical impairments. For example, Madera and Hebl (2012) compared applicants with or without facial stigmas via interview videos. Despite the wide range of potential diversity variables and its suitability for experimental study, little research has focused on these aspects.

Nationality and Spoken Accents

In this context, accented spoken language signals that a person is a non-native speaker and likely hails from a nation other than the local or dominant nationality. Nationality as signaled by language can play a unique role compared to race as described above. For example, Huang and colleagues (2013) compared the political skill perceptions of White and Asian nonnative speakers of English using a recorded set script. In a classic example, Rubin (1992), provided participants with a recording of the same lecture presented with a picture of a White or an Asian instructor; merely the perception that the lecture could be accented influenced participant ratings. A logical extension of this research could include a more explicit consideration of a person′s migration history and the cultural or physical distance between the focal and the target nationalities. As Harrison et al. (in press) readily acknowledge, a field study would be logistically challenging, but for this reason, it is ripe for experimental study.

Pregnancy & Parenthood

Pregnancy and/or parenthood are less often considered under the umbrella of workplace diversity. Pregnancy is rather unique, as this type of diversity is temporary and transitional. In other words, pregnancy lasts only about 9 months, and depending on the stage of development, it can be invisible (i.e., typically up to about 3 months) or visible (i.e., typically after about 6 months). Thus, these features present particularly unique considerations for pregnancy research. Similar to obesity, prostheses have been used to simulate a pregnancy around 6 months (Hebl et al., 2007). Furthermore, the confederate was dressed in a business causal maternity outfit and wore a wedding band. In a follow up study, the researchers used an experimental vignette study through which they subtly indicated that the woman was expecting a child (i.e., "preparing a new nursery at home;" Hebl et al., 2007). Although non-parent, pregnant, and parent seem like distinct categories of people, the lines may be blurring, particularly for childbearing-aged women (see Gloor, Li, Lim, & Feierabend, 2018).

Race

Race or ethnicity refers to a class or kind of people unified by a common origin. Racial diversity has been studied using similar approaches as gender diversity. In a recent article, Hernandez and colleagues (in press) investigated both participants′ race and target race. In their first vignette study, participants were presented with a prospective job seeker′s profile that only differed in the image at the top (e.g., a Black/African American or White/Caucasian photo). As in this study, research on racial diversity often focuses on White/Caucasian compared with another group (e.g., Black/African American or Asian). However, there remains a need to understand the nuances not only between, but also within racial groups. For example, Ma and colleagues (2018) used images of Black men and women that varied in the degree to which they fit the prototypical ideal (i.e., racial prototypicality). These findings show a need for future work to consider more discrete racial features when designing their stimuli, such as skin tone and facial features that may be distinct to a particular racial group. For a more in-depth discussion of research on race, see Stone and colleagues (2008).

Religion

Religion refers to "a personal set or institutionalized system of religious attitudes, beliefs, and practices" (Merriam-Webster). To date, the literature on stereotypes of religious differences

has investigated differences between Muslim and non-Muslim individuals (for example) as signaled by their attire. In a study on reactions to job applications, King and Ahmad (2010) provided photos of applicants wearing traditional Muslim attire or non-religious attire. In another study, Everett and colleagues (2015) used photos of women either wearing a complete veil, a head scarf, or no head covering. Some earlier work by Greenberg and colleagues (1990) studied impressions of Christians and Jewish people by explicitly providing participants with the target's religious affiliation of the target person.

Sexual Orientation

Sexual orientation refers to an individual's preference for sexual partners or identification. To date, most of the research focuses on comparing perceptions of (perceived) heterosexual individuals to (perceived) homosexual individuals. Rule and colleagues (2014) explicitly labeled faces in the stimulus materials with "straight" or "gay." In another study, Rule and colleagues (2016) used a database of photos that had been pre-tested for being perceived as gay or straight to examine the effect of sexual orientation on suitability for a leadership position. They also adapted field-sourced online profiles of gay and straight men. A particular challenge within the scope of sexual orientation research relates to its rather "invisible" nature, which may make it particularly hard to convey via experimental stimuli. Furthermore, the number of categories under the topic of sexual orientation are continuing to increase as the general awareness and understanding increases. (i.e., LGBTI).

In summary, there is a large selection of experimental manipulations that can be used to investigate various forms of diversity in a range of organizational settings and applications. We commented on some challenges for each attribute, but we summarize some point again in Table 1, and expand on others with a bit more detail in the following section.

Overlooked or Undercooked?

The details of designing high-quality diversity stimuli and experiments are not always comprehensively or explicitly listed in published works. Thus, it may be difficult for early career scholars and non-experimentalists to discern how to conduct high-quality work in this field, even with well-published exemplars. To fill this gap and reduce these knowledge-based inequalities, we highlight several–but not all–of these more implicit assumptions here.

Demand Effects

Demand effects, or how subjects infer experimenters' expectations of them via cues in experimental settings, are among the classical issues in experimental methods (Klein et al., 2012). Demand effects can elicit upward- or downward-biased effects, as participants respond in ways to prove or disprove hypotheses (respectively). To reduce demand effects, scholars can use several strategies that are often relatively easy to implement. For example, refrain from explaining the detailed purpose of the study at the beginning of the session.[5] Filler items that are unrelated to the research question(s) can also occlude the true purpose of the experiment. Scholars can also ask participants at the end of the experiment what they think was the purpose of the study, possibly excluding or controlling for these participants

[5]Yet, we strongly advise against misrepresenting or deceiving participants, as it is a questionable practice from an ethical perspective and can confound effects (for a critical discussion, see Lonati et al., 2018).

later. Social desirability scales can also be included to identify and possible control for the extent to which participants may have answered in a self-favoring manner (see Uziel, 2010).

Diversity of Outcomes and Designs

When possible, scholars should measure a range of outcomes in their experiments. We recommend including both perceptual and attitudinal measures along with behavioral measures. In the context of diversity experiments, this could include how a diversity attribute is perceived by a participant (e.g., how warm is a female target compared with a male target) as well as how a participant's behavior is influenced by the target's attribute (e.g., active facilitation behavior such as helping in response to perceived warmth). By considering real behaviors, scholars can demonstrate diversity's more tangible consequences, thereby increasing external validity. Perceptions and attitudes remain important though, as they provide insights into the underlying processes driving the observed behavior.

In a similar vein, lab- and web-based experiments may offer advantages compared to field studies when it comes to establishing causality, but they also have drawbacks, such as a limited generalizability to real-world settings. To capitalize on the strengths of both designs and limit problems with one or the other, scholars can increase the overall rigor of their program of research by combining lab- and/or web-based experiments and field studies.

Manipulation Checks

Manipulation checks, which are also sometimes characterized as attention checks or comprehension checks, refer to the questions that experimentalists ask to ensure that participants noticed the experimental manipulation and interpreted it as it was intended. In other words, manipulation checks are an indication of construct validity: did participants notice Jane's name on the CV, and thus, recognize and remember that they saw a woman's CV? These questions are essential, because unless the stimuli are already pre-tested with a non-overlapping sample[6], how else can one be sure that they are manipulating what they intended to manipulate? These questions should be asked at the end of the study to reduce demand effects (to be described in more detail later). Some scholars use these as exclusion criteria (i.e., filtering out participants with incorrect responses), while others use these responses to indicate the strength of the manipulation and/or as control variables in analyses.

Paper People

All too often, scholars manipulate categories of interest with a name and/or gendered pronouns in a CV, without a photograph or any visual indication of the target's appearance. Such designs may inherently breed confounds that cloud scholars' ability to draw appropriate causal claims from their experiment, due to the reasons we mentioned above in the section on names (see Kasof, 1993; Simonsohn, 2015). Although such a design may be very easy to implement, is relevant for some organizational decisions (e.g., interview decisions in a hiring process), and maintains an air of vagueness and sense of increased generalizability (i.e., by not describing or showing a target's race, the target may be assumed to hail from different races), it lacks the richness of real-world interactions. We recommend manipulations that are as rich and as real as possible (e.g., videos or virtual reality; see Latu et al., 2013). Another strategy is to use multiple types of different kinds of manipulations to ensure one elicits

[6]That is, a different set of participants than those who will take part in the main survey

similar effects regardless of the design, simultaneously reducing the possibility that certain effects may be an artifact of the method (e.g., see Gloor et al., 2018, who used both names and manipulated photos to indicate gender and body weight in Study 1, but gendered pronouns and textual descriptions of body weight in Study 2).

Participant Diversity and Data Quality

Web-based platforms such as Mturk provide access to samples with similar demographic and statistical properties as compared with other sampling approaches (Buhrmester, Kwang, & Gosling, 2011). In terms of participant diversity, workers such as Mturkers provide access to working individuals across different regions and to diverse industries (Woo, Keith & Thorton, 2015). However, this might also increase the variability in your responses, so including a few extra questions can provide you with important insights into your data quality. For example, ask participants if they were distracted while taking the survey and what they think was the true purpose of the study. With these questions, you can identify distracted participants who may add noise to your data and potentially eliminate those who figured it out, which might (consciously or unconsciously) influence their responses.

Stimuli

Stimuli include the materials provided to participants with which scholars seek to elicit perceptions, attitudes, behavioral intentions, and/or actual behaviors. They can take on a range of different forms including CVs, photos, or videos. One's stimuli are perhaps the most important parts of experiments, yet key aspects are often overlooked. For example, in a study of the effects of gender on leadership ratings, a scholar might use one name to represent the female and male leader (e.g., Jennifer and John, respectively). This is problematic, because by having only one example for each group, scholars cannot show if their effects are driven by a comparison of the broader gender categories, women and men, or just the two fictitious people, Jennifer and John. To avoid this issue, we recommend that scholars use multiple stimuli for each category. Similarly, scholars should be thoughtful when selecting names, as they imply (or the ratee/participant infers) gender, race/ethnicity, and socio-economic status, which may undermine one's results and implications (see Kasof, 1993; Simonsohn, 2015).

Time

The vast majority of experiments pay little thought and attention to time, assessing the effects of the manipulation on dependent variables after one brief exposure. This means that time is only incorporated in terms of the few quick minutes it happened to take for the participant reviewed the stimuli and answered the questions. Although such a design may be relevant for some organizational decisions (e.g., hiring), multiple exposures are also often relevant even in these cases (e.g., an initial screening, followed by a first or informal meeting, then an interview). Indeed, many organizational phenomena include multiple interactions, which are often not present, yet are absolutely possible to achieve via experimental designs. Repeated measures also allow scholars to delve into dynamic effects (see Ployhart & Vandenberg, 2010). Repeated measures may be easier to collect via experiments, offering the advantage to measure and control for participant fixed effects (which we explain in more detail later).

Ethical Considerations

Finally, there are also several ethical considerations related to the approach, design, and dissemination of experimental research on diversity in organizations. Consistent with the general approach of this research methods anthology, we highlight five issues here.

Binary Approaches

Experiments in particular lend themselves to conceptualizing diversity with binary categories. This is perhaps most clearly evident in the organizational diversity space with empirical studies of gender bias and stereotyping, as a binary consideration of gender has dominated empirical investigations (i.e., "men" and "women"). However, this oversimplifies the continuous construct of gender, transforming it instead into dichotomous, biological sex, thereby also overlooking individuals who identify outside of these two categories. As gender (and other forms of diversity) may be best understood as a spectrum, and even though manipulations need to be categorical given experimental design constraints, scholars can include more than two categories and integrate continuous measures as manipulation checks. For a more in-depth discussion of this topic, see Morgenroth and Ryan (2018).

Intersectionality

The idea of intersectionality originated in black feminism scholarship by Crenshaw, (1989) and can be defined as "overlapping social categories, such as race and gender, that are relevant to a specific individual or group's identity and create a unique experience that is separate and apart from its originating categories" (Rosette, Ponce de Leon, Koval, & Harrison, 2019, p. 3). Intersectionality threatens the generalizability of diversity research's implications because the dominant approach is to focus on one category or a single aspect of diversity. Although certainly a valuable approach that has produced many important insights, such single category approaches inherently mean that scholars overlook how belonging to multiple categories may completely alter a person's identity and experience. In other words, being a female or being Black elicits qualitatively different leadership ratings than being a Black female (see Hall, Hall, Galinsky, & Phillips, in press; Rosette et al., 2019). In the case of gender, for example, well-intended researchers may think they are studying a phenomenon applicable to nearly half the population when they study the effects of gender. However, for topics such as emotions, pay, and leadership potential, intersectional is the more accurate *and* more appropriate approach to understand the reactions to–and thereby, also the experiences of–individuals who belong to multiple social groups or categories.

To illustrate intersectionality, and consistent with the general approach of this research methods anthology, we describe a recently published experimental paper by the first author (Gloor et al., 2018). In three experiments, the authors examined the effects of employee gender and body weight on coworker support for parental leave, using different manipulations for gender and body weight, including names and manipulated photos (Study 1 and 3), and gendered pronouns and textual descriptions of body weight (Study 2). In this way, the authors replicated evidence of their predicted effects as well as demonstrating converging results across different types of manipulations. Of note, since the two diversity categories also significantly interacted with each other, this also provides evidence of an intersectional effect, such that being a woman or being a person with obesity were both qualitatively *and* measurably distinct from being an obese woman, at least in terms of the outcomes studied here (e.g., parenting expectations and received parental leave support).

Endogeneity

Although it is out of scope to go into great detail, and endogeneity has been reviewed at length elsewhere (e.g., Antonakis, Bendahan, Jacquart, & Lalive, 2010; 2014) the basic gist is that your claims may no longer be causal even though you conducted an experiment. In designs with mediators, for example, one often manipulates the independent variable, but then measures both the mediator and the outcome(s) as self-reported measures from the same source. This means that endogeneity could potentially threaten the accuracy and stability of the estimate between the mediator and the outcome, highlighting an ethical concern if scholars inaccurately think they have causal findings.

Thus, to test for and ideally eliminate endogeneity then, we suggest four different strategies that can be incorporated already at the design stage. Perhaps the simplest way to reduce (but not entirely eliminate) endogeneity is to collect responses from different sources. Another option is to conduct a causal chain design (Spencer, Zanna, & Fong, 2005), wherein a series of experiments is conducted to test a model with a mediator rather than just one experiment; for an illustrative example in the context of organizational diversity research, see Windscheid, Bowes-Sperry, Kidder, Cheung, Morner, and Lievens (2016). Panel designs (i.e., data collected multiple times from the same participants) can also reduce endogeneity, because scholars can calculate a fixed effect associated with the participant that captures unobserved sources of variance. Finally, if data is already collected, then you might need an instrumental variable approach to check and correct for it (see Antonakis et al., 2010; 2014).[7]

Compensation Practices

People should be incentivized participate in experiments to make responses more consequential and reduce demand effects (see Lonati et al., 2018). Economics offers financial compensation, which is also common in management, but participants often receive course credit in psychology. To our knowledge, direct evidence explicitly comparing these approaches in terms of their effects on outcomes is lacking. Yet there may be concerns of data quality due to low paid workers who begrudgingly or hastily participate (Buhrmeister et al., 2011). We extend these arguments to propose that experimenters also pay participants a fair wage. If using the "livable wage" as a reference, one can generally estimate a minimum of $15USD per hour (although amounts vary by location; see wageindicator.org).

Bias

Last but not least, of all the topics to study, diversity researchers in particular may be drawn to the topic by personal, philosophical, and/or political motives. This is reflected in the scholarly adage, "research is me search" (Gloor, 2014). We believe that passion for and personal experience with a topic can facilitate important insights. Yet, one's motives for pursuing a certain scientific area may also threaten the transparency or accuracy of said science. In the case of gender, for example, scholars may–at varying levels of consciousness–aim to uncover a bias indicative of male advantage and/or female disadvantage, brushing off findings that are inconsistent with these beliefs as attributable to external reasons (e.g., measurement error). Similarly, when disseminating findings, scholars may oversimplify the science in favor of their beliefs. To illustrate again with gender, scholars may advertise a broad, female leadership advantage, which science clearly shows is more contextual and

[7]However, this approach requires an instrumental variable (i.e., exogenous variables that do not depend on other variables or disturbances in the system of equations; Antonakis et al., 2010), which could also present a challenge if data is already collected. Thus, it may be worthwhile to include variables in your design that could serve as instruments later (on how to find instruments, see Antonakis et al., 2010).

complex (see Eagly; 2016, 2018). For these reasons, scholars should stay vigilant throughout the research process to remain "honest brokers" of diversity science, ensuring that the discovery of favorable *and* unfavorable information is subject to scrutiny (King et al., 2013).

Conclusions

Experiments have great potential in management research to bring clarity and causal claims to an important area of organizational research. Until now, experiments have been underutilized and undervalued in management research, perhaps because they have been largely misunderstood (e.g., see Highhouse, 2009; and Podsakoff & Podsakoff, in press). We hope this chapter encourages more diversity scholars to contribute important causal evidence to improve our knowledge of diversity in organizations through experimental studies, particularly with an ounce of prevention to avoid the pound of pitfalls we have raised here.

References

Agerström, J., & Rooth, D.-O. (2011). The role of automatic obesity stereotypes in real hiring discrimination. *Journal of Applied Psychology*, *96*(4), 790–805.

Aguinis, H., & Bradley, K. J. (2014). Best practice recommendations for designing and implementing experimental vignette methodology studies. *Organizational Research Methods, 17*(4), 351-371.

Antonakis, J., Bendahan, S., Jacquart, P., & Lalive, R. (2014). Causality and endogeneity: Problems and solutions. In D. V. Day (Ed.), *The Oxford Handbook of Leadership and Organizations* (pp. 93-117). New York: Oxford University Press.

Antonakis, J., Bendahan, S., Jacquart, P., & Lalive, R. 2010. On making causal claims: A review and recommendations. *The Leadership Quarterly, 21*(6), 1086-1120.

Bodner, T. E., & Bliese, P. D. (2018). Detecting and differentiating the direction of change and intervention effects in randomized trials. *Journal of Applied Psychology, 103*(1), 37-53.

Brosi, P., Spörrle, M., Welpe, I. M., & Heilman, M. E. (2016). Expressing pride: Effects on perceived agency, communality, and stereotype-based gender disparities. *Journal of Applied Psychology*, *101*(9), 1319–1328.

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk. *Perspectives on Psychological Science*, *6*(1), 3–5.

Colella, A., DeNisi, A. S., & Varma, A. (1998). The impact of ratee's disability on performance judgments and choice as partner: The role of disability–job fit stereotypes and interdependence of rewards. *Journal of Applied Psychology*, *83*(1), 102–111.

Crenshaw, K. W. (1989). Demarginalizing the intersection of race and sex: A black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics. *The University of Chicago Legal Forum*, 139-167.

Eagly, A. H. (2016). When passionate advocates meet research on diversity, does the honest broker stand a chance? *Journal of Social Issues, 71*(1), 199-222.

Eagly, A. H. (2018). The shaping of science by ideology: How feminism inspired, led, and constrained scientific understanding of sex and gender. *Journal of Social Issues, 74*(4), 871-888.

Eden, D. (2017). Field experiments in organizations. *Annual Review of Organizational Psychology and Organizational Behavior, 4*, 91-122.

Everett, J. A. C., Schellhaas, F. M. H., Earp, B. D., Ando, V., Memarzia, J., Parise, C. V., … Hewstone, M. (2015). Covered in stigma? The impact of differing levels of Islamic head-covering on explicit and implicit biases toward Muslim women. *Journal of Applied Social Psychology*, *45*(2), 90–104.

Gloor, J. L. (2014). Bullying and harassment in the workplace: Developments in theory, research, and practice. *Academy of Management: Learning & Education, 13*(1), 145-148.

Gloor, J. L., Li, X., Lim, S., & Feierabend, A. (2018). An inconvenient truth? Interpersonal and career consequences of "maybe baby" expectations. *Journal of Vocational Behavior, 104,* 44-58.

Gloor, J. L., Li, X., & Puhl, R. M. (2018). Predictors of parental leave support: Bad news for (big) dads and a policy for equality. *Group Processes & Intergroup Relations, 21*(5), 810-830.

Gloor, J. L., Morf, M., Paustian-Underdahl, S., & Backes-Gellner, U. (in press). Fix the game, not the dame: Restoring equity in leadership evaluations. *Journal of Business Ethics*, 1-15.

Grant, A. M., Gino, F., & Hofmann, D. A. (2011). Reversing the extraverted leadership advantage: The role of employee proactivity. *Academy of Management Journal, 54*(3), 528-550.

Greenberg, J., Pyszczynski, T., Solomon, S., Rosenblatt, A., & et al. (1990). Evidence for terror management theory II: The effects of mortality salience on reactions to those who threaten or bolster the cultural worldview. *Journal of Personality and Social Psychology*, *58*(2), 308–318.

Hall, E., Hall, A. V., Galinsky, A., & Phillips, K. W. in press. MOSAIC: A model of stereotyping through associated and intersectional categories. *Academy of Management Review*.

Harrison, D. A., & Klein, K. J. (2007). What's the difference? Diversity constructs as separation, variety, or disparity in organizations. *Academy of Management Review, 32*(4), 1199-1228.

Hauser, O. P., Linos, E., & Rogers, T. (2017). Innovation with field experiments: Studying organizational behaviors in actual organizations. *Research in Organizational Behavior, 37*, 185-198.

Hebl, M. R., King, E. B., Glick, P., Singletary, S. L., & Kazama, S. (2007). Hostile and benevolent reactions toward pregnant women: Complementary interpersonal punishments and rewards that maintain traditional roles. *Journal of Applied Psychology, 92*, 1499- 1511.

Hernandez, M., Avery, D. R., Volpone, S. D., & Kaiser, C. R. (2018). Bargaining while Black: The role of race in salary negotiations. *Journal of Applied Psychology*. https://doi.org/10.1037/apl0000363

Highhouse, S. (2009). Designing experiments that generalize. *Organizational Research Methods, 12*(3), 554-566.

Huang, L., Frideger, M., & Pearce, J. L. (2013). Political skill: Explaining the effects of nonnative accent on managerial hiring and entrepreneurial investment decisions. *Journal of Applied Psychology*, *98*(6), 1005–1017.

Kasof, J. (1993). Sex bias in the naming of stimulus persons. *Psychological Bulletin, 113*, 140- 163.

Kaufmann, M. C., Krings, F., & Sczesny, S. (2016). Looking too old? How an older age appearance reduces chances of being hired. *British Journal of Management*, *27*(4), 727–739.

King, E. B., & Ahmad, A. S. (2010). AN experimental field study of interpersonal discrimination toward Muslim job applicants. *Personnel Psychology, 63*(4), 881-906.

King, E. B., Hebl, M. R., Morgan, W., & Ahmad, A. S. (2013). Field experiments on sensitive organizational topics. *Organizational Research Methods, 16*(4), 501–521.

Klein, O., Doyen, S., Leys, C., Magalhães de Saldanha da Gama, P. A., Miller, S., Questienne, L., & Cleeremans, A. (2012). Low hopes, high expectations: Expectancy effects and the replicability of behavioral experiments. *Perspectives on Psychological Science*, *7*(6), 572-584.

King, E. B., Shapiro, J. R., Hebl, M. R., Singletary, S. L., & Turner, S. (2006). The stigma of obesity in customer service: A mechanism for remediation and bottom-line consequences of interpersonal discrimination. *Journal of Applied Psychology, 91*(3), 579-593.

Krings, F., Sczesny, S., & Kluge, A. (2011). Stereotypical Inferences as Mediators of Age Discrimination: The Role of Competence and Warmth. *British Journal of Management*, *22*(2), 187–201.

Latu, I. M., Mast, M. S., Lammers, J., & Bombari, D. (2013). Successful female leaders empower women's behavior in leadership tasks. *Journal of Experimental Social Psychology*, *49*(3), 444–448.

Lonati, S., Quiroga, B. F., Zehnder, C., & Antonakis, J. (2018). On doing relevant and rigorous experiments: Review and recommendations. *Journal of Operations Management*, *64*, 19-40.

Madera, J. M., & Hebl, M. R. (2012). Discrimination against facially stigmatized applicants in interviews: An eye-tracking and face-to-face investigation. *Journal of Applied Psychology*, *97*(2), 317–330.

Morgan, W., Singletary, S., Hebl, M.R., & King, E.B. (2013). A field experiment: Reducing discrimination toward pregnant job applicants. *Journal of Applied Psychology, 98*, 799-809.

Morgenroth, T., & Ryan, M. (2018). Gender trouble in social psychology: How can Butler's work inform experimental social psychologists' conceptualization of gender? *Frontiers in Psychology, 9*.

Pearsall, M. J., Ellis, A. P., & Evans, J. M. (2008). Unlocking the effects of gender faultlines on team creativity: Is activation the key?. *Journal of Applied Psychology*, *93*(1), 225-234.

Ployhart, R. E., & Vandenberg, R. J. (2010). Longitudinal research: The theory, design, and analysis of change. *Journal of Management*, *36*(1), 94-120.

Podsakoff, P. M., & Podsakoff, N. P. (in press). Experimental designs in management and leadership research: Strengths, limitations, and recommendations for improving publishability. *The Leadership Quarterly*.

Prior, K., & Sargent-Cox, K. A. (2014). Students' expectations of ageing: An evaluation of the impact of imagined intergenerational contact and the mediating role of ageing anxiety. *Journal of Experimental Social Psychology*, *55*, 99–104.

Puhl, R. M., & Peterson, J. L. (2014). The nature, consequences, and public health implications of obesity stigma. In P. Corrigan (Ed.), *The stigma of disease and disability: Understanding causes and overcoming injustices* (pp. 183-203). Washington, DC: American Psychological Association.

Rosette, A. S., Ponce de Leon, R., Koval, C. Z., & Harrison, D. A. (2019). Intersectionality: Connecting experiences of gender and race at work. *Research in Organizational Behavior, 18*.

Rubin, D. L. (1992). Nonlanguage factors affecting undergraduates' judgments of nonnative English-speaking teaching assistants. *Research in Higher Education*, *33*(4), 511–531.

Ruggs, E. N., Hebl, M. R., & Williams, A. (2015). Weight isn't selling: The insidious effects of weight stigmatization in retail settings. *Journal of Applied Psychology, 100*(5), 1483-1496.

Rule, N. O., Bjornsdottir, R. T., Tskhay, K. O., & Ambady, N. (2016). Subtle perceptions of male sexual orientation influence occupational opportunities. *Journal of Applied Psychology*, *101*(12), 1687–1704.

Rule, N. O., Tskhay, K. O., Freeman, J. B., & Ambady, N. (2014). On the interactive influence of facial appearance and explicit knowledge in social categorization. *European Journal of Social Psychology*, *44*(6), 529–535.

Scandura, T. A., & Williams, E. A. (2000). Research methodology in management: Current practices, trends, and implications for future research. *Academy of Management Journal, 43*, 1248-1264.

Shaughnessy, B. A., Mislin, A. A., & Hentschel, T. (2015). Should He chitchat? The benefits of small talk for male versus female negotiators. *Basic and Applied Social Psychology*, *37*(2), 105–117.

Simonsohn, U. (2015). How to study discrimination (or anything) with names; if you must. Data Colada. Retrieved from www.datacolada.org.

Spencer, S. J., Zanna, M. P., & Fong, G. T. (2006). 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.
University of Chicago Legal Forum, 139-167.

Stone, D. L., Hosoda, M., Lukaszewski, K. M., & Phillips, N. T. (2008). Methodological problems associated with research on unfair discrimination against racial minorities. *Human Resource Management Review, 18*, 243-258.

Uziel, L. (2010). Rethinking social desirability scales: From impression management to interpersonally oriented self-control. *Perspectives on Psychological Science*, *5*(3), 243-262.

Van Hoye, G., & Lievens, F. (2003). The effects of sexual orientation on hireability Ratings: An experimental study. *Journal of Business and Psychology*, *18*(1), 15–30.
Windscheid, L., Bowes-Sperry, L., Kidder, D. L., Cheung, H. K., Morner, M., & Lievens, F. (2016). Actions speak louder than words: Outsiders' perceptions of diversity mixed messages. *Journal of Applied Psychology, 101*(9), 1329-1341.

Table 1

Overview of key categories of diversity, example papers and manipulations

Diversity variable of interest	Types of manipulations	Examples of dependent variables	Example studies	Advantages	Disadvantages
Age	Young vs. old as reported age by birthdate (i.e., 29 years vs. 50 years)	Competence and warmth; interview intentions; selection decisions	Krings et al., 2011	Simple, yet effective	Differences could be subtle and leave out the non-working age older adults (i.e., 65 and older)
	Young vs. old as indicated by photo on resume		Kaufmann et al., 2015	Disentangles age from perceptions of being old vs. young	May confound with attractiveness
Body Weight	Costume (e.g., obesity prosthetic devices)	Customer service quality	King et al., 2006	Can be conducted as a field experiment	Very time consuming and cost intensive
	Labels (e.g., obese vs. normal weight)	Hiring decisions	Agerström & Rooth, 2011	Simple and effective manipulation	Encourages extreme examples; manipulated photos may affect other factors (e.g., attractiveness)
Gender	Names	Negotiation outcomes	Shaughnessy et al., 2015	Easy to make similar sounding male and female names	Potential cultural influences; participants must read carefully

	Images	Agency and communality	Brosi et al., 2016	Richer medium for manipulating gender	Need to pretest the materials; also need manipulation checks and control variables (e.g., attractiveness)
Nationality and Spoken Accent	Recordings (e.g., native vs. nonnative speakers of a particular language)	Political skill perceptions	Huang et al., 2013	Captures another element beyond demographic race	There is a wide range of accents; the degree of accent may also play a role, as well as who is rating it
	Recordings (e.g., French vs. Chinese vs. American accented English)	Consumer choice	Livingston et al., 2017	Considers the implications of different nationalities and their accents	May confound some racial stereotypes in the manipulation
Mental and/or Physical Ability	Mention vs. no mention of non-visible disability (i.e., Dyslexia)	Performance and expected salary	Colella et al., 1998	Simple and overt	Might require prior knowledge from the participants of the particular disability
	Images	Applicant ratings and memory recall of interview	Madera & Hebl, 2012	High external validity	Potentially costly for confederate and for realistic and consistent manipulation

Pregnancy	Prosthesis of a pregnant stomach	Interpersonal treatment at the workplace	Hebl et al., 2007	Provides a standardized and externally valid manipulation	Costly for both the prosthesis and confederate(s)
		Incompetence, lack of commitment, inflexibility, need for accommodation, provision of a job application form	Morgan et al., 2013		
	Implicit indicators (e.g., preparing a nursery)	Evaluations of future expected work behavior	Hebl et al., 2007	Subtle indicator; simple and easy application	Requires participants to read carefully to properly interpret stimulus materials
Race	Images	Salary	Hernandez et al., 2018	A richer medium with less ambiguity	May introduce confounds (i.e., attractiveness or prototypicality)
	Racial prototypicality	Stereotype activation	Ma et al., 2018	Highlights the degrees of difference within racial categories	Introduces complexity into study designs May require larger sample sizes
Religion	Attire (e.g., Muslim-identified vs. non-religious)	Reactions towards applicants	King & Ahmad, 2010	Strong manipulation of differences	Requires an understanding that clothing indicates religion

	Attire (e.g., head scarf vs. no head scarf)	Outcome expectancy	Everett et al., 2015		May need to the local laws governing wearing head scarves take into consideration
Sexual Orientation	Images (e.g., faces, which pre-tests indicated as being viewed as "gay" or "straight")	Occupational opportunities and hireability	Rule et al., 2016	Relies on perceived sexual orientation provides a subtler manipulation than labels	Could also confound other factors (e.g., attractiveness, attributions of sophistication or cultured background)
	Description of family situation (i.e., living with another man)	Hireability	Van Hoye & Lievens, 2003	More overt, but also subtler than labels	May trigger social desirability in participants
	Labels (e.g., "straight" or "gay")	Social categorization into group	Rule et al., 2014	Unambiguous	