Organizational Research
Methods

Field Experiments on Sensitive Organizational Topics
Eden B. King, Michelle R. Hebl, Whitney Botsford Morgan and Afra Saeed Ahmad

The online version of this article can be found at:

A more recent version of this article was published on - Aug 30, 2013

Additional services and information for *Organizational Research Methods* can be found at:

Email Alerts: http://orm.sagepub.com/cgi/alerts

Subscriptions: http://orm.sagepub.com/subscriptions

Reprints: http://www.sagepub.com/journalsReprints.nav

Permissions: http://www.sagepub.com/journalsPermissions.nav

Version of Record - Aug 30, 2013

>> OnlineFirst Version of Record - Oct 25, 2012

What is This?

Field Experiments on Sensitive Organizational Topics

Eden B. King[1], Michelle R. Hebl[2],
Whitney Botsford Morgan[3],
and Afra Saeed Ahmad[1]

Abstract

Organizational scholars study a number of sensitive topics that make employees and organizations vulnerable to unfavorable views. However, the typical ways in which researchers study these topics—via laboratory experiments and field surveys—can be laden with problems. In this article, the authors argue that the difficulties in studying sensitive topics can be overcome through the underutilized method of field experiments, detail strategies for conducting high-quality experimental field studies, and offer suggestions for overcoming potential challenges in data collection and publishing. As such, this article is designed to serve as a guide and stimulus for using the valuable methodological tool of field experiments.

Keywords
field experiments, sensitive topics, discrimination, bullying, unethical behavior

Bob Cialdini (2009), a social psychologist, master field researcher, and keynote speaker at the Society for Industrial Organizational Psychology's annual conference in 2010, announced his retirement and "break up" with psychological science. He cited his frustration with the field's narrow focus on cognitive processes and mediating mechanisms measured in laboratory studies and lamented that "truly natural human activities don't lend themselves to the collection of the kinds of secondary data on which to base meditational analyses" (p. 4). Cialdini joins a long line of critics of both laboratory experiments and survey methods for studying human behavior (e.g., Berkman & Lieberman, 2011; Beyer & Trice, 1984; Gordon, Slade, & Schmitt, 1986; Greenberg & Folger, 1988; Lewandowski & Strohmetz, 2009; Medin, 2011; Mitchell, 1985; Mook, 1983; Paluck, 2006; Sears, 1986) who have called for more rigorous, behaviorally based, and externally valid research. Similarly, researchers suggest the importance of going beyond the study of undergraduate subject pools and suggest

[1]Department of Psychology, George Mason University, Fairfax, VA, USA
[2]Department of Psychology, Rice University, Houston, TX, USA
[3]School of Business, University of Houston–Downtown, Houston, TX, USA

Corresponding Author:
Eden B. King, Department of Psychology, George Mason University, 4400 University Drive, Fairfax, VA 22030, USA
Email: eking6@gmu.edu

psychology must become "bolder in its choices of study populations if it is to be relevant to real-world problems" (Ginges, Atran, Sachdeva, & Medin, 2011).

At the intersection of these seemingly distinct and fallible approaches lies opportunity in the form of field experiments. The dilemma—or set of "interlocking choices" (McGrath, 1981)—imposed by the need to maximize precision of measurement, realism of context, and generalizability may be addressed by conducting experiments in the field. By manipulating behavioral variables in a field setting (Scandura & Williams, 2000), field experiments maintain both rigor and relevance (responding to Mitchell, 1985) and allow examination of research questions concerning both what *can* happen and what *does* happen (responding to Mook, 1983). Moreover, Locke (1986) suggests field studies provide opportunity to test the generalizability of laboratory studies, thus increasing confidence in scientific contributions. Indeed, though rarely implemented, field experiments have yielded important conclusions on organizational topics, including prosocial behavior (Holland, Aarts, & Langendam, 2006), job attitudes (Brief, Butcher, & Roberson, 1995; Griffin, 1983), and occupational health interventions (May, Reed, Schwoerer, & Potter, 2004).

A recent study on employment discrimination exemplifies this opportunity. Agerstrom and Rooth (2011) created and submitted job applications in response to 985 job openings in Sweden and tracked interview offers. These applications were identical with the exception of a photograph of the applicant that varied randomly with regard to gender and weight (but were pretested to match with regard to attractiveness). After interview decisions were made, the researchers recruited the targeted hiring managers to complete both explicit and implicit measures of stereotypes of obese individuals. The results demonstrated that implicit, but not explicit, obesity stereotypes were associated with discriminatory interview decisions about ostensible job applicants. This example demonstrates that field experiments may be particularly useful in studying workplace issues that are unfavorable, interpersonally discrediting, or otherwise taboo. The sensitivity around such topics is amplified in organizational settings wherein impression management concerns are salient as employees work to keep their jobs and leaders strive to appease their stakeholders. Typical laboratory studies are insufficient in creating the real-world pressures around these tabooed topics, and accurate reporting in traditional (correlational) field studies would be obscured (Mitchell, 1985). Field experiments can overcome these challenges and build important knowledge around sensitive organizational topics.

Thus, the goal of this article is to describe field experimentation as it can be applied to sensitive subject matter for individuals and their organizations. We will discuss practical suggestions for enacting field experiments and for overcoming challenges that arise in conducting and publishing this type of research. In doing so, we hope to galvanize high-quality field experiments and increase appreciation for the meaningful conclusions that can be derived from this approach.

What Are Field Experiments?

We define field experiments as a method that uses random assignment to implement a manipulation relevant to working adult participants engaging in genuine tasks or with genuine outcomes in natural settings. Such research, we argue, is rare and underutilized. For instance, in their review of prominent methodological approaches taken in three major management journals (*Academy of Management Journal*, *Administrative Science Quarterly*, and *Journal of Management*), Scandura and Williams (2000) indicated the stark number of studies that relied on field experiments. Specifically, 3.9% of studies published between 1985 and 1987 and 2.2% of studies published between 1995 and 1997 involved "collecting data in a field setting but manipulating behavioral variables" (p. 1260). Instead, the majority of scholars publishing in these journals relied on nonexperimental field (i.e., survey) research. As another example of the underrepresentation of field experiments, only 11% of nearly 1,000 studies on prejudice reduction were classified as experimental field research in a

recent review of the literature (Paluck & Green, 2009). Understanding that nonexperimental research is limited in establishing causal inference, it is important to consider whether field experiments can improve upon these limitations.

Judging this method by Cook and Campbell's (1976) criteria, Scandura and Williams (2000) described field experiments as moderately high with regard to precision of measurement and realism of context but low with regard to generalizability. This conclusion was based on the idea that sampling in field experiments is systematic rather than representative and that the setting is contrived. Indeed, the studies cited as examples of field experiments by Scandura and Williams are certainly subject to these criticisms. Uncited in this article are numerous field experiments that overcome many of these limitations.

For example, Brief et al. (1995) randomly assigned nurses who volunteered to participate in a study to a control or positive mood-inducing condition (giving them cookies and a small gift) and measured job satisfaction ratings. As another example, Buckley and colleagues (2002) randomly assigned genuine job applicants for telemarketing positions to one of four job preview conditions that were implemented by an organization member who was blind to the study's hypotheses to examine the effect of realistic job previews on acceptance of a job offer and number of days worked (turnover). Ganster, Mayes, Sime, and Tharp (1982) examined workers' biological and self-report indictors of stress as a function of random assignment to a stress management or control condition. Combs and Luthans (2007) compared pre- and posttreatment measures for diversity training and control group participants to explore the role of self-efficacy on diversity training outcomes.

In his influential studies of leadership, teams, and self-fulfilling prophecies, Eden (e.g., Dvir, Eden, Avolio, & Shamir, 2002; Eden, 1985, 1990) has often utilized field experiments. In a single article, Eden (1985) described a field experiment with three dependent variables of varying degrees of rigor to address the methodological limitations of previous work. Specifically, Eden examined 18 army units, a random half of which attended a team development program. Responses from a random set of the participants suggested positive evaluations of the program. Conversely, comparisons between treatment and control group evaluations of teamwork and pre- and postevaluations demonstrated that the intervention did not ultimately improve team functioning. In response to Scandura and Williams's (2000) critiques, these studies clearly demonstrate how the use of random assignment supports the generalizability of the findings.

Indeed, Carpenter, Harrison, and List (2005) specified that

> although it is tempting to view field experiments as simply less controlled variants of laboratory experiments, this would be a serious mischaracterization. What passes for "control" in laboratory experiments might in fact be precisely the opposite if it is artificial to the subject or context of the task. (p. 1)

These economists described five factors that typify field experiments: the nature of participants (students vs. others), relevance of information and experience that participants bring to the task, the commodity (genuine or artificial), task or institutional rules applied (domain specificity), and the environment in which the participant operates (realism). These factors underlie the overarching taxonomy of experiments Carpenter and colleagues ultimately defined:

> A *conventional lab experiment* is one that employs a standard subject pool of students, an abstract framing, and an imposed set of rules
>
> An *artefactual field experiment* is the same as a conventional lab experiment with a nonstandard subject pool

A *framed field experiment* is the same as an artefactual field experiment but with a field context in either the commodity, task, or information set that the subjects can use

A *natural field experiment* is the same as a framed field experiment but where the environment is one where the subjects naturally undertake these tasks and where subjects do not know they are in an experiment. (p. 6)

Thus, research involving randomly assigning participants from online survey vendors who use volunteer participant panels such as Mechanical Turk or Study Response to experimental conditions might be considered an artifactual or framed field experiment since participants are engaging in tasks that are not naturally undertaken and aware that they are in a study. Labor economists often attempt to achieve high-quality experimental field research in audit studies that involve confederates or trained actors attending job interviews, applying for jobs over the phone, or responding to job ads in written form (Riach & Rich, 2002). In our view, the particular labels derived from this taxonomy are less important than our general definition of high-quality organizational field experiments. Drawing on the research of Carpenter and colleagues (2005), we define high-quality organizational field experiments as: *studies involving random assignment to manipulated conditions of working adult participants* (the population of interest to organizational scholars) *who engage in genuine tasks or with genuine outcomes in natural settings*. These are the characteristics of what we consider to be high-quality experimental field studies that can be particularly useful in exploring sensitive organizational topics.

What Are Sensitive Topics?

When people gather for a special occasion with a mix of friends, family, and in-laws of varying degrees of familiarity, interactants generally know what topics to avoid discussing: religion, sex, money, and politics. People recognize that bringing up these topics might lead to discomfort, tension, and long-lasting disfavor. Socially shared knowledge of appropriate conversation extends beyond holiday meals within one's personal sphere to professional domains and organizational settings. That is, there are simply some issues that workers fear expressing, are embarrassed to admit, or suppress to the point of denial.

In accordance with Sieber and Stanley (1988), socially sensitive research involves "studies in which there are potential consequences or implications, either directly for the participants in the research or for the class of individuals represented by the research" (p. 49). More broadly, Lee (1993) defines sensitive research as "research which potentially poses a substantial threat to those who are or have been involved in it" (p. 4). Such sensitive organizational topics might include those focusing on unethical or scandalous behavior, bullying, harassment, discrimination, the use of unsafe procedures or errors, "coming out" at work, illicit drug use, and abusive relationships. It is also possible that there are context-specific norms about what is discussed in a particular organization or job; for example, it may be that workers in downtown New York implicitly agreed not to talk about planes that fly past tall buildings after September 11, 2001. Another example may be workers who ignore unsanitary working conditions in establishments that serve food and beverage. The common theme underlying these topics is that it would be socially undesirable (and sometimes legally prosecutable) for employees or for organizations as entities to be associated with these behaviors. Reputations, both individual and organizational, could be negatively impacted through linkages to incidents (James & Wooten, 2006). As a result, research participants may be unwilling or unable to report incidents of any one of these behaviors accurately. We argue that this is precisely why field experiments can be so useful.

Why Should We Use Field Experiments to Study Sensitive Topics?

Stone, Hosoda, Lukaszewski, and Phillips (2008) highlighted problems inherent in studying sensitive topics in their discussion of the methodological problems of research on racial discrimination, concluding that the obtrusive measures, student samples, and demand characteristics common to this area of research might yield larger effects than actually exist in the population at large. It follows that field experiments that overcome these problems could uncover important effects on sensitive organizational topics.

The reasons why field experiments might be so potent can be understood by considering each of field experiments' three important components: (a) the necessity of random assignment, (b) the use of genuine tasks, and (c) the reliance on natural settings. First, as Cook and Campbell (1979) discuss in detail, random assignment is a critical aspect of experimental design because it functions to "draw samples that are comparable to each other within known limits of sampling error" (p. 341). This is particularly critical when it comes to organizationally sensitive topics because equivalence across experimental conditions determines the validity of the results. For example, conclusions from a study of an intervention designed to curb unethical behavior by rewarding whistleblowing would be undermined if assignment to the treatment condition corresponded with operational divisions or teams; findings could be attributed to differences between teams or divisions (e.g., a particularly unethical supervisor) rather than the intervention itself.

Experimental manipulation with random assignment allows for causal inference. This is particularly important in studies—like those of sensitive organizational topics—wherein causality is questioned. For example, scholars and practitioners often question whether gender differences in pay are due to discrimination by decision makers or behaviors by men and women (see Blau & Kahn, 2000; King, 2008, Lyness & Thompson, 1997). An experimental approach to understanding gender differences in pay might involve randomly assigning participants to evaluate either the résumé of a man or a woman. Because the résumé is identical in every way except the gender of the target, differences in evaluations (and pay expectations) could only be due to gender. Thus, experimental manipulations in conditions of random assignment can nullify such questions by providing a single causal explanation for effects.

Random assignment to experimental conditions produces particularly compelling results to the extent that the general conditions of the experiment are ecologically valid or possess "mundane realism" (Stone et al., 2008). The strength of the study depends on whom the participants are and to whom the sample might be expected to generalize. One characteristic that might affect the validity of research on sensitive organizational topics is participant age. Attempts to understand incidences of sexual harassment, for example, may look very different among young, male undergraduates who are embedded in a college culture of fraternity parties and hazing than among older, working men embedded in an organizational culture with formalized reporting procedures and salient legal considerations.

Second and similarly, the degree to which the tasks or behaviors in which participants engage are genuine ("real") also influences the validity of research designs. One example of this is the difference in performance appraisals that are done for the purposes of "research" versus those that are done for genuine organizational purposes (i.e., "the performance appraisal purpose effect"; Jawahar & Williams, 1997). When participants engage in activities they know to be false, they may indeed exhibit false (or at least unrepresentative) patterns of responding.

Third and finally, of critical relevance to examining sensitive topics is the degree to which the setting of research is natural. Participants are generally aware that they are participants in carefully designed laboratory experiments or survey studies. Thus, they may be particularly susceptible to socially desirable responding patterns. In the case of sensitive topics such as discrimination, participants can often monitor their overt behaviors or responses to reflect socially acceptable standards

of egalitarianism. In high-quality field experiments, participants are in natural settings outside the constraints of the lab or a survey, may be unaware that they are involved in a study of any kind, and are blind to the intent of the study. Thus, actions may be less filtered by social desirability and behaviors that emerge are likely more representative of naturally occurring incidents.

Another overarching issue about sensitive organizational topics that demands high-quality field experiments is that discovery of unfavorable information is likely subject to a high level of scrutiny. People are less concerned about causal explanations when they find $20 in their pocket that they forgot about compared to when they lose $20. Similarly, organizations and employees may raise more questions about information that potentially harms their reputation than about information that boosts their reputation. Because field experiments address many of the criticisms typical to laboratory experiments and field surveys, it gives researchers strong ground on which to base and defend their conclusions. Taken together, these points underscore the relevance of field experiments for studying sensitive organizational topics. The question, then, becomes how to do it.

How to Do Field Experiments

One explanation for the lack of field experiments in top outlets for management research is that scholars do not know how or feel able to conduct such research (see Borsboom, Kievit, Cervone, & Hood, 2009; Highhouse, 2009). Whereas survey design and laboratory methods are commonly taught in graduate courses on research methods or psychometrics, students in organizational sciences are not typically trained in practical strategies for enacting field experiments. The SIOP Guidelines for doctoral training in industrial and organizational psychology (Society for Industrial and Organizational Psychology, Inc., 1999) state that students should learn experimental, quasi-experimental, and nonexperimental design but do not specify that these research design features should be considered in the field. The updated version of Cook and Campbell's (1979) book on experimentation in field settings no longer has "field settings" in the title (Shadish, Cook, & Campbell, 2002). Recent research acknowledges the neglect of quasi-experimental research and advocates for renewed interest (Grant & Wall, 2009). It follows that a primary goal of the current article is to provide practical guidance for conducting experimental field studies in both ideal circumstances and in situations that are less than ideal.

Go for Gold

To maximize the benefits of field experiments, the attributes that characterize high-quality research must be fully addressed and should be sought in every field experiment. This means the use of random assignment and "real" participants in "real" situations. However, there are also some additional preliminary things to consider and multiple sequential steps to engage in before pursuing field experiments.

Determine fit of research question to field experiment methodology. Field experimentation begins with a research question that is difficult to address in a lab setting or with survey methodology (see McGrath, 1981). For example, perhaps the question requires behavioral measurement and simply cannot be garnered via self-report and/or survey methodology. Perhaps the laboratory does not offer the realism requisite to embrace the validity and generalizability of the answer to the research question. Or perhaps participants realize they are being studied and feel constrained to act in socially desirable ways, thereby leading researchers to be unsure of how valid the results are to a posed question. We can think of a number of such questions: How can organizations encourage employees to recycle (Holland et al., 2006)? Are Muslim women discriminated against in selection decisions (King & Ahmad, 2010)? How do people respond to over- and underpayment inequity (Greenberg, 1982)? How can pregnant women overcome sexism in the hiring procedures (Botsford Morgan,

Singletary Walker, Hebl, & King, 2012)? How can people be happier at work (Kaplan et al., 2012)? What are the individual and situational determinants of employee theft (Greenberg, 2002)?

Our advocacy of field experiments should not be interpreted to suggest that it is appropriate for addressing every question. Instead, we acknowledge that some topics are better suited to this kind of field experiments than other topics. Cook and Campbell (1979) suggested that some specific conditions are conducive for randomized experiments in the field. For example, field experiments may be feasible when lotteries are expected (as in the draft or dorm assignments), when people are unlikely to communicate about their experiences (e.g., spatial separation), or when researchers can create their own organizations in which to exert total control (Cook & Campbell, 1979). We do not dispute these conditions, but recognize substantial practical and ethical challenges of creating an organization through which data are collected and are pleased to report that high-quality field experiments can be conducted without having to create a whole company. Drawing upon Locke's (1986) approach, we argue that it is critical to identify the nonessential and essential features of a research paradigm before embarking on a particular study. Specifically, Locke argued that it is not plausible or necessary to replicate every feature of a particular setting to which results should generalize, but rather to focus on the essential features that must be replicated in order to demonstrate validity with respect to a particular phenomenon.

Lay the groundwork. The groundwork that is necessary for a study will be based on the particular research question being posed. If conducted within a single organization, researchers typically prepare for their study by gaining consent, obtaining psychological and financial support, recruiting manpower, and/or gaining other forms of buy-in and knowhow from various organizational constituents. There are a number of general considerations that all researchers should consider (e.g., institutional review board, deliverables to the organization, cost burdens, authorship), but experimental field researchers need to focus particular attention on a few issues that are central to studying sensitive topics such as: (a) specific requirements of employees (including whether or not informed consent is planned), (b) anonymity and/or confidentiality of individuals and the organization, and (c) publishing agreements and accessibility of data and publishing.

Field experiments also are conducted in settings that are open to the public such as sidewalks, parks, shopping malls, auto dealerships, hospital parking garages, and amusement parks (Cialdini, 2009). For field experiments conducted in these public settings, laying the groundwork typically involves gathering information about the location of interest. For example, in the research conducted on hiring discrimination in retail stores (e.g., Hebl, Foster, Mannix, & Dovidio, 2002; King & Ahmad, 2010; King, Mendoza, Madera, Hebl, & Knight, 2006), the researchers began by obtaining lists of all the shopping malls in a major metropolitan area, lists of each store within each mall, and maps of each mall and parking area. Then, the researchers updated such lists according to stores that had gone out of business and/or brand new stores that had opened, called each store and asked if they were currently hiring salespeople, and created a physical plan of how to best get to and from each store locale. Finally, the researchers had to consult legal professionals to ensure their research plans were within the legal constraints of the location of interest. This step was imperative since the research design involved a confederate carrying a hidden audio recorder. The researchers learned that it is legal in some states (but not others) to take an audio recording of a conversation so long as one party is aware of the recording. Clearly, the relevant institutional review boards must also sign off on any given experiment, so it is useful for researchers to discuss their plans with their respective review board officials before submitting the final, formal application for review. This step can help researchers proactively address potential concerns of the review board before plans are finalized.

Gather a team. Presuming the researchers have formulated a research question that is well suited for field experimentation and established the groundwork for beginning the study, a next important step is to gather the large number of committed, conscientious individuals who are often required to carry out field research. Substantial personpower is often required for doing high-quality field experiments, particularly when experiment facilitators (those instructing the data collectors), data collectors (these individuals may include experimenters, confederates, and/or observers), and independent coders (those who are unaware of the study methodology and who simply listen and code audiotapes after data collection is complete) are all needed to conduct just a single study. Add to this the fact that effect field experimentation often requires pilot testing of materials; training of experimenters, confederates, observers, and coders; and the need for coding and entering of data (as SurveyMonkey and other online data collection sites are not typically used), and the critical need for gathering effective team members becomes clear. In a study conducted by Hebl, King, Glick, Singletary, and Kazama (2007), for instance, the researchers relied on a team comprised of three professors, four graduate students, and 20 undergraduate research assistants.

With this often large number of research personnel, it is imperative for the team leader to have organizational skills and management abilities and be able to engender enthusiasm for each research participant's task as some are more tedious to enact (e.g., coding data trials) than are others (e.g., serving as a confederate with a prosthesis). It may also be critical to filter information carefully and differentially (according to study roles such as experimenters, confederates, and/or observers) to ensure that the needed research personnel remain naïve or blind to relevant elements of the study.

Training sessions. In addition to engendering enthusiasm for field experiments, it is also essential to engender professionalism in, preparedness for, and standardization for testing in research assistants who will be collecting data in the field. Situations may arise in field experiments that do not ever or very uncommonly arise in the lab (e.g., third parties insert themselves into ongoing interactions that are being studied or participants do not follow the expected patterns of behavior on which experimental scripts were based). For this reason, it may be necessary to hold training sessions in an attempt to maximize the utility of field experiments and to prepare and protect the field experiment personnel. These sessions should focus on teaching research assistants how to enact targeted behaviors in a standardized way across trials and preparing research assistants in how to deal with unexpected sorts of behaviors that are more likely to happen with field experiments than other types of research.

To maximize success with field experimentation, then, researchers might plan to meet several times (prior to collecting any data at all) with the research group to review the goals (as much as can be allowed by not simultaneously forfeiting the integrity of keeping personnel blind to study hypotheses), review the logistics associated with the planned data collection, and practice scripts and standardized behaviors several times until a level of comfort is reached. As an example, in a study examining discrimination toward pregnant job applicants (Hebl et al., 2007), confederates memorized specific scripts and behavioral cues to use in their interactions. The experiment facilitators trained confederates not only to use consistent words and phrases across conditions, but also to move and walk in such a way that their pregnancy prosthesis would appear realistic. Then, the researchers played the role of store personnel and began asking several types of questions that employers might ask to prepare the confederates for what they might encounter.

When conducting training, researchers should stress to field experiment personnel the importance of professionalism, of maintaining the utmost respect for the participants who are often naïve, and of considering and practicing a large variety of unexpected behavioral situations that might possibly arise within any given script. While these goals are generally characteristic of most research, it is particularly paramount in field experimentation to instill and promote the use of good judgment and development of action plans as sometimes unintended situations arise (e.g., a security officer or store manager approaches a confederate and is suspicious about prostheses).

The other requisite skill that training should provide is reinforcement of displaying consistent behaviors within and across conditions to ensure standardized behavior. This is particularly true for studies that utilize experimenters and confederates—the power of field experiments is the results that emerge from reactions to a consistent manipulation among a sea of other standardized behaviors. So, researchers should carefully train experimenters and confederates, keeping them naïve to the study's purpose whenever possible but *always* ensuring that their behaviors remain stable so as to avoid systematic error such as experimenter bias or the self-fulfilling prophecy.

Choosing appropriate stimuli. In his article on experimental methods, Highhouse (2009) described the importance of choosing or designing appropriate stimuli that are strong with regard to both construct validity (that the manipulation is interpreted as intended and does not have unintended effects) and external validity (that the manipulation is representative of the construct of interest). With regard to construct validity, it is particularly important that experimental manipulations in field experiments are realistic and ordinary so as not to raise participants' suspicion. Recall, for example, the Agerstrom and Rooth (2011) study on discrimination toward obese job applicants described in the introduction of the article. The primary manipulation involved photographs of individuals who varied with regard to weight. In the United States, it is uncommon for photos to be shared with potential employers and would have appeared quite strange for job applications to be accompanied by photographs. Agerstrom and Rooth tell readers, however, that photographs are expected to accompany applications in the context in which the study took place (Sweden). Realism is a particular concern with regard to field experiments on sensitive topics since any suspicion is likely to contaminate responding; if participants know that they are being studied, they are likely to engage in or report socially desirable behaviors. It is important to be aware that this need for realism sometimes necessitates a relatively weak manipulation and thus often yields a small effect size that should be interpreted in light of the manipulation strength (Highhouse, 2009).

With regard to external validity, Highhouse (2009) specified that an often ignored but critical aspect of experimental manipulation is the extent to which a particular stimulus (e.g., a résumé) is reflective of a larger class of stimuli to which the findings are argued to generalize. Thus, it can be helpful to identify a typical stimulus within a population of stimuli representing a particular construct. If a researcher were interested in studying reactions to unethical behavior, for example, it would be appropriate to randomly select several typical behaviors from a large set of unethical behaviors. As another example, Bertrand and Mullainathan's (2004) field experiment on racial discrimination in interview decisions required unique stimuli in the form of names that were representative yet distinctively male and female from Caucasian and African American racial categories. They gathered frequency data from birth certificates regarding the most common names and conducted a survey in the relevant geographic regions to determine perceived distinctiveness of these names. Moreover, consistent with Highhouse's suggestion, the experimenters used multiple exemplars (names) of each level of the manipulated independent variable (gender and race) across participants to further strengthen external validity.

In field experiments on sensitive organizational topics, the construction of representative stimuli can be particularly challenging given that less control may be available—sometimes researchers have to work with whichever trainer, manager, confederate, or information that happens to be available. In such cases, it is important to keep as much constant across conditions as possible, eliminate unnecessary information, track participants' reactions to the stimuli in order to determine their likely representativeness, and use multiple stimuli to represent the larger class of stimuli when possible.

Participant recruitment, selection, and assignment. Shadish (2002) specified that the logistics of selection of participants and the assignment of individuals to conditions could make or break the study. He offered a number of suggestions for overcoming potential problems at this stage of field

experimentation. For example, the location of a representative sample of potential participants could be enhanced through preexperimental surveys and aggressive recruitment. In addition, the selection and retention of participants might be improved by conducting pilot tests to determine who will agree to participate, pipeline studies to ensure that participants will remain engaged over time, and examination of intervention or stimuli features that might cause withdrawal from the study. Of utmost importance is that random assignment is implemented appropriately (see also Shadish et al., 2002). Issues can arise unpredictably in the field—particularly when it comes to sensitive issues that can interfere with random assignment. An employee assigned to one condition might be transferred to a new position or division, a confederate might drop out halfway through an experiment, or history can intervene—clear procedures for ensuring the implementation of random assignment that include back-up plans are necessary for conducting high-quality field experiments.

Pilot testing. Since the investment of time and effort in field experiments is typically substantial, pilot testing of materials is critical. Of particular importance is conducting checks of the intended manipulation (see Highhouse, 2009). In laboratory experiments, manipulation checks can be embedded in experimental materials that are presented to participants. In field experiments, however, such direct assessment of manipulations cannot always be achieved. For example, Bertrand and Mullainathan (2004) could not reasonably contact each of the applicant recruiters who placed ads in newspapers and ask if they believed the names "Lakisha" and "Jamal" to belong to African American job applicants. Instead, presenting study materials (e.g., the names "Lakisha" and "Jamal") to naïve raters prior to running the actual study might be necessary to determine whether the manipulations effectively convey the intended constructs.

Pilot testing can also be useful for gauging general reactions to manipulations such as realism or believability. In pretest data collection, for instance, Williams (2010) learned that her male obesity prosthetics were not believable when a security officer and two store managers asked male confederates point blank what they were doing in "their outfits." Such pretesting allowed Williams to alter the male obesity prostheses significantly to be more believable, a condition that was strengthened in additional pretesting. Finally, pilot testing may also be useful in determining comparability of confederates. If it is necessary to compare two different people (e.g., a male confederate and a female confederate), pretesting can indicate whether they are perceived to be similar with regard to potential confounds (e.g., age, attractiveness, and ethnicity). In a review of field experiments (audit studies) in the labor literature, Riach and Rich (2002) concluded that the first way to enhance the validity of field experiments is to test for homogeneity across aggregated stimuli. Thus, researchers, when relying on confederates, should use more than one confederate to represent a category (e.g., male and female, or Black and White) because multiple exemplars help to reduce or eliminate the possibility that results might be due to some idiosyncratic feature about an individual.

Data collection. Field experiments typically require active involvement of researchers in the data collection process. Whether tracking interview callbacks (Bertrand & Mullainathan, 2004; Hebl et al., 2002), observing and noting interpersonal behaviors (King & Ahmad, 2010; King et al., 2006), or negotiating payments (Cialdini, Bickman, & Cacioppo, 1979), experimenters or confederates are often right in the middle of field experiments. As mentioned previously, efforts should be taken when possible to keep confederates and other experimenters blind to experimental manipulations and study hypotheses to avoid experimenter bias. For example, in a study of discrimination toward gay and lesbian job applicants (Hebl et al., 2002), confederates remained unaware whether they appeared to be gay or were assumed to be heterosexual. To accomplish this, confederates wore (but could not see) hats that identified them as gay ("Gay and Proud") or presumably heterosexual ("Texan and Proud"). In addition, observations of the confederates or experimenters themselves also might be undertaken to assess consistency (and lack of bias) of behavior. Another research

assistant observed each of the interactions the confederate job applicant enacted and assessed the believability and consistency of that confederate's behavior. Shadish (2002) also highlighted the importance of ensuring that the intervention or manipulation is *actually* integrated. Since intention does not necessitate action, it is not sufficient to ask participants or experimenters to engage in a particular behavior—the behavior itself must be observed.

Whenever possible, collection of observable behavioral indicators is ideal (see Lewandowski & Strohmetz, 2009). An example of this is the "lost letter technique" that involves observing whether or not members of the public mail stamped and addressed envelopes that they inadvertently find (Bolton, 1974). A more complex, and potentially more potent, design would involve measuring multiple behaviors. Such behaviors can be assessed through direct observation or through video or audio recording. The latter forms of behavior are generally preferable because they allow for in-depth analysis by coders who are blind to the condition and experimental hypotheses.

Reporting the study. When it comes to describing the procedures used in field experiments, it is key for researchers to focus on clarity and comprehensiveness. Whereas survey research can be described simply by stating that "participants indicated their agreement with X items using a Y-point Likert scale," field experiments are more complex, particularly to reviewers who have not personally done such research. Indeed, it has been argued that reviewers in the organizational sciences have little experience conducting experiments, much less field experiments (Highhouse, 2009). It is critical that readers (and reviewers) fully understand the actors, manipulations, measures, and setting. Because field experiments can be such a valuable methodological tool, authors have the opportunity to showcase their work by carefully describing each decision point and providing compelling data (e.g., the manipulation checks and observer ratings previously described) that justify each decision. The method section should make it possible for readers to not only replicate the study but also to understand why the study was conducted in one (and not another) particular way. Such detailed reporting may result in longer methods sections, but until field experiments are more common, it is often essential for explanatory purposes.

In addition, scholars reporting high-quality experimental field studies should also highlight the strengths of their approach. Authors become accustomed to noting the limitations of our methodological choices in the discussion section of manuscripts, but they also should weigh these against its strengths. Remind readers that in spite of any limitations, the study makes it possible to draw causal inferences regarding real people in real situations. These strengths certainly apply when field experiments are conducted in an ideal setting but can also apply to less ideal situations.

Settle for Silver

Sometimes circumstances are not conducive for achieving the ideals described in the previous section. For example, a company or institutional review board may insist that individuals obtain informed consent, which may cue socially desirable responding and thus contaminate the findings. As additional examples, researchers may gain access only to a small number of participants or be able to gauge only a few dependent variables. This does not mean that field experiments should be abandoned. Instead, these are instances when researchers should be opportunistic; in the interest of advancing science and practice, researchers should take advantage of the more limited opportunity presented.

Such opportunities often involve naturally occurring events. For example, Greenberg (1988) took advantage of the seemingly mundane refurbishment of office space to investigate an important question about the effects of over- and underpayment inequity on performance. He was able to (temporarily) randomly assign workers to offices that were "better," "worse," or "equal" to their typical office. Although Greenberg noted limitations such as the inability to track the effects of inequity

over a long period of time or to determine which specific aspects of new offices created status perceptions, Greenberg's opportunistic view—seeing office refurbishment as an opportunity for scholarship—yielded important information about inequity.

Another example of opportunism is evident in a recent study that examined framing effects on diversity training effectiveness. Upon learning that all of the incoming students at a university (and their mentors) were required to attend a diversity training session, researchers negotiated a brief manipulation in which students (and mentors) received a piece of paper that described a goal-setting or neutral introduction to the diversity training workshop (Madera, King, & Hebl, 2012). In the months that followed, students completed questionnaires about their diversity-related behaviors. Thus, the goals of random assignment and naturalistic tasks were achieved that allowed useful information about the role of mentorship and goal setting in diversity training even though students (rather than working adults) were sampled, the manipulation was exceptionally brief (less than 10 minutes), and participants were aware that they were in a study but blind to its purpose.

These kinds of studies take advantage of the benefits of field experiments under realistic constraints that often accompany field studies. While such field experiments may not fully embody the ideals of field experimentation, they are nonetheless useful in pushing organizational research forward.

Common Challenges to Field Experiments and Strategies to Overcome Them

Cook and Campbell (1979) detailed several methodological challenges in conducting field experiments, such as treatment-related attrition, small sample sizes, and treatment-related refusals to participate in the research. There is one factor that seems to be very characteristic of field experiments from a practical perspective: They are sometimes messy. A range of challenges emerges when conducting field experiments that concern ethical, legal, measurement, interpersonal, and publishing issues. These challenges are likely substantial barriers to conducting field experiments that may partially explain the relative paucity of this type of research.

Ethical and institutional review board considerations. Experimental field studies typically involve changing people's behavior or environment, collecting data on individuals, or giving them false information, often without their knowledge and sometimes without their consent (Grant & Wall, 2009). These issues, which can apply to participants or to experimenters themselves, raise moral red flags and institutional review board (IRB) headaches.

In terms of participants, plans to alter behavior or experience must be justified with regard to the potential risk:benefit ratio. You will need to convince your research team and then the IRB that the benefits of your manipulation(s) far outweigh the risks that might be imposed to participants. Does the manipulation (or involvement in the study more generally) pose greater psychological or physical risk to participants than it would benefit science or society? Is the experience something that might adversely affect participants or would be outside the realm of what they might normally encounter in their organizations and daily life? The *Chronicle of Higher Education* (Basken, 2010) reported on an experimental field study (Milkman, Akinola, & Chugh, 2012) in which professors received a meeting request from a (fictitious) prospective student who did not ultimately show up for a scheduled meeting. At least one angry professor filed a complaint that the experience of being a participant in the study (without their informed consent) posed unreasonable constraints on their time and energy. The *New York Times* reported on a field experiment in which restaurant owners received a (fictitious) letter detailing a complaint of food poisoning (Kifner, 2001). Business professor Francis Flynn was sued by a number of these "participants" who claimed the research was an egregious overstepping that caused psychological and financial harm. While most researchers believe their research has substantial benefits or else they would not be doing it, "risk" is clearly

a subjective term that should be considered conservatively in the interest of protecting the rights and health of participants.

One element common to some field experiments studies that particularly deserves ethical consideration is the use of deception. As Kelman (1967) stated, it is important to consider the negative implications of deception and use it only when clearly justified. Kelman describes deception as potentially being harmful because it can make participants vulnerable, it can ruin the reputation of the experimenter and or the experimenter/participant relationship, and it has potentially bad consequences for the field if things go awry. All of these valid points imply that a cost/benefit analysis must be conducted when using deception as a part of field experiments.

More generally, there are four primary questions that can help guide risk evaluations in field experiments. Written answers to these questions might be provided as part of the formal IRB application to help IRB members understand the risks and benefits involved in a particular study. First, researchers should consider whether informing participants about their involvement in a study may affect the responses they provide. In the case of sensitive organizational topics, the answer is almost always "yes"—sensitive topics yield socially desirable responding. A description of research demonstrating social desirability biases in the topic area may be sufficient to convince IRBs that informed consent would prohibit a test of the research question.

Particularly when informed consent is not provided, researchers must ask a second question: Could answers to the research question be garnered through a milder or less extreme method? In the case of the food poisoning study, for example, could the research question be answered if the experimenter simply asked restaurateurs to describe their likely reactions to a case of food poisoning? If a milder test would be equally effective, it should be seriously considered.

Third, researchers should question whether they are requiring participants to (unknowingly) engage in any behaviors that are outside of their typical daily activities. According to our definition of high-quality field experiments, validity is greatest when the activities of participants directly mirror those that are typical. Take for example a study on the interpersonal experiences of shoppers when they appeared to be obese (wearing an obesity prosthesis) compared to when they appeared to be thin (not wearing the prosthesis). The confederate-shoppers interacted with sales personnel for, at most, a few minutes in the same way that sales people interact with shoppers every minute of their typical work days. IRBs would likely weigh the risks in such cases as minimal.

Fourth and finally, researchers should evaluate whether informing participants that they were in a study at the conclusion of the research would likely yield benefit or harm. In a substantial proportion of studies of sensitive organizational topics, informing participants that their unfavorable behaviors were monitored might create more distress than it would relieve. Should debriefing be necessary, particularly when informed consent was not given prior to data collection, it may be helpful to empower participants to make personal decisions about whether or not their data can be included in the study. This protects the rights of human participants and thus is aligned with the mission of IRBs.

Although IRB purview tends to end with the participants, field experiments may call for special ethical considerations for experimenters—there is a high degree of involvement of researchers in field experiments. With this high degree of involvement in studying sensitive topics sometimes comes vulnerability. For example, confederates or experimenters may be subject to the harassment, bullying, or discrimination of focus in the study. As a result, these individuals should be fully informed of these potential risks prior to participating in the study and debriefed and supported after any negative interaction. Indeed, some researchers give formal informed consent statements to students involved as confederates so they fully consider the seriousness and potential consequences of the experiments in which they will be participating. And researchers might additionally stress to experimenters that they, just like participants, can withdraw from taking part in the study at any

time without any penalty. Locations for research should be chosen with physical safety in mind—researchers might choose public settings in which security or police officers are available and make sure that researchers always work in pairs or triads to enhance physical and psychological safety. Finally, senior investigators should always be immediately available on site and by phone to discuss any problems that are encountered.

Legal concerns. The sensitive issues of focus here, such as unethical behavior, harassment, and discrimination, are often subject to legal consideration. Consider that it is very possible that research findings may constitute discoverable evidence (i.e., evidence that can be presented in a court of law). When research involves employment decisions that are potentially litigable, assessments that are potentially discoverable, or other actions about which there are local and federal laws (e.g., audio and video taping), researchers should talk to a university lawyer before collecting data via field experimentation (or any methodology).

Measurement issues. To preserve realism, measurement of key variables is often constrained. For example, in Correll, Benard, and Paik's (2007) audit study, the authors tracked callbacks to fictitious job applications by "Allison" or "Brad" who either appeared to be a parent (i.e., member of a "Parent Teacher Association") or not (i.e., member of an "Alumni Association"). They found that women without children received job callbacks at twice the rate as mothers with identical credentials. Though these results are both important and compelling, the only dependent variable included was whether or not fake job applicants received calls in response to their applications, which can constrain/limit the findings.

Similarly, in a field experiment examining the impact of condoning or condemning discrimination on the behaviors of others, Zitek and Hebl (2007) had confederates voice opinions and then participants were asked to give theirs. Researchers sought answers from participants on only seven attitudinal items and no accompanying individual difference items were measured. These constrained dependent measures raise questions about the reliability of measurement and about whether the variables of interest might impact other outcomes and behaviors.

It is possible to contend with and/or overcome some of these constraints. For instance, in the Zitek and Hebl (2007) study, the researchers emailed original participants several weeks following their initial participation. Experimenters stated that due to an error, they needed to recollect the data. This strategy allowed for reliability of measurement by reassessing responses to the original items and the collection of responses to three additional dependent measures, thereby ensuring that participants were not just simply reproducing their original numbers.

Another measurement issue involves the use of confederates as data gatherers who complete questionnaires after participating in trials. There are times when they are not always blind to the conditions in which they participate and may intuit the study's purpose or speculate on the study's true intent. For instance, when participants are wearing obesity prosthetics or are posing as disabled individuals, it is likely that they anticipate compromised interactions.

Our general approach to some of these limitations in field experiments has been to strive for multiple behavioral measures from multiple perspectives whenever possible. Researchers might focus on interpersonal interactions, wherein innumerable verbal and nonverbal cues exist. Researchers might also analyze not only confederate and observer ratings of these interactions but also have separate coders listen to and rate video or audio recordings. Furthermore, researchers can assess the behaviors not only of the targets of interest (those with whom the confederates and experimenters are interacting) but also assess the consistency and standardization in behavior of the confederates and experimenters themselves. Convergence in findings across these measures provides additional confidence in any conclusions drawn.

Research explosions. The complexity of field experiments is particularly obvious in the case of what can only be labeled research explosions. These are incendiary incidents that occur between stakeholders in the experiment, such as the experimenter and participants or organizational decision makers and employees, that might make researchers yearn for the safety and predictability of the laboratory. In the real world, the unpredictable often happens. A hiring manager may touch a pregnant belly and realize it is a prosthetic. A CEO may be indicted on ethics charges while a research team is doing a whistleblowing intervention. Layoffs may hit the fan while researchers are manipulating interactional justice.

Our best advice here is to expect the unexpected and be prepared by having a general sort of action plan. As such, research teams might spend time thinking through various contingencies (ranging from mild to serious) and decide on a course of action that they will follow ahead of time if disaster erupts. The importance of having such initial and continued dialogue is critical, as are open lines of communication between the field experiment researcher and the entire research team, the institutional review board, and any partnering organizations. In the case of some explosions, such as the discovery of egregious, ongoing, and prosecutable sexual harassment in an organization, it may be necessary to consult an attorney. Researchers simply must be professional and constantly weigh the safety of those involved. Additionally, the researcher must also be flexible, creative, and opportunistic in responding to these explosions as the findings that accompany such challenges hopefully result in important research breakthroughs.

Publishing problems and tradeoffs. While the benefits of field experiments are many, the perception of these strengths may not be universally shared. Instead, when researchers submit their field experiments for scholarly review, they may receive criticism that is usually reserved for survey studies *and* criticism that is typical for laboratory studies. Salovey and Williams-Piehota (2004) described three primary objections to field experiments and offered counterarguments. The first objection is that scholars often believe that it is easier to draw causal inferences from laboratory experimentation. Salovey and Williams-Piehota countered that random assignment is the keystone of high-quality laboratory *and* field research that determines attributions of causality. The second objection is that it can be difficult to study mechanisms or explanations for phenomena in the field. This is related to beliefs that field experiments are applied in nature and make little theoretical contribution—Salovey and Williams-Piehota countered this argument by offering examples of work in social psychology that has applied value but is based in sound theory (i.e., prospect theory and framing of health communication). Indeed, field experiments can involve manipulations that explicitly test theoretical propositions—for example, comparing different strategies of enhancing well-being helps to clarify the factors that give rise to well-being (Kaplan et al., 2012). Finally, critics suggest that findings from one field setting (e.g., the North) might not generalize to another field setting (e.g., the South). This pot-calling-the-kettle-black objection is reminiscent of McGrath's (1981) claim that "methodological discussions should not waste time arguing about which is the right strategy, or the best one: they are all lousy!"

Indeed, reviewers have asked us about potential differences in how Northerners might respond to pregnancy manipulations compared to Southerners (from Hebl et al., 2007) and about whether there is bias toward people who wear hats to apply for jobs (from Hebl et al., 2002). The former criticism may have some theoretical relevance given small geographical differences in attitudes toward women but ignores the important data presented in the article, dismisses the practical challenges of replicating the study across the country, and most importantly, would likely apply to the vast majority of laboratory and field studies that are not nationally representative but are nevertheless published in organizational journals every day. The second criticism implies a lack of understanding of the experimental methodology—because confederates in the Hebl et al. (2002) study wore hats in

every condition, the effect of hat-wearing was methodologically controlled. In our opinion, these kinds of criticisms reflect a lack of knowledge of and respect for the power of field experiments.

The most common criticisms we as field experimenters have received in our submissions include: (a) "but you didn't control for . . . ," (b) "what about individual differences?" and (c) "did you examine other moderators of the effect?" Some of these concerns are genuine limitations of field experiments—researchers cannot usually measure personality traits of participants who are unaware that they are in a study (cf. Agerstrom & Rooth, 2011), researchers cannot control for how much experience participants have with a particular task, and researchers are often limited in the number of categories that they can successfully manipulate in one study in the field. But part of the criticism may also be that reviewers lack specific training in field experimentation. Thus, doing field experiments may be particularly difficult because other reviewers and readers may not fully understand or appreciate the complexities of field experimentation. In fact, recent research has shown that reviewers in organizational sciences tend to have limited understanding of experimental design more generally (see Borsboom et al., 2009; Highhouse, 2009). As field experiments gain more popularity in use, it is likely there also will be more refinement in practice and knowledge and respect in its use.

As there may be tension between the expectations of reviewers and the realities of field experiments, it is important to increase reviewers' awareness of some of the tradeoffs that accompany field experiments. One such tradeoff might involve deciding whether access to an externally valid sample that is rich in detail offsets a small sample size (e.g., Eden, 1995). Indeed, field experiment sample sizes are sometimes smaller than those achieved with laboratory studies or organizational surveys since the procurement of participants in field experiments can require a substantial amount of time and energy (e.g., use of confederates, logistics of running sessions, requirement of second source data). Similarly, sometimes there simply are not large samples available to study. If, for instance, a researcher gets access to assessing the impact of different types of diversity training initiatives within an organization, the researcher may be constrained by the fact that the training is voluntary, has a limited number of participants, and will not necessarily be repeated in the future. However, Shadish and colleagues (2002) noted that improving research is not just about increasing sample size, stating that the robustness of findings from any given sample should be confirmed through replication and meta-analysis. Ideally, reviewers should consider that the realism of actually studying what happens in the field might outweigh concerns of small sample sizes.

A similar tradeoff involves deciding what to do when the setting rather than the participants are treated as the randomly assigned variable. For instance, Greenberg (2002) studied the extent to which actual employees were willing to steal from their organizations as a function of several variables, one of which was whether they came from an organization that either did or did not have an ethics program in place. Given that participants were drawn from only two organizations (one had and one did not have an ethics program), it is difficult to know how reviewers should treat this design element. Greenberg's efforts are laudable (and he did have other participant variables that were randomly assigned), but this is a case in which causal conclusions are limited—maybe differences between the two organizations were more plentiful than just the presence or absence of an ethics program; hence it is difficult to interpret the power of the ethics programs. Thus, reviewers should be sensitive to limitations in what can and cannot be manipulated and randomly assigned, but also hold very high expectations for these variables as they are the essence of what we consider to be necessary for quality field research.

Another tradeoff is deciding whether measuring only a few authentic behaviors offsets getting a larger number of self-reported (and sometimes second-source) responses on multiple standardized scales and inventories. When conducting field experiments, it is sometimes not possible to ask participants to complete surveys about their individual differences, opinions and reactions, and ratings on scaled inventories. Indeed, some of the participants in field experiments do not even know that they are in a study; informing them of their participation would change substantially their behaviors

and reactions in a way that made their data artificial and potentially misleading. Thus, field experimenters might examine whether people are willing to mail a letter for someone else, how close in proximity they are willing to sit next to each other, and/or how many company pencils they take from the supply closet—these and other sorts of behaviors might be wonderfully valid measures of helping behavior, social distance, and stealing from the company—and researchers simply might not be able to so accurately approximate the measurement of these constructs by simply asking individuals to report on their intentions. Thus, it may be appropriate for reviewers to favorably evaluate solid behavioral measures even when self-reported or individual difference measures are not available.

There is also sometimes a tradeoff between testing moderation or mediation and showcasing a novel phenomenon. In conducting field experiments, researchers often lose the ability to measure a great many constructs but also often gain the ability to take a novel approach to measuring workplace behaviors in original ways. Bullock, Green, and Ha (2010) called attention to the paucity of research that conducts experimental analysis of mediation. Such novel approaches often unveil findings that provide rich details into phenomenon, provoke new considerations for theory building or application, and provide greater understanding of the associated complexities (e.g., participant emotions, situational variables). For instance, Singletary and Hebl (2009) showed that gay and lesbian applicants were able to alter the amount of interpersonal discrimination they received by engaging in one of three strategies (i.e., acknowledgment, compensatory positivity, individuation), but they did not show the mechanisms, reasons, or mediators for such findings. In fact, the Singletary and Hebl study was relatively short on dependent measures; however, the study importantly revealed that strategies used in real job application situations could reduce authentic expressions of discrimination. Thus, reviewers might more carefully consider manuscripts involving experimental field research in which authors might not be able to fully explain the mechanisms behind every outcome when the research might provide remarkably novel data and/or insights.

One way that researchers might successfully respond to the criticisms that they receive is to pair experimental field studies with other more traditional methods. That is, a laboratory study might help point to the mechanism of interest or a survey study that underscores the outcomes of an effect found in a field experiment. Indeed, McGrath's conclusion that all methods are "lousy" was followed by the suggestion that "such discussions might better engage in questions of how best to combine multiple strategies" (see also Oakley, 2000). Locke (1986) agreed and enlisted scholars to examine the degree to which laboratory studies were replicated in the field within industrial-organizational, organizational behavior, and human resource management, illustrating the essential features that required further study. Similarly, Carpenter et al. (2005) suggested precisely this path:

> The obvious solution is to conduct experiments both ways: with and without naturally occurring field referents and context. If there is a difference, then it should be studied. If there is no difference, one can conditionally conclude that the field behavior *in that context* travels to the lab environment. (p. 12)

Conclusion

Although field experimenters should not "break up" with organizational psychology just yet, it is important to heed Cialdini's (2009) insistence that scholars must

> reassign substantially more value to field research than has been the case in recent times. It should be taught regularly in our graduate methods classes across organizational sciences, there should be prestigious awards (such as the Cialdini Award for Field Research) designated for the best of it, and it should be given more grace (and space) in the loftiest of journals. (p. 6)

This argument should be further specified to focus on *field experiments*. Through field experiments, researchers can understand causal effects that emerge in real-world workplaces among genuine working people. Since this is precisely the goal of our science, experimental field research is the direction organizational research must take.

Declaration of Conflicting Interests

Funding

References

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

Basken, P. (2010). Academe hath no fury like a fellow professor deceived. *Chronicle of Higher Education*. Retrieved from http://chronicle.com/article/Academe-Hath-No-Fury-Like-a/65466/

Berkman, E. T., & Lieberman, M. D. (2011). What's outside the black box? The status of behavioral outcomes in neuroscience research. *Psychological Inquiry*, *22*, 100-107.

Bertrand, M., & Mullainathan, S. (2004). Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination. *American Economic Review*, *94*, 991-1013.

Beyer, J. M., & Trice, H. M. (1984). A field study of the use and perceived effects of discipline in controlling work performance. *Academy of Management Journal*, *27*, 743-764.

Blau, F. D., & Kahn, L. M. (2000). Gender differences in pay. *The Journal of Economic Perspectives*, *14*, 75-99.

Bolton, G. M. (1974). The lost letter technique as a measure of community attitudes toward a major social issue. *The Sociological Quarterly*, *15*, 567-570.

Borsboom, D., Kievit, R. A., Cervone, D., & Hood, S. B. (2009). The two disciplines of scientific psychology, or: The disunity of psychology as a working hypothesis. In J. Valsiner, P. C. M. Molenaar, M. C. D. P. Lyra, & N. Chaudary (Eds.), *Dynamic process methodology in the social and developmental sciences* (pp. 67-98). New York, NY: Springer.

Botsford Morgan, W., Singletary Walker, S., Hebl, M. R., & King, E. B. (2012). *Remediating interpersonal discrimination towards pregnant job applicants*. Unpublished manuscript.

Brief, A. P., Butcher, A. H., & Roberson, L. (1995). Cookies, disposition, and job attitudes: The effects of positive mood-inducing events and negative affectivity on job satisfaction in a field experiment. *Organizational Behavior and Human Decision Processes*, *62*, 55-62.

Buckley, M. R., Mobbs, T. A., Mendoza, J. L., Novicevic, M. M., Carraher, S. M., & Beu, D. S. (2002). Implementing realistic job previews and expectation lowering procedures: A field experiment. *Journal of Vocational Behavior*, *61*, 263-278.

Bullock, J. G., Green, D. P., & Ha, S. E. (2010). Yes, but what's the mechanism? (Don't expect an easy answer). *Journal of Personality and Social Psychology*, *98*, 550-558.

Carpenter, J. P., Harrison, G. W., & List, J. A. (2005). Field experiments in economics: An introduction. *Research in Experimental Economics*, *10*, 1-16.

Cialdini, R. B. (2009). We have to break up. *Perspectives on Psychological Science*, *4*, 5-6.

Cialdini, R. B., Bickman, L., & Cacioppo, J. T. (1979). An example of consumeristic social psychology: Bargaining tough in the new car showroom. *Journal of Applied Social Psychology*, *9*, 115-126.

Combs, G. M., & Luthans, F. (2007). Diversity training: Analysis of the impact of self-efficacy. *Human Resource Development Quarterly*, *18*, 91-120.

Cook, T. D., & Campbell, D. T. (1976). The design and conduct of quasi-experiments and true experiments in field settings. In M. Dunnette (Ed.), *Handbook of industrial and organizational psychology*. Skokie, IL: Rand McNally.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design & analysis issues for field settings*. Boston, MA: Houghton Mifflin Company.

Correll, S. J., Benard, S., & Paik, I. (2007). Getting a job: Is there a motherhood penalty? *American Journal of Sociology*, *112*, 1297-1339.

Dvir, T., Eden, D., Avolio, B. J., & Shamir, B. (2002). Impact of transformational leadership on follower development and performance: A field experiment. *Academy of Management Journal*, *45*, 735-744.

Eden, D. (1985). Team development: A true field experiment at three levels of rigor. *Journal of Applied Psychology*, *70*, 94-100.

Eden, D. (1990). Pygmalion without interpersonal contrast effects: Whole groups gain from raising manager expectations. *Journal of Applied Psychology*, *75*, 394-398.

Eden, D. (1995). Seasickness as a self-fulfilling prophecy: Raising self-efficacy to boost performance at sea. *Journal of Applied Psychology*, *80*, 628-635.

Ganster, D. C., Mayes, B. T., Sime, W. E., & Tharp, G. D. (1982). Managing organizational stress: A field experiment. *Journal of Applied Psychology*, *67*, 533-542.

Ginges, J., Atran, S., Sachdeva, S., & Medin, D. (2011). Psychology out of the laboratory: The challenge of violent extremism. *American Psychologist*, *66*, 507-519.

Gordon, M. E., Slade, L. A., & Schmitt, N. (1986). The "science of the sophomore" revisited: From conjecture to empiricism. *The Academy of Management Review*, *1*, 191-207.

Grant, A. M., & Wall, T. D. (2009). The neglected science and art of quasi-experimentation: Why-to, when-to, and how-to advice for organizational researchers. *Organizational Research Methods*, *12*, 653-686.

Greenberg, J. (1982) Approaching equity and avoiding inequity in groups and organizations. In J. Greenberg & R. L. Cohen (Eds.), *Equity and justice in social behavior* (pp. 389-435). New York, NY: Academic Press.

Greenberg, J. (1988). Equity and workplace status: A field experiment. *Journal of Applied Psychology*, *73*, 606-613.

Greenberg, J. (2002). Who stole the money, and when? Individual and situational determinants of employee theft. *Organizational Behavior and Human Decision Processes*, *89*, 985-1003.

Greenberg, J., & Folger, R. (1988). *Controversial issues in social research methods*. New York, NY: Springer-Verlag Publishing.

Griffin, R. W. (1983). Objective and social sources of information in task redesign: A field experiment. *Administrative Science Quarterly*, *28*, 184-200.

Hebl, M. R., Foster, J. B., Mannix, L. M., & Dovidio, J. F. (2002). Formal and interpersonal discrimination: A field study of bias towards homosexual applicants. *Personality and Social Psychology Bulletin*, *28*, 815-825.

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.

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

Holland, R. W., Aarts, H., & Langendam, D. (2006). Breaking and creating habits on the working floor: A field-experiment on the power of implementation intentions. *Journal of Experimental Social Psychology*, *42*, 776-783.

James, E. H., & Wooten, L. P. (2006). Diversity crises: How firms manage discrimination lawsuits. *Academy of Management Journal*, *49*, 1103-1118.

Jawahar, I. M., & Williams, C. R. (1997). Where all the children are above average: The performance appraisal purpose effect. *Personnel Psychology*, *50*, 905-925.

Kaplan, S., Bradley, J., Ahmad, A., Anderson, A., Hargrove, A., & Lindsey, A. (2012). *Testing different programs to improve workplace well-being*. Unpublished manuscript.

Kelman, H. C. (1967). Human use of human subjects: The problem of deception in social psychological experiments. *Psychological Bulletin*, *67*, 1-11.

Kifner, J. (2001, September 10). A recipe for distress. *The International Herald Tribune*, p. 3.

King, E. B. (2008). The effect of bias on the advancement of working mothers: Disentangling legitimate concerns from inaccurate stereotypes as predictors of advancement in academe. *Human Relations*, *61*, 1677-1711.

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

King, E. B., Mendoza, S., Madera, J., Hebl, M. R., & Knight, J. L. (2006). What's in a name? A multiracial investigation of the role of occupational stereotypes in selection decisions. *Journal of Applied Social Psychology*, *36*, 1145-1159.

Lee, R. M. (1993). *Doing research on sensitive topics*. Thousand Oaks, CA: Sage.

Lewandowski, G. W., & Strohmetz, D. B. (2009). Actions can speak as loud as words: Measuring behavior in psychological science. *Social and Personality Psychology Compass*, *3*, 992-1002.

Locke, E. A. (Ed.). (1986). *Generalizing from laboratory to field settings*. Lexington, MA: Lexington Books.

Lyness, K. S., & Thompson, D. E. (1997). Above the glass ceiling? A comparison of matched samples of female and male executives. *Journal of Applied Psychology*, *82*, 359-375.

Madera, J. M., King, E. B., & Hebl, M. R. (2012). Enhancing the effects of sexual orientation diversity training: The effects of setting goals and training mentors on attitudes and behaviors. *Journal of Business and Psychology*. Retrieved from http://rd.springer.com/article/10.1007/s10869-012-9264-7#

May, D. R., Reed, K., Schwoerer, C. E., & Potter, P. (2004). Ergonomic office design and aging: A quasi-experimental field study of employee reactions to an ergonomics intervention program. *Journal of Occupational Health Psychology*, *9*, 123-135.

McGrath, J. E. (1981). Dilemmatics: The study of research choices and dilemmas. *American Behavioral Scientist*, *25*, 179-210.

Medin, D. L. (2011). Fields for psychology. *American Psychological Society: Observer*, *24*(7). Retrieved from http://www.psychologicalscience.org/index.php/publications/observer/2011/september-11/fields-for-psychology.html

Milkman, K. L., Akinola, M, & Chugh, D. (2012). Temporal distance and discrimination: An audit study in academia. *Psychological Science*, *23*, 710-717.

Mitchell, T. R. (1985). An evaluation of the validity of correlational research conducted in organizations. *Academy of Management Review*, *10*, 192-205.

Mook, D. G. (1983). In defense of external invalidity. *American Psychologist*, *38*, 379-387.

Oakley, A. (2000). *Experiments in knowing: Gender and method in the social sciences*. Cambridge, MA: Polity Press.

Paluck, E. L. (2006). Diversity training and intergroup contact: A call to action research. *Journal of Social Issues*, *62*, 577-595.

Paluck, E. L., & Green, D. P. (2009). Prejudice reduction: What works? A review and assessment of research and practice. *Annual Review of Psychology*, *60*, 339-367.

Riach, P. A., & Rich, J. (2002). Field experiments of discrimination in the market place. *The Economic Journal*, *112*, F480-F518.

Salovey, P., & Williams-Piehota, P. (2004). Field experiments in social psychology: Message framing and the promotion of health protective behaviors. *American Behavioral Scientist*, *47*, 488-505.

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

Sears, D. O. (1986). College sophomore in the laboratory: Influences of a narrow data base on psychology's view of human nature. *Journal of Personality and Social Psychology*, *51*, 515-530.

Shadish, W. R. (2002). Revisiting field experimentation: Field notes for the future. *Psychological Methods*, *7*, 3-18.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inferences*. Boston, MA: Houghton Mifflin.

Sieber, J., & Stanley, B. (1988). Ethical and professional dimensions of socially sensitive research. *American Psychologist*, *43*, 49-55.

Singletary, S. L., & Hebl, M. R. (2009). Compensatory strategies for reducing interpersonal discrimination: The effectiveness of acknowledgements, increased positivity, and individuating information. *Journal of Applied Psychology*, *94*, 797-805.

Society for Industrial and Organizational Psychology, Inc. (1999). *Guidelines for education and training at the doctoral level in industrial-organizational psychology*. Bowling Green, OH: Author. Retrieved from http://www.siop.org/PhDGuidelines98.aspx

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.

Williams, A. (2010). *The stigma of obesity and customer service: What about men?* (Undergraduate senior honors thesis). Rice University, Houston, TX.

Zitek, E. M., & Hebl, M. (2007). The role of social norm clarity in the influenced expression of prejudice over time. *Journal of Experimental Social Psychology*, *43*, 867-876.

Bios

Eden B. King joined the faculty of the Industrial-Organizational Psychology program at George Mason University after earning her PhD from Rice University in 2006. Dr. King is pursuing a program of research that seeks to guide the equitable and effective management of diverse organizations. Her research integrates organizational and social psychological theories in conceptualizing social stigma and the work-life interface.

Michelle R. Hebl is an applied psychologist who is part of the Industrial-Organizational program at Rice University. Her research focuses on issues related to diversity and discrimination. She is particularly interested in examining subtle ways in which discrimination is displayed, and how such displays might be remediated by individuals and/or organizations.

Whitney Botsford Morgan is an Assistant Professor in the College of Business at the University of Houston-Downtown. Her scholarship focuses on the experiences of women and mothers in organizations. In particular, she has studied the interpersonal experiences of pregnant women and new mothers to build understanding of factors that influence their career decisions.

Afra Saeed Ahmad is a doctoral student in Industrial-Organizational Psychology at George Mason University. Her research considers the ways people manage concealable aspects of their identities in workplace interactions. More specifically, she is interested in understanding the experiences of religious minorities in contemporary organizations.