Organizational Research Methods
Volume 12 Number 3
July 2009 554-566

10.1177/1094428107300396
http://orm.sagepub.com
hosted at
http://online.sagepub.com

Designing Experiments That Generalize

Scott Highhouse
Bowling Green State University

Organizational research has relied too heavily on methods characterized by passive observation, likely because there is a widespread belief that experimental research has limited generalizability. However, this is often because researchers (and reviewers or editors) misunderstand the nature of generalizability and what it requires. This article reiterates the importance of experimental research for understanding organizational phenomena and separates the legitimate concerns about experimental generalizability from the irrelevant ones. Whereas most criticisms of experiments focus on sample characteristics and mundane realism (i.e., superficial resemblance to the real world), more attention needs to be paid to the degree to which the treatment manipulation is valid, representative, and strong.

Keywords: *experimental design; generalizability; construct validity; stimulus sampling*

Randomized experiments are the most potent research design for determining whether or not *x* causes *y*. Even the most elegant of statistical procedures will not enable a passive-observational study to meet the basic requirements for inferring causation (i.e., evidence of concomitant variation, time-ordering, and elimination of alternative explanations).[1] Despite the benefits of randomized experimentation (hereafter referred to as experimentation) for making causal inferences, organizational scholarship has historically been characterized by an unhealthy over-reliance on passive observation (Dipboye, 1990; Fromkin & Streufert, 1976; Greenberg & Tomlinson, 2004; Podsakoff & Dalton, 1987; Weick, 1965). Austin, Scherbaum, and Mahlman (2002) analyzed the content of the *Journal of Applied Psychology* over the past several decades and noted the instances in which articles used methods characterized by passive observation, experimentation, or something else (e.g., archival data analysis). The authors found that, in the year 2000, only 31% of the studies published were experimental; only 22% were conducted in the lab. This represented a 40% decline in the publication of experimental studies since 1970. In contrast, Austin and his colleagues observed a 15% increase in the use of passive-observational studies, which constituted 54% of the studies published in 2000.

Stopping short of banning laboratory experiments, some organizational journals explicitly discourage submissions that fail to take into account the richness of the organizational setting (*Journal of Organizational Behavior*) or that use students as subjects (*Journal of Occupational and Organizational Psychology*). Although the mission statement of the

Author's Note: I am very grateful for the constructive and thoughtful input of Margaret Brooks, Mike Doherty, Jen Gillespie, Dave Harrison, Paul Sackett, and Mike Zickar. I take full responsibility, however, for any errors of logic or common sense. Correspondence concerning this article should be addressed to Scott Highhouse, Department of Psychology, Bowling Green State University, Bowling Green, OH 43403. Electronic mail may be sent to shighho@bgnet.bgsu.edu.

Academy of Management Journal welcomes all empirical methods, an inspection of the content of the 2005 volume revealed that only 3 out of 58 articles (5%) used experimental manipulations of any sort. Many scholars have observed the common stereotypes of experiments—particularly ones that are conducted in the laboratory (see Dobbins, Lane, & Steiner, 1988; Greenberg & Tomlinson, 2004; Griffin & Kacmar, 1991; Ilgen, 1986; Scandura & Williams, 2000; Stone-Romero, 2002). Experiments are seen as contrived, irrelevant, and even misleading. Most of the concerns are aimed at limits to external validity or the ability to generalize inferences across populations, settings, and variables (D. T. Campbell & Stanley, 1967). These critiques often overlook the fact that field-based passive observation may involve taking workers from one specific organization away from their jobs to self-report about things they have not thought about before. Or, it may involve correlating organization-level data (e.g., budgetary slack and organizational innovation) that is rife with problems of deficiency and contamination.

The point is not that one method is superior to other methods for investigating organizational phenomena. It is simply that experimentation, despite seemingly occupying lower status in the pecking order of organizational research methods, is a uniquely superior method to investigate certain kinds of questions (Haack, 2003). Nevertheless, there are some genuine shortcomings of experiments that have been conducted in the organizational literature, particularly ones that limit our ability to generalize inferences to behavior in organizations. The goal of this article is to separate the legitimate criticisms of experiments from the specious ones and to provide suggestions for how researchers can design experiments that generalize to behavior in organizations.

Ecological Validity Versus External Validity

I purposely did not focus this article on the design of experiments with ecological validity, which is concerned with the realism of the experimental methods, materials, and settings.[2] This is because experiments do not need to mirror the external environment for us to generalize inferences across populations, settings, and variables. There is a real danger in confusing ecological validity (generalizing to) with external validity (generalizing across).[3] Indeed, it would be impossible to design an experiment that generalizes to the "typical" organizational situation. What is a typical organization? Is a fast-food restaurant typical? How about a *Fortune 50* corporation? Can we translate findings from one to the other?

Focusing on the surface similarity between an experimental situation and an organizational situation often misses the point (Cook & Campbell, 1979; Shapiro, 2002). It diverts attention away from underlying theoretical principles, which are often more powerful for generalizing across organizational situations (J. P. Campbell, 1990). Generalizability happens when we understand the process by which a result comes about. This means understanding the causal process. When we get caught up in the distinctiveness of the research setting, it implies that we are testing effects in settings rather than testing theories that should apply to multiple (especially organizational) settings.

When researchers design experiments, they rarely presume to find effects that are immediately applicable to organizational situations. An exception to this might be an examination of training effects that includes random assignment of participants to either a training group

or a control group. The goal here is not only to determine the effects of training, but to determine the effect *size* of the specific training intervention. When a determination of effect size from a single experiment is the goal, then ecological validity is necessary.

The design of experiments is usually done, however, with the goal of finding whether one operationalization of a construct (an independent variable) causes a change in another operationalization of a construct (a dependent variable)—holding all else constant. In other words, the goal is to test a theorized effect, not a statistical effect (Calder, Phillips, & Tybout, 1981; Chow, 1996). The experimenter is not immediately interested in generalizing the size of the effect observed in the experiment. Rather, the experimenter is interested in assessing the status of the theory. As Chow (1997) noted, generalizability is not a property of a study; it is a property of a theory.

Most experiments are designed with the goal of generalizing theoretical explanations beyond the specific experimental circumstances. Theoretical propositions are the vehicles for generalizing to the real world (Schlenker & Bonoma, 1978). The ability to generalize from one situation to another requires an understanding of underlying principles and recognizing which principles apply in which situations (Shapiro, 2002). Further field research, involving passive observation, can be used to identify boundary conditions of the theoretical explanation as it is directly applied to specific organizational circumstances (Brinberg & McGrath, 1985; Fromkin & Streufert, 1976). Because the goal of experiments is usually to apply the theory beyond the research setting, the degree to which the specific sample represents the population of interest is of less importance. Any sample within the theory's domain is a relevant sample (Calder et al., 1981). The real concern for generalizing theoretical explanations is ensuring that the operationalizations of the constructs allow generalizable inferences.

To this point, I have attempted to establish that experiments are underutilized in organizational research. I have suggested that this underutilization is at least partly a result of a general misunderstanding of what is required for an experiment to generalize. It is unproductive and misleading for organizational researchers to say that experiments must generalize *to* organizations, when the ultimate goal is to generalize *across* organizations. Generalizing across organizations requires theory testing, and theory testing requires generalizable causes and effects. My focus in the following sections is on causes (i.e., independent variables). This is not because effects (i.e., dependent variables) are unimportant but because measurement issues receive the bulk of attention in organizational research. Very little attention is given to the design of treatments that are valid, representative, and strong.

Designing Better Treatments

Checking Manipulations

Although construct validity receives considerable attention in correlational research, less attention is given to the topic in experimentation. This is ironic, because D. T. Campbell and Fiske's (1959) notion of convergent validity was an adaptation of Garner, Hake, and Eriksen's (1956) notion of converging operations in experimentation (Grace, 2001). The use

of manipulation check procedures dates as far back as Farnsworth and Misumi's (1931) examination of the effects of fame on judgments of the quality of artwork. Festinger (1953), however, is credited with emphasizing their systematic use in experimentation involving abstract concepts. For an organizational researcher, determining whether a treatment is representative of the latent independent variable should be as important as determining whether a collection of items is a reliable and valid indicator of the latent dependent variable.

Perdue and Summers (1986) stressed the need for investigators to consider both convergent and discriminant validity as they apply to experimental manipulations. From a convergent validity perspective, a manipulation check attempts to directly measure the independent variable of interest. For example, an experimenter studying the effects of management sincerity on reactions to negative feedback would want to show that participants in the sincere condition perceived the communication to be more heartfelt and genuine than did participants in the insincere condition. From a discriminant validity perspective, the manipulation check should measure possible unintended effects of the independent variable. For example, the above experimenter might want to show that the information communicated by the manager in the sincere condition is not seen as more positive or as less personal—attributes not conceptualized as part of the sincerity construct. Although manipulation checks are often thrown in to assuage reviewer concerns about manipulation strength and fidelity, too often they do not even get at the question of whether the manipulation worked as intended. Furthermore, they rarely get at the issue of unintended consequences. Good manipulation checks require thought, precision, and creativity.

Manipulation checks are critical to the process of determining whether the manipulation worked as intended. They say nothing, however, about the *typicality* of the chosen stimulus within the population of stimuli that represent the construct. Whereas the construct validity of the manipulation is crucial for generalizing inferences from an experiment (i.e., internal validity), equally important is the degree to which the manipulation is representative of the class of manipulations that represent the construct (i.e., external validity). This usually requires stimulus sampling.

Sampling Stimuli

More than 60 years ago, Egon Brunswik (1944) observed a double standard in the practice of statistical sampling. Brunswik thought it odd that psychological researchers demanded that subjects be sampled for generalization to the population but paid no attention to whether experimental stimuli were representative of a defined population of stimuli (see Dhami, Hertwig, & Hoffrage, 2004; Hammond, 1948, 1996). Since then, a number of authors have pointed out the need to use multiple instances of a stimulus category when operationalizing a construct for use as an independent variable (e.g., D. T. Campbell & Stanley, 1967; Clark, 1973; Rosenthal & Rosnow, 1991; Wells & Windschitl, 1999). This issue has rarely been raised, however, within the context of organizational research (cf. Fontenelle, Phillips, & Lane, 1985).

Consider a researcher interested in studying the effects of irrelevant information in résumés on predictions of job performance. The researcher is familiar with a judgment

phenomenon called the "dilution effect" (Nisbett, Zukier, & Lemley, 1981), in which predictions (e.g., grade point average) based on a valid predictor (e.g., SAT score) are diluted by cues having no validity (e.g., enjoys houseplants). The researcher, thus, sets about creating an experiment that manipulates résumé information relevance. A common method for designing a study like this would be to have one résumé containing only relevant information and one that includes both relevant and irrelevant information. The irrelevant information might be a section of the résumé that lists the applicant's hobbies and interests (i.e., science fiction writing, puzzles, ballroom dancing).

The problem with this traditional experimental design is that it ignores the unmeasured and uncontrolled aspects of the stimuli chosen to represent the category—that is, irrelevant résumé information. It would be impossible to conclude from this study that irrelevant information leads to dilution in performance predictions. One can only conclude that this particular set of hobbies and interests leads to dilution. Perhaps the hobbies and interests chosen were so atypical as to lead to an atypical image of the stimulus person. Although a dilution effect for atypical résumé information may be interesting, the original theory did not make any delimiting statements about the nature of the irrelevant information. A more appropriate test of the phenomenon would be to use stimulus sampling, in which the specific irrelevant information is varied across subjects *within* the experimental condition.[4] For example, the researcher might obtain a large sample of résumés from the university career center and examine what kinds of hobbies and interests are typically listed. The researcher might choose from these a random sample of the most commonly listed items and create multiple versions of the hobbies and interests section in the résumé used for the irrelevant information condition.

Consider two real-world examples of failures to sample stimuli that I arbitrarily pulled from recent issues of the *Journal of Applied Psychology*. In one study, the authors were interested in how people react to offers of help from coworkers. The authors of this article were concerned with ecological validity, such that they designed a work area that mirrored a typical office in a modern organization. To enhance mundane realism, the investigators equipped the office with computers, telephones, and Internet connections, and the office was furnished with desks, chairs, and file cabinets. The participants were actual temporary workers, recruited to work for 4 hours as administrative assistants.

The problem with the generalizability of their experiment concerned the manipulation of their primary independent variable, which was an operationalization of the construct "imposed support." Imposed support fundamentally involves providing help to a coworker without asking if he or she wants the help. The support manipulation in this study involved staffing the office with a female confederate who said, in a friendly tone, that she would help get missing information for a file the participant was working on. The confederate was instructed to provide the help in lieu of a request for help from the participant and to disregard any protests in a friendly manner.

Aside from the fact that the researchers used only a female confederate, the use of the same female confederate for every participant did not allow us to generalize the effects of support beyond that offered by this particular person. There is no way for us to know if this confederate had idiosyncratic vocal or facial characteristics that could have made this imposed support more or less abrupt or condescending. Moreover, the behavior of the confederate may be idiosyncratic in terms of the timing of the support or in terms of the type

of help provided. Thus, the broader construct "support" is operationalized by use of a single case that may or may not be representative of the central tendencies of the population of cases that are contained within the construct (see Wells & Windschitl, 1999). The solution to this problem is not to replicate the study using a different confederate. The solution is to appropriately sample the cases contained within the construct. The ideal situation would have been to completely cross participants with confederates. Short of that, the experimenters could have rotated among several confederates who presumably vary on the otherwise peculiar features any one confederate may bring to the experiment.

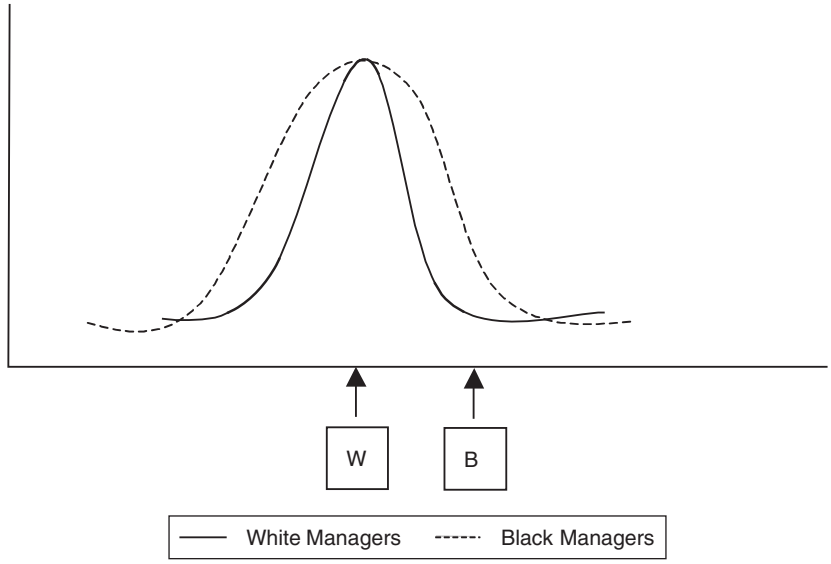
A less obvious example of failure to sample stimuli occurred in a study of reactions to recruitment Web sites that varied in the representation of minorities as employees. In one condition, all managerial employees pictured were White. In another condition, one of the White managers in the image was replaced with a Black manager. Although the investigators reported that they matched the White and Black managers on rated attractiveness, the experiment uses only one stimulus person to represent an entire category—African American manager. The use of only one Black manager confounds the unique characteristics of this stimulus with the category. If one accepts the idea that there is variance in the degree to which Black managers appear to be desirable coworkers or managers, then one must also accept that finding one person who is representative for all Black managers is nearly impossible.

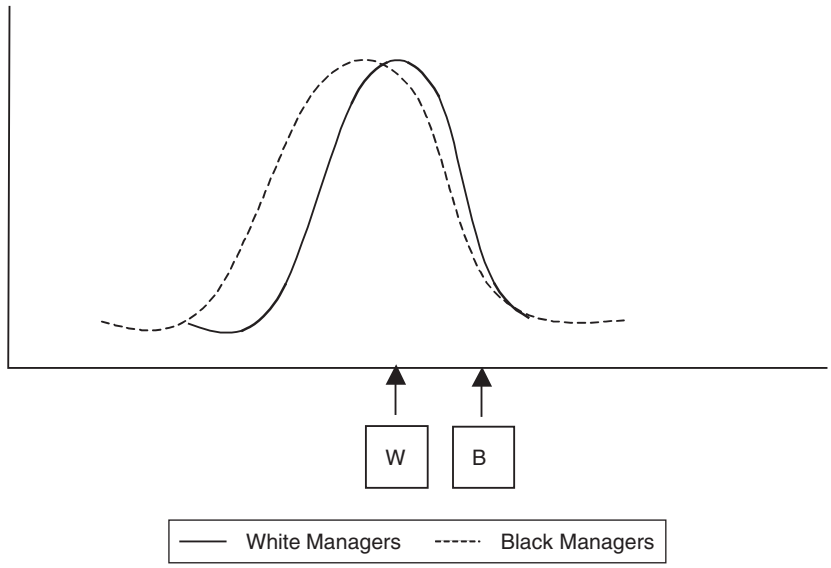
Figure 1 shows two hypothetical distributions of Black managers and White managers in terms of desirability as coworkers in the eyes of White applicants. Note that the hypothetical distributions overlap considerably and show a wide degree in variability in desirability for both Black and White managers. The top graph shows a study that chooses an advertisement depicting a Black manager who is about one-half standard deviation above the mean in desirability, and one White manager who is at the mean. This study would result in a Type I error because the stimulus person chosen to represent the category of African American manager is not representative of the typical instance of that category. The bottom graph shows another hypothetical example. This time, the mean desirability of White managers as coworkers is higher than the mean desirability of Black managers in the eyes of White applicants. The Black manager depicted in the advertisement is, again, above the mean in desirability. The White manager is at the mean. A study based on these stimuli would result in a Type III error, which is when an experimenter gets the direction wrong (Kaiser, 1960).

Avoiding Type III Errors

With Type III errors, the null hypothesis is false, and the null is rejected. However, the direction of the true population difference is opposite of the direction of the observed difference (Kaiser, 1960; Leventhal & Huynh, 1996). Experimenters are particularly prone to Type III errors when they fail to sample stimuli in situations where it is necessary. Recall the dilution phenomenon described earlier in the hypothetical study of résumé judgments. Researchers are finding that the dilution effect holds up only under a limited set of conditions and that the effect can even be reversed depending on the type of nondiagnostic stimuli used (Labella & Koehler, 2004; Peters & Rothbart, 2000). For example, Peters and Rothbart (2000) showed that, unless the nondiagnostic information is highly

Figure 1
Two Hypothetical Frequency Distributions of White Managers (W) and Black Managers (B) on Desirability as Coworkers in the Eyes of White Applicants





atypical, the most common finding is enhancement, not dilution. This may be an example where failure to sample stimuli resulted in a Type III error that endured for nearly 20 years.

The greater the overlap between two stimulus distributions, the greater the chance of a Type III error. For example, imagine one was interested in studying biases in the perceived

customer service performance of older workers versus younger workers. Imagine further that the experimenter creates videotapes depicting a customer service episode, one with an older worker and one with a younger worker. The risk of a Type III error is contingent on the degree to which the distribution of perceived older-worker friendliness overlaps the distribution of perceived younger-worker friendliness. The greater the overlap, the more likely it is that the researcher will choose exemplars that do not represent the typical member of each category. It is unclear how often Type III errors occur in the organizational literature, but occasional debates over the directionality of effects suggest that these errors probably do happen (e.g., Bandura & Locke, 2003; Highhouse, Luong, & Sarkar, 1999; Johns, Xie, & Fang, 1992; Kluger & DeNisi, 1996; Ofir & Mazursky, 1997; Slattery & Ganster, 2002; Vancouver, 2005).

Enhancing Power

Although I have concentrated on design problems that can cause Type I and Type III errors, investigators need to be concerned with Type II errors as well. Whereas power is primarily an issue of sample size for correlational studies, experimenters also need to concern themselves with manipulation strength and fidelity. I suspect that there are two reasons that organizational researchers, as opposed to basic experimentalists, ignore the issue of manipulation strength. First, organizational researchers are likely (rightfully) worried about accusations of demand effects, or the possibility that participants are acting in response to knowledge that the manipulation is occurring. Strong manipulations, however, are not synonymous with demand characteristics. Second, organizational researchers are often so concerned with mundane realism that they ignore the fact that manipulations need to be strong enough to cause an effect. Just as drug trials need to ensure that the dosage is strong enough to have an effect on the patients, investigators need to ensure that the manipulated leadership style, for example, is experienced by the followers. It makes no sense to manipulate leadership style subtly, in a realistically "noisy" work environment, if the experimental participant does not perceive the intended leadership behaviors. To the extent that making an experiment similar to the real world interferes with the ability to draw inferences from the results, one should generally sacrifice real-world authenticity for internal validity (Anderson & Bushman, 1997; Berkowitz & Donnerstein, 1982; Mook, 1983).

Power, therefore, has as much to do with how well designed an experiment is as it does with how big the sample is. Those who evaluate experiments are rightly concerned with Type I errors more than with Type II errors. Those who *design* experiments, however, need to be more concerned with Type II errors. Experiments with null results are often poorly designed, with weak manipulations that lack construct validity. This is probably why the limitations of null hypothesis significant testing are of concern to fields that rely heavily on passive observation but not to fields that are characterized more by experimentation. As Abelson (1997) noted, the preoccupation with effect size ignores the important issue of "cause size" in experimentation.

The problem with applying correlational logic to experimental research is that it ignores the fact that passive-observational researchers are often dedicated to their operationalizations

of constructs (e.g., NEO conscientiousness scores as predictors of job performance), making interpretation of effect size in terms of importance perfectly appropriate. In other words, the investigator is interested in the research question for its own sake. The resulting effect size gives a good indication of the strength and utility of a typical conscientiousness measure for predicting job performance. It is rare in experimental research that the treatment of interest is used as the experimental manipulation (one exception might be an organizational development intervention) or that the investigator is interested in the experimental question for its own sake (Chow, 1988). Experiments are generally concerned with testing implications from theories, not with the direct effects of the independent variables.

Considering the importance of manipulation strength in experimentation, therefore, concern with effect size estimation in this context is erroneous (Abelson, 1997; Chow, 1996; Fern & Monroe, 1996; Prentice & Miller, 1992). Effect size estimates for experiments are only useful when the levels of the independent variable are selected randomly and are thus representative of the levels in the relevant population (Fern & Monroe, 1996). Even then, small effects may be more important than larger ones (Prentice & Miller, 1992). The relevant consideration for an experimenter is whether the obtained result is consistent with that predicted by the theory.

Conclusion

Twenty years ago, Locke (1986) gathered a group of organizational scholars to examine evidence of lab versus field differences and concluded that instances in which findings differ in the two settings are rare (see also Anderson & Bushman, 1997; Peterson, 2001). Nevertheless, the bias against experiments in organizational research has certainly not abated and has probably grown even stronger (Stone-Romero, 2002). Those who reject experimental findings as irrelevant are ignoring the fact that experiments are the only way to test many key propositions about behavior in organizations.

I have suggested in this article that the generalizability of experiments has little to do with the degree to which they take place in, or mirror, the real world. Rather, the generalizability of experiments is dependent on the degree to which the operationalizations of the constructs are true to the constructs themselves—this means true not only in the sense of psychometric validity but also in the sense of domain representativeness. This requires careful attention to research design issues that are seldom discussed in the organizational literature, such as measuring confounds, sampling stimuli, and strengthening manipulations.[5]

My goal in raising these issues is not to set a ridiculously high bar for experimental research. One might argue, for example, that the list of stimuli that need to be sampled is so long as to present a hopeless situation for the experimenter. With the imposed support study, for instance, one could argue that appropriately sampling the type of imposed support is just as important as sampling the confederates. For the job advertisement example, one might argue that just as there should be multiple conditions with different Black managers, there should be multiple conditions with different groups of White employees. One might further argue that there are other aspects of the stimuli that could create confounds

(nature of the job, nature of the images, etc.). An awareness of stimulus sampling does not necessitate an obsessive concern with ensuring the representativeness of every nuance of a treatment. Experimenters should merely be prepared to defend the representativeness of their manipulations. This sometimes requires sampling stimuli, but it also might involve reasoned argument for why the manipulation represents a typical instance of the phenomenon. Indeed, the point of this article is to direct consumers of experimental research to the relevant aspects of generalizability, not to create even more barriers to getting experimental research into the public domain.

Referees and editors often condemn submissions for failing to take into account surface similarities between the experiment and the organizational setting to which the investigator wishes to generalize. The realism of the task and stimuli is not relevant to external validity unless the experiment is unrealistic on a dimension that interacts with the manipulations of interest. In addition, the nature of the sample is not relevant to external validity unless there is some reason to suspect that the observed effect is moderated by a property of the population that is not represented in the sample. Referees should be expected to explain why a particular sample or setting nullifies the theoretical contribution of the submission, and authors should be expected to explain why their sample or setting is relevant to the theoretical question they are addressing. For experiments to find acceptance in the organizational literature, experimenters need to do a better job of designing and communicating the value of their experiments, and gatekeepers need to gain a better understanding of the logic of experimentation.

Notes

1. The only exception is the rare case in which a correlation-based causal model can be fully specified in a longitudinal design. Experiments can be properly specified without being fully specified (Schwab, 2005). See James, Mulaik, and Brett (1982) for further treatment of the requirements for drawing causal inference from nonexperimental data.

2. I use the term *ecological validity* as it is commonly used in reference to experimental realism, rather than in the Brunswikian sense of cue prediction (see Hammond, 1996).

3. I am grateful to an anonymous reviewer for articulating this better than I could. A similar distinction between physical fidelity and psychological fidelity is made in the training and human factors literatures.

4. It may be useful to think about this in terms of the distinction between fixed effects and random effects models. Random effects models assume generalization beyond the specific stimuli used (also referred to as variance component models).

5. As one reviewer noted, however, some constructs may be so difficult to manipulate (e.g., transformational leadership) that experimentation may not be the appropriate method to use.

References

Abelson, R. P. (1997). On the surprising longevity of flogged horses: Why there is a case for the significance test. *Psychological Science*, *8*, 12-15.

Anderson, C. A., & Bushman, B. J. (1997). External validity of "trivial" experiments: The case of laboratory aggression. *Review of General Psychology*, *1*, 19-41.

Austin, J., Scherbaum, C., & Mahlman (2002). History of research methods in industrial organizational psychology: Measurement, design, analysis. In S. G. Rogelberg (Ed.), *Handbook of research methods in industrial and organizational psychology* (pp. 77-98). Malden, MA: Blackwell.

Bandura, A., & Locke, E. A. (2003). Negative self-efficacy and goal effects revisited. *Journal of Applied Psychology*, *88*, 87-99.

Berkowitz, L., & Donnerstein, E. (1982). External validity is more than skin deep: Some answers to criticisms of laboratory experiments. *American Psychologist*, *37*, 245-257.

Brinberg, D., & McGrath, J. E. (1985). *Validity and the research process*. Beverly Hills, CA: Sage.

Brunswik, E. (1944). Distal focusing of perception: Size constancy in a representative sample of situations. *Psychological Monographs*, *56*, 1-49.

Calder, B. J., Phillips, L. W., & Tybout, A. M. (1981). Designing research for application. *Journal of Consumer Research*, *8*, 197-207.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, *56*, 81-105.

Campbell, D. T. & Stanley, J. C. (1967). *Experimental and quasi-experimental designs in research*. Chicago: Rand McNally.

Campbell, J. P. (1990). Modeling the performance prediction problem in industrial and organizational psychology. *Handbook of industrial and organizational psychology*. Palo Alto, CA: Consulting Psychologists Press.

Chow, S. L. (1988). Significance test or effect size? *Psychological Bulletin*, *103*, 105-110.

Chow, S. L. (1996). Statistical significance: Rationale, validity, and utility. Thousand Oaks, CA: Sage.

Chow, S. L. (1997). Science, ecological validity and experimentation. *Journal of the Theory of Social Behaviour*, *17*, 181-194.

Clark, H. H. (1973). The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. *Journal of Verbal Learning and Verbal Behavior*, *12*, 335-359.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation design and analysis*. New York: Guilford.

Dhami, M. K., Hertwig, R., & Hoffrage, U. (2004). The role of representative design in an ecological approach to cognition. *Psychological Bulletin*, *130*, 959-988.

Dipboye, R. L. (1990). Laboratory vs. field research in industrial-organizational psychology. In C. L. Cooper & I. T. Robertson (Eds.), *International review of industrial and organizational psychology* (Vol. 5, pp. 1-34). New York: Wiley.

Dobbins, G. H., Lane, I. M., & Steiner, D. D. (1988). A note on the role of laboratory methodologies in applied behavioural research: Don't throw out the baby with the bath water. *Journal of Organizational Behavior*, *9*, 281-286.

Farnsworth, P. R., & Misumi, I. (1931). Further data on suggestion in pictures. *American Journal of Psychology*, *43*, 632.

Fern, E. F., & Monroe, K. B. (1996). Effect-size estimates: Issues and problems in interpretation. *Journal of Consumer Research*, *23*, 89-105.

Festinger, L. (1953). Laboratory experiments. In L. Festinger and D. Katz (Eds.), *Research methods in the behavioral sciences* (pp. 136-172). New York: Rinehart and Windston.

Fontenelle, G. A., Phillips, A. P., & Lane, D. M. (1985). Generalizing across stimuli as well as subjects: A neglected aspect of external validity. *Journal of Applied Psychology*, *70*, 101-107.

Fromkin, H. L., & Streufert, S. (1976). Laboratory experimentation. In M. D. Dunnette (Ed.), *Handbook of industrial and organizational psychology* (pp. 415-465). Chicago: Rand McNally.

Garner, W. R., Hake, H. W., & Eriksen, C. W. (1956). Operationism and the concept of perception. *Psychological Review*, *63*, 149-159.

Grace, R. C. (2001). On the failure of operationism. *Theory & Psychology*, *11*, 5-33.

Greenberg, J., & Tomlinson, E. C. (2004). Situated experiments in organizations: Transplanting the lab to the field. *Journal of Management*, *30*, 703-724.

Griffin, R., & Kacmar, K. M. (1991). Laboratory research in management: Misconceptions and missed opportunities. *Journal of Organizational Behavior*, *12*, 301-311.

Haack, S. (2003). *Defending science–within reason: Between scientism and cynicism*. Amherst, MA: Prometheus.

Hammond, K. R. (1948). Subject and object sampling: A note. *Psychological Bulletin*, *45*, 530-533.

Hammond, K. R. (1996). Upon reflection. *Thinking and Reasoning*, *2*, 239-248.

Highhouse, S., Luong, A., & Sarkar, S. (1999). Research design, measurement, and effects of attribute range on job choice: More than meets the eye. *Organizational Research Methods*, *2*, 37-49.

Ilgen, D. R. (1986). Laboratory research: A question of when, not if. In E. A. Locke (Ed.), *Generalizing from laboratory to field settings* (pp. 257-267). Indianapolis, IN: D.C. Heath.

James, L. R., Mulaik, S. A., & Brett, J. (1982). *Causal analysis. Models, assumptions and data*. Beverly Hills, CA: Sage.

Johns, G., Xie, J. L., & Fang, Y. (1992). Mediating and moderating effects in job design. *Journal of Management*, *18*, 657-676.

Kaiser, H. F. (1960). Directional statistical decisions. *Psychological Review*, *67*, 160-167.

Kluger, A. N., & DeNisi, A. (1996). The effects of feedback interventions on performance: Historical review, meta-analysis and a preliminary feedback intervention theory. *Psychological Bulletin*, *119*, 254-284.

Labella, C., & Koehler, D. J. (2004). Dilution and confirmation of probability judgments based on nondiagnostic evidence. *Memory & Cognition*, *32*, 1076-1086.

Leventhal, L., & Huynh, C. L. (1996). Directional decisions for two-tailed tests: Power, error rates, and sample size. *Psychological Methods*, *1*, 278-292.

Locke, E. A. (1986). Generalizing from laboratory to field: Ecological validity or abstraction of essential elements? In E. A. Locke (Ed.), *Generalizing from laboratory to field settings* (pp. 257-267). Indianapolis, IN: D.C. Heath.

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

Nisbett, R. E., Zukier, H., & Lemley, R. E. (1981). The dilution effect: Nondiagnostic information weakens the implications of diagnostic information. *Cognitive Psychology*, *13*, 248-277.

Ofir, C., & Mazursky, D. (1997). Does a surprising outcome reinforce or reverse the hindsight bias? *Organizational Behavior and Human Decision Processes*, *69*, 51-57.

Perdue, B. C., & Summers, J. O. (1986). Checking the success of manipulations in marketing experiments. *Journal of Marketing Research*, *23*, 317-326.

Peters, E., & Rothbart, M. (2000). Typicality can create, eliminate, and reverse the dilution effect. *Personality and Social Psychology Bulletin*, *26*, 177-187.

Peterson, R. A. (2001). On the use of college students in social science research: Insights from a second-order meta-analysis. *Journal of Consumer Research*, *28*, 450-461.

Podsakoff, P. M., & Dalton, D. R. (1987). Research methodology in organizational studies. *Journal of Management*, *13*, 419-441.

Prentice, D. A., & Miller, D. T. (1992). When small effects are impressive. *Psychological Bulletin*, *112*, 160-164.

Rosenthal, R., & Rosnow, R. L. (1991). Essentials of behavioral research: Methods and data analysis. New York: McGraw-Hill.

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.

Schlenker, B. R., & Bonoma, T. V. (1978). "Fun and games": The validity of games for the study of conflict. *Journal of Conflict Resolution*, *22*, 7-38.

Schwab, D. P. (2005). *Research methods in organizational research* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum.

Shapiro, M. A. (2002). Generalizability in communication research. *Human Communication Research*, *28*, 491-500.

Slattery, J. P., & Ganster, D. C. (2002). Determinants of risk taking in a dynamic uncertain context. *Journal of Management*, *28*, 89-106.

Stone-Romero, E. F. (2002). The relative validity and usefulness of various empirical research designs (77-98). In S. G. Rogelberg (Ed.), *Handbook of research methods in industrial and organizational psychology*. Malden, MA: Blackwell.

Vancouver, J. B. (2005). The depth of history and explanation as benefit and bane for psychological control theories. *Journal of Applied Psychology*, *90*, 38-52.

Weick, K. E. (1965). Laboratory experimentation with organizations. In J. G. March (Ed.), *Handbook of organizations* (pp. 194-260). Chicago: Rand McNally.

Wells, G. L., & Windschitl, P. D. (1999). Stimulus sampling and social psychological experimentation. *Personality and Social Psychology Bulletin*, *25*, 1115-1125.

Scott Highhouse is a professor in the Department of Psychology at Bowling Green State University. His research interests include judgment and decision making in the workplace, employee recruitment and selection, corporate reputation, and the history of applied psychology.