Journal of Applied Research in Memory and Cognition 7 (2018) 11–15

Contents lists available at ScienceDirect

Journal of Applied Research in Memory and Cognition

journal homepage: www.elsevier.com/locate/jarmac

Commentary

Connecting Laboratory and Field Research in Judgment and Decision Making: Causality and the Breadth of External Validity

Daniel M. Bartels*, Reid Hastie, Oleg Urminsky

University of Chicago Booth School of Business, United States

Markman (2018) provides a thought-provoking perspective on the relationships between lab and field research, using his own research program on regulatory fit as an instructive example (e.g., Otto, Markman, Gureckis, & Love, 2010; Worthy, Maddox, & Markman, 2007). He discusses characteristics of typical lab and field studies, identifies some examples of productive relationships, and analyzes the failure to communicate between lab and field researchers, focusing on the relationship between academic cognitive psychology research programs and naturalistic decision making programs (e.g., Lipshitz, Klein, Orasanu, & Salas, 2001; abbreviated as NDM hereafter). He also provides useful advice about how to facilitate interactions between researchers of each type.

We elaborate on Markman's advice on how to promote lab–field interactions in research by exploring distinctions in the goals and benefits of different kinds of field research and noting some additional ways of effectively relating lab and field settings. We add two important distinctions to Markman's conceptual analysis: narrow versus broad external validity and descriptive-observational versus causal-experimental field studies.

Narrow Versus Broad External Validity

Markman refers to a tradeoff between internal and external validity (Campbell & Stanley, 1963; Cook & Campbell, 1979). As we interpret it, internal validity refers to the validity of causal claims specifically in the setting where they are discovered. Markman indicates that external validity refers to the usefulness of a description or theoretical construct to elucidate and provide control over behavioral phenomena in a specific naturally occurring situation. NDM studies of firefighters, military personnel, or airplane crews *in situ* exhibit high external validity because their conclusions are valid in one specific naturally occurring, non-laboratory setting.

Markman "types" NDM studies as high in external validity and low in internal validity and implies that much of laboratory-based judgment and decision-making research is low in external validity but high in internal validity. We think taking a broader view of research that engages with the field may lead to different conclusions about the nature and inevitability of tradeoffs between lab and field.

There is an important distinction between narrow and broad external validity: the generalizability of a conclusion or finding only to one specific naturally-occurring target situation versus generalizability to many situations beyond the one where the finding was originally discovered. We would conjecture that much NDM research has aimed to answer questions about decisions in one specific non-laboratory setting like firefighting incidents, commercial aircraft cockpits, the bridge of a Navy cruiser, or the radar room on a Navy destroyer. The motivation and funding for much of this research was aimed to improve individual and team decision processes in these specific settings.

This type of narrow external validity is an appropriate objective for many applied, domain-specific research programs. Whether these kinds of findings generalize will depend on how similar the factors related to the causal processes are in the original study setting and the setting that is the target of generalization. By conducting studies in the settings and with the actors to which they wanted to draw conclusions, NDM researchers followed the optimal strategy to produce conclusions with narrow external validity.

Author Note

* Correspondence concerning this article should be addressed to Daniel M. Bartels, University of Chicago Booth School of Business, United States. Contact: bartels@uchicago.edu

In contrast, most scientific laboratory research is motivated to discover more basic—and, for this reason, potentially more broadly valid—principles of human nature. This kind of research aims for broad external validity. The vehicle for generalization is not similarity between the situation where the research is conducted and the target setting. Rather, generalization will be supported if an identified causal mechanism represents a general characteristic of human nature that extends to settings beyond the original studies, and therefore can be broadly applied to understand and control behavior (Mook, 1983; Pearl & Bareinboim, 2014). If that scientific goal is achieved, it will be possible to generalize the theoretical conclusions to many different settings. It is less likely that research conducted in only one naturally occurring setting will achieve high levels of this type of broad external validity.

Many major successes in the scientific enterprise of inducing general causal principles for decision behaviors have been produced by using a variety of controlled tasks as the base of the analysis. Some prime examples would be Kahneman and Tversky's (1979) prospect theory (see also Tversky & Kahneman, 1992); Anderson's (1996) information integration theory; Payne, Bettman, and Johnson's (1993) adaptive decision maker; Brunswik and Hammond's lens model (Hammond & Stewart, 2001); Edwards's intuitive statistician (Weiss & Weiss, 2009); Gigerenzer's (2000) simple adaptive heuristics; and many others.

We also believe that broad external validity will be promoted by research in field settings, but that general conclusions will require studies that span a variety of relevant, naturally occurring domains. Determining whether prospect theory captures broadly generalizable causal mechanisms, for example, required testing its predictions in a broad range of naturally occurring (or laboratory) contexts, such as goal striving (Allen, Dechow, Pope, & Wu, 2016), consumer purchasing (Bell & Lattin, 2000), financial trading (Haigh & List, 2005), taxi driving (Camerer, Babcock, Loewenstein, & Thaler, 1997; Thakral & Tô, 2017), and gambling (Camerer, 2000).

Descriptive Versus Causal Research in the Field

A second key distinction is between research in the field that is primarily descriptive and based on observations versus research that attempts to identify causal relationships in the field based on interventions or complex statistical modeling (e.g., Angrist, Imbens, & Rubin, 1996). Most of the studies conducted in the NDM research tradition have been aimed at description rather than at directly testing causal mechanisms.

As Markman points out, descriptive research "encourages researchers to think about global theories of how choices are made that take into account context, expertise, time pressure, and team performance," and even descriptive findings can call into question simplifying assumptions often made in the lab. For example, Markman cites important findings (fire-fighting, military tactics) where the traditional one-shot, fixed choice-set decision conceptual framework, developed to illuminate laboratory-based research (as well as many everyday examples in medical, legal, and consumer decision making), does not seem to apply. We agree that these are instructive cases in which research in the field can inform conclusions about limitations on the scope of theories and models developed for decisions in different settings. We are a bit less optimistic than Markman, however, about the power of NDM approaches to consistently generate overarching, integrative theories, given their practical focus on specific applications.

A key point here, that we reiterate from Markman's target article, is that NDM research, as well as immersive observational methods popular in sociology and anthropology, are primarily descriptive and are therefore unlikely to produce strong conclusions about causal relationships. (Some critics endorse a more extreme position, that descriptive methods lack methodological power to resolve theoretical conflicts and to falsify precise causal hypotheses.) So, many—but not all—field research programs entail a tradeoff between internal and external validity. Some approaches to field research test the causal effects of real-world interventions, which can provide both internal and external validity.

There is a long tradition of real-world research, rooted in medical research, developmental economics, and social psychology, that relies on "field experiments" or "randomized controlled trials" (Campbell, 1991; Harrison & List, 2004). Field experiments can identify causal effects of an intervention, testing the causal predictions of theories, whether originally based on lab or field research. Field experiments often lack the degree of control and opportunities for process measurement achievable in the lab, limiting the ability of the researcher to dissect the multiple psychological processes that may mediate or moderate an outcome of interest.

When predictions are not confirmed in the field, it can be difficult to zero-in on the specific theoretical claim than should be revised. When faced with lack of confirmation, a researcher whose theory was not confirmed will often note that theoretical preconditions for the hypothesis may not have been instantiated in the less controlled field setting (e.g., maybe people were not attentive to the theoretically significant information, or other incentives conflicted with performing the theoretically relevant task).

Nevertheless, field experiments can be extremely useful for questioning simplifying assumptions in the lab and highlighting the need for more global theories that take the full set of relevant factors into account to make externally valid, broadly generalizable predictions. Field experiments are often ideal for identifying the causal effect of a change (Benartzi et al., 2017) under real-world conditions, and therefore can be very powerful for testing theories that do make strong, falsifiable predictions.

To pick one historical example, in the 1850s when the physician John Snow hypothesized that water contamination could cause cholera, he did not bemoan the lack of internal validity in the field and the lack of external validity in the lab. Instead, he removed the handle of the Broad Street water pump in London and ended the local cholera epidemic, providing compelling experimental evidence for a causal link. Snow's experiment was not definitive in terms of the causal mechanism (cholera was identified in the lab 35 years later), but this crude field experiment provided high levels of both internal and external validity.

Furthermore, it contributed to the later research that identified water-borne bacteria as the causal mechanism producing outbreaks of cholera. (For another early example, see how citrus fruits were tested against other treatments for scurvy on an 18th century British Royal Navy ship [White, 2016][1] or, for a more sophisticated behavioral example, see the extensive multilaboratory research program on prejudice in social judgments conducted in the field [Bertrand & Duflo, 2012].)

Tradeoffs Between Internal and External Validity

It is a mistake to assume that internal validity and external validity are necessarily in conflict and must trade off in individual empirical projects. Many research projects that implement experiments in field settings sidestep the tradeoff between internal and external validity. We suggest, proceeding from the initial premise, that the most conclusive scientific studies for testing causal predictions, in terms of both internal and external validity, would be well-controlled experiments conducted in the relevant field settings (Levitt & List, 2007) with measurement of major mediating and moderating factors. For broad external validity, it would be necessary to conduct a series of such field experiments in a representative sample of relevant field settings, as findings from the same field experiment can vary substantially across settings (Allcott, 2015).

There are, of course, practical constraints that often prevent behavioral researchers from achieving this ideal. The most important obstacles are logistical costs. Field experiments consume more time, money, and labor than lab studies exploring the same questions. Researchers often lack access to an appropriate field setting, particularly when the institutional decision-makers whose approval is needed see little benefits to allowing field research or are worried about challenges to conventional practices. Field research that may reveal trade secrets, poor management, or unethical behavior is likely to be blocked by those who control access. Ethical imperatives may also prevent conducting field experiments, particularly when the question involves vulnerable populations.

Another key set of limitations that can reduce the value of field experiments is the lack of control over the research environment in many non-laboratory settings. Accurate measurement can be difficult to achieve in the field, and, in the lab, the researcher can "control out" many background factors that would confound precise causal claims. The lab researcher can ensure that participants are paying attention, are not interacting with each other, have the desired level of experience with the task, and so on.

In field studies, when confounding factors cannot be carefully controlled or are deliberately preserved to maximize narrow external validity, these factors may reduce internal validity by providing alternate causal interpretations of results. This means that researchers seeking broad external validity are often well-served by studying abstracted, controlled tasks in the lab. Those are the tasks that maximize the chances of reaching a more general conclusion about causal mechanisms. All of these factors impose the typical, but not necessary, practical tradeoff between internal and external validity.

Finding a Way Forward That Integrates Lab and Field

Based on his view of the relationship between laboratory-based judgment and decision making and research in the NDM tradition, Markman makes three prescriptions for future research: (a) use computational models like those in the ACT and GOMS families to help generalize from lab results to the field; (b) design more laboratory tasks to simulate naturally occurring tasks; (c) "encourage more discussion across researchers from the laboratory research and naturalistic decision making communities." We agree with these ideas, and we would like to suggest a broader agenda for bridging lab and field.

Initial progress on scientific projects seems to occur most often in the behavioral sciences when studies are conducted in artificial, highly controlled tasks. Well-designed laboratory tasks can provide precise manipulations of the most important causal variables, providing strong tests of causal claims about these abstract variables. A resulting understanding of fundamental causal mechanisms can then potentially achieve broad external validity. For example, Kahneman and Tversky's claim that "losses loom larger than gains" was originally examined in laboratory tasks involving artificial gambles, but the claim generalized to many other artificial and natural decision situations. Payne et al.'s (1993) model of strategic flexibility and adaptive selection of choice strategies as a function of decision importance has been applied across dozens of naturally occurring situations. The effects of working memory capacity, discovered in many studies of highly artificial stimuli in the lab, are extremely general across decision tasks. Again, the basis of generality is not similarity on causally relevant factors between two domains, but rather the discovery of a more basic causal mechanism that is general across many tasks and actors.

Starting with a good descriptive understanding of real-world settings is essential for designing tasks aimed at discovering basic and generalizable causal mechanisms in the lab. In our view, NDM research is one of many viable approaches to shaping laboratory tasks that map onto real-world phenomena. Expert guidance, other descriptive research traditions (see, e.g., Beller, Bender, & Medin, 2012) and even "flawed" field experiments can also shape significant lab research. In particular, field experiments with inherent confounds or measurement limitations, whether reporting positive or null findings, can and should give rise to more precise lab studies that investigate the psychological mechanism underlying an observed field outcome. For example, the failure of intrinsic-motivation theories to generalize to the field motivated lab-based reassessment and revision of those theories (Goswami & Urminsky, 2017).

We also endorse Markman's point that a process interpretation of the laboratory phenomena, ideally in the form of a computational model, is helpful for making the transfer from a laboratory result to application in a specific field setting. As noted by Markman, a good formal model identifies the most important causal factors discovered in the laboratory setting and

[1] This cure also reportedly brought about the gimlet, which is a nice bonus. See https://kindredcocktails.com/review/the-gimlet.

provides suggestions for how to measure empirical variables and estimate important parameters. Process models also make unequivocal statements about how the causal variables are integrated to produce the outcomes of interest, providing specificity and falsifiability.

In the case of dynamic process models, there is also a commitment to the ordering of underlying causal events, which can facilitate model application and testing. Because of the enormous adaptive flexibility of human behavior, the approach exemplified by ACT, SOAR, and GOMS applications seems the most effective: a general overarching framework, essentially a programming language for human cognition, within which task-specific models can be defined.

Conclusions

To summarize, we recommend paying attention to the distinction between narrow and broad external validity, keeping in mind that methodology will depend on which is the primary objective. Narrow external validity is promoted by studies in the target setting of interest and high-fidelity laboratory simulations. Internal validity is less important when narrow external validity is the goal. Broad external validity is promoted by studies in multiple settings, both controlled and naturally occurring, and by maximizing internal validity to identify fundamental causal mechanisms.

External and internal validity are not necessarily in conflict. Field experiments represent one exemplar of a methodology that can achieve high levels of both, but tend to be more expensive, require cooperation of authorities, and make controlled and precise measurement more challenging. In our view, the ideal research program involves comprehensive and careful descriptive studies of a naturally occurring phenomenon, shifts to the creation of controlled laboratory tasks that capture the essential causal mechanisms hypothesized to underlie the phenomenon (summarized by a mathematical model), and field experiments to verify causal conclusions from the laboratory studies, which may in turn raise new questions.

Thus, we see the ideal relationship between lab and field as bidirectional. It seems indubitable that research in a high-fidelity controlled simulation task will yield the fastest insights at a reasonable cost into behavior in one specific naturally occurring decision process. But, there is often a tradeoff with the loss of scientific generality. In many, perhaps most, cases in the behavioral sciences, scientific progress seems greater when we first start with phenomena and theoretical hypotheses from the lab and then test them in field settings. This can occur on a descriptive level, when theories are assessed for completeness and plausibility by contrasting assumptions and predictions with observed facts. And, significant progress can be made on scientific projects when causal hypotheses are tested with experiments and structural models in non-laboratory settings.

Hypotheses about human behavior hatched and raised in the lab often languish in the field. In some research traditions, these outcomes are brushed aside, with unscientific fatalism, viewing behavior in the field as simply inexplicable and too complex to predict. But, in agreement with Markman, we believe that "failures" in the field provide guidance to fruitful directions for future research. When a robot fails to perform as predicted and crashes into walls, researchers set aside their plans for developing additional capabilities and refocus their efforts on identifying the current bugs in the robot's program. When humans fail to behave as psychological theories predict, it is equally important to set aside our plans for building on those theories, and instead to go back to the lab to identify the flaws in the theories, even starting from scratch if need be.

Author Contribution

All authors contributed equally.

Conflict of Interest Statement

The authors declare no conflict of interest.

Keywords: Field studies, Lab studies, External validity, Internal validity, Causality, Judgment, Decision making

References

Allcott, H. (2015). Site selection bias in program evaluation. *The Quarterly Journal of Economics*, *130*, 1117–1165.

Anderson, N. H. (1996). *A functional theory of cognition* (pp. 1–360). Hillsdale, NJ: L. Erlbaum Associates.

Angrist, J. D., Imbens, G. W., & Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, *91*, 444–455.

Allen, E. J., Dechow, P. M., Pope, D. G., & Wu, G. (2016). Reference-dependent preferences: Evidence from marathon runners. *Management Science*, *63*, 1657–1672.

Bell, D. R., & Lattin, J. M. (2000). Looking for loss aversion in scanner panel data: The confounding effect of price response heterogeneity. *Marketing Science*, *19*, 185–200.

Beller, S., Bender, A., & Medin, D. L. (2012). Should anthropology be part of cognitive science? *Topics in Cognitive Science*, *4*, 342–353.

Benartzi, S., Beshears, J., Milkman, K. L., Sunstein, C. R., Thaler, R. H., Shankar, M., et al. (2017). Should governments invest more in nudging? *Psychological Science*, *28*, 1041–1055.

Camerer, C. F. (2000). Prospect theory in the wild: Evidence from the field. In D. Kahneman, & A. Tversky (Eds.), *Choices, values, and frames* (pp. 288–300). Cambridge: Cambridge University Press (Chapter 16)

Bertrand, M., & Duflo, E. (2012). Field experiments on discrimination. In A. V. Banerjee, & E. Duflo (Eds.), *Handbook of economic field experiments* (pp. 309–393). Amsterdam: North Holland, Elsevier (Chapter 8).

Camerer, C., Babcock, L., Loewenstein, G., & Thaler, R. (1997). Labor supply of New York City cabdrivers: One day at a time. *The Quarterly Journal of Economics*, *112*, 407–441.

Campbell, D. T. (1991). Methods for the experimenting society. *American Journal of Evaluation*, *12*, 223–260.

Campbell, D. T., & Stanley, J. C. (1963). *Experimental and quasi-experimental designs for research* (pp. 1–84). Chicago, IL: Wadsworth.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis issues for field settings* (pp. 1–405). Boston, MA: Houghton Mifflin Harcourt.

Gigerenzer, G. (2000). *Adaptive thinking: Rationality in the real world* (pp. 1–360). New York, NY: Oxford University Press.

Goswami, I., & Urminsky, O. (2017). The dynamic effect of incentives on postreward task engagement. *Journal of Experimental Psychology: General*, *146*(1), 1–19.

Haigh, M. S., & List, J. A. (2005). Do professional traders exhibit myopic loss aversion? An experimental analysis. *The Journal of Finance*, *60*, 523–534.

Hammond, K. R., & Stewart, T. R. (2001). *The essential Brunswik: Beginnings, explications, applications* (pp. 1–560). Oxford: Oxford University Press.

Harrison, G. W., & List, J. A. (2004). Field experiments. *Journal of Economic Literature*, *42*, 1009–1055.

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, *47*, 263–291.

Levitt, S. D., & List, J. A. (2007). What do laboratory experiments measuring social preferences reveal about the real world? *The Journal of Economic Perspectives*, *21*, 153–174.

Lipshitz, R., Klein, G., Orasanu, J., & Salas, E. (2001). Taking stock of naturalistic decision making. *Journal of Behavioral Decision Making*, *14*, 331–352.

Markman, A. B. (2018). Combining the strengths of naturalistic and laboratory decision-making research to create integrative theories of choice. *Journal of Applied Research in Memory and Cognition*, http://dx.doi.org/10.1016/j.jarmac.2017.11.005

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

Otto, A. R., Markman, A. B., Gureckis, T. M., & Love, B. C. (2010). Regulatory fit and systematic exploration in a dynamic decision-making environment. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*, 797–804.

Payne, J. W., Bettman, J. R., & Johnson, E. J. (1993). *The adaptive decision maker* (pp. 1–330). Cambridge: Cambridge University Press.

Pearl, J., & Bareinboim, E. (2014). External validity: From do-calculus to transportability across populations. *Statistical Science*, *29*, 579–595.

Thakral, N., & Tô, L. T. (2017). *Daily labor supply and adaptive reference points. Working paper*. Cambridge, MA: Harvard University.

Tversky, A., & Kahneman, D. (1992). Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty*, *5*, 297–323.

Weiss, J. W., & Weiss, D. J. (Eds.). (2009). *A science of decision making: The legacy of Ward Edwards* (pp. 1–511). New York, NY: Oxford University Press.

White, M. (2016). James Lind: The man who helped to cure scurvy with lemons. *BBC News*,. Retrieved from http://www.bbc.com/news/uk-england-37320399

Worthy, D. A., Maddox, W. T., & Markman, A. B. (2007). Regulatory fit effects in a choice task. *Psychonomic Bulletin and Review*, *14*, 1125–1132.

Received 31 December 2017;
accepted 3 January 2018