

## **Internal and External Validity in Economics Research:**

### **Tradeoffs between Experiments, Field Experiments, Natural Experiments and Field Data\***

**Brian E. Roe and David R. Just**

January 2009

Accepted to the 2009 Proceedings Issue, *American Journal of Agricultural Economics*

*Abstract:* In the realm of empirical research, investigators are first and foremost concerned with the validity of their results, but validity is a multi-dimensional ideal. In this article we discuss two key dimensions of validity – *internal* and *external* validity – and underscore the natural tension that arises in choosing a research approach to maximize both types of validity. We propose that the most common approaches to empirical research – the use of naturally-occurring field/market data and the use of laboratory experiments – fall on the ends of a spectrum of research approaches, and that the interior of this spectrum includes intermediary approaches such as field experiments and natural experiments. Furthermore, we argue that choosing between lab experiments and field data usually requires a tradeoff between the pursuit of internal and external validity. Movements toward the interior of the spectrum can often ease the tension between internal and external validity but are also accompanied by other important limitations, such as less control over subject matter or topic areas and a reduced ability for others to replicate research. Finally, we highlight recent attempts to modify and mix research approaches in a way that eases the natural conflict between internal and external validity and discuss if employing multiple methods leads to economies of scope in research costs.

---

\*Roe is professor in the Department of Agricultural, Environmental and Development Economics, Ohio State University. Just is an associate professor in the Department of Applied Economics and Management, Cornell University. Roe acknowledges support of the Ohio Agricultural Research and Development Center.

This article was presented in an invited paper session at the 2009 ASSA annual meeting in San Francisco, CA. The articles in these sessions are not subjected to the journal's standard refereeing process. The authors thank Stephan Marette, Ernst Fehr and Rachel Croson for helpful suggestions.

Economics is about tradeoffs, and our approach to designing and executing economics research is not immune to this core economic principle. In the realm of empirical research, investigators are first and foremost concerned with the validity of their results, but validity is a multi-dimensional ideal. In this article we discuss two key dimensions of validity – *internal* and *external* validity – and underscore the natural tension that arises in choosing a research approach to maximize both types of validity. We propose that the most common approaches to empirical research – the use of naturally-occurring field/market data and the use of laboratory experiments – fall on the ends of a spectrum of research approaches, and that the interior of this spectrum includes intermediary approaches such as field experiments and natural experiments. Furthermore, we argue that choosing between lab experiments and field data usually requires a tradeoff between the pursuit of internal and external validity. Movements toward the interior of the spectrum can often ease the tension between internal and external validity but are also accompanied by other important limitations, such as less control over subject matter or topic areas and a reduced ability for others to replicate research. Finally, we highlight recent attempts to modify and mix research approaches in a way that eases the natural conflict between internal and external validity and discuss if employing multiple methods leads to economies of scope in research costs.

## **Validity**

Validity within empirical economics is generally concerned with whether a particular conclusion or inference represents a good approximation to the true conclusion or inference (i.e. whether our methods of research and subsequent observations provide an adequate reflection of the truth). (See Trochim 2008 for a more general discussion of validity in social science research.) Validity

is not a single-dimensional effort but rather requires integrated effort on several fronts to develop conclusions that may be defended as valid. We focus on two dimensions: internal and external validity.<sup>1</sup>

We define internal validity as the ability of a researcher to argue that observed correlations are causal. The credo of ‘correlation does not imply causation’ continues to dominate the discussion and critique of empirical work based on naturally-occurring market and field data. Econometric identification issues often dominate researchers’ efforts in the empirical analysis of uncontrolled field data because internal validity must be established.<sup>2</sup>

We define external validity as the ability to generalize the relationships found in a study to other persons, times and settings. Even if internal validity is convincingly established, the general validity of a result or conclusion must be determined by the context in which the result will be applied. For instance proving that a behavioral result in one market is causal may require a different approach than estimating the size of that behavioral effect in another market.

A related but distinct concept is ecological validity. We will say that a study has ecological validity to the extent that the context in which subjects cast decisions is similar to the context of interest. If the researcher is conducting the study in the setting of direct interest and is able to conduct the study with minimal disturbance to the contextual ecology of that setting, then internally valid references will be meaningful for that setting. However, if the researcher must radically disturb or alter the ecological context in order to establish internal validity, the

---

<sup>1</sup> Construct and conclusion validity are also critical for establishing general validity, but are de-emphasized in the remainder of this paper due to space constraints. Conclusion validity concerns whether there is a robust empirical correlation between the measured variables (e.g. are household income and consumption correlated?). Construct validity deals with the appropriateness of a particular empirical measurement in reflecting its theoretical counterpart (e.g. is household income the best measure of income for the theory being tested?). It is our conjecture that the additional control provided laboratory researchers can provide greater construct validity (thanks to Rachel Croson for first suggesting this insight during audience discussion of this work).

<sup>2</sup> For example Just and Just (in press) note a tradeoff between validity and parameter identification. Put simply, the need to identify parameters of a model can limit one’s ability to discern which of several models is valid.

relevance of inferences for that setting must still be established. Regardless of the degree of ecological validity established by the researcher, external validity still requires that the results be generalizable to other meaningful contexts. Intuition suggests that studies with high ecological validity have already established meaning in at least one relevant context and have a lower burden of proof for establishing external validity than studies with low ecological validity.

### **The Research Methodology Spectrum**

One aim of this paper is to catalog the major economic research design approaches and articulate the strengths and weakness of each approach with a particular emphasis on internal and external validity. Here, we introduce and define a spectrum of approaches to generating data used in economic research, including data collected from laboratory experiments, field experiments, natural experiments, and naturally-occurring field and market settings.

This spectrum is defined by the degree of verifiable exogenous variation within the economic context that produces the data. At one end of the spectrum are laboratory experiments, in which the researcher purposefully imposes one or more exogenous changes (i.e. treatments) on a randomly chosen subset of subjects while holding all other elements of the context identical for a control group with the express purpose of establishing a causal relationship between the altered elements and observed outcomes. At the other end are naturally-occurring or uncontrolled data in which the researcher observes market or field behavior that transpires regardless of the researcher's existence and largely independent of the researcher's activity.

From these disparate ends, we consider two interior points. One is the field experiment, which involves researcher manipulation of a naturally-occurring context to induce relevant exogenous variation (see Harrison and List 2004 for a recent review and Herberich, Levitt and

List [this issue] for a historical perspective). Compared to laboratory experiments, in which the researcher has control over nearly all aspects of the economic and institutional context, field experiments allow for less researcher control because much of the context is independent of the researcher's effort.<sup>3</sup>

Another intermediate point is the natural experiment (see reviews by Rosenzweig and Wolpin 2000; Meyer 1995). As with field data, a researcher cannot manipulate the stimulus or influence the data generation process. Rather, the researcher takes advantage of a change in context or setting that occurs for some subjects due to natural causes or social changes beyond the researcher's and subjects' influence. A natural experiment generates a control group (i.e. a group of similar individuals who avoid the natural treatment) who can be compared to the treatment group.

### **Tradeoffs along the Spectrum**

Comparing ends of the methodology spectrum, there is a clear tradeoff between internal and external validity: laboratory experiments provide greater internal validity than field data, while field data provide greater ecological validity and, we argue, a lower burden for establishing external validity. To understand these rank orderings, we turn to common threats to both internal and external validity and discuss how each method addresses these threats (see figure 1 for a summary).

Threats to internal validity are several:

---

<sup>3</sup> We note that there may be several intermediate steps between a laboratory and a field experiment, such as a laboratory experiment conducted with subjects recruited from the field population of interest. For the purposes of discussion, we will focus on existing economic institutions and settings in which the researcher has manipulated some facet for a randomly assigned group of individuals (e.g. altering the information provided or sales mechanism within randomly chosen stores).

*Lack of temporal clarity.* Field data – particularly cross-sectional data – may not record the timing of exposure and potential response, which is often a key element to establishing causality. In the case of field data, it may be difficult to discern the direction causality runs due to coarseness in the data collection time frame or due to recall bias during interviews of participants. Adept lab researchers, on the other hand, control the structure of the experiment such that the timing of stimulus and response are clear and data collection is not reliant upon subject memory.

*Systematic differences in treatment groups.* In order to isolate the impact of a stimulus, a researcher must be able to argue that the treatment and control groups are otherwise identical. If these groups differ in other preconditions, these other factors will confound inference concerning the treatment. Data describing the characteristics of treatment and control groups in field settings may not be rich enough to rule out such systematic differences, while in lab work, researchers can randomly assign treatment to participating subjects to rule out systematic differences or even attempt to match subjects with respect to key characteristics such as gender, age, or race prior to randomly assigning one pair member to the treatment condition.

*Concurrent third elements that confound the outcome.* The presence of uncontrolled variation in unobserved variables diminishes the field researcher's ability to isolate causal effects. Any correlation may be purely spurious, as the third element may affect both the hypothesized stimulus and the hypothesized response. Lab researchers' command over the subjects' context and setting can limit the chance that the other elements systematically vary between control and treatment groups.

*Maturation over time/structural change.* The relationship between the stimulus and response may change over time (e.g., due to learning or physical changes among subjects).

Analysis of field data may be unable to identify such effects independently of other correlations, which further retards causal identification. Laboratory experiments are normally of limited time, which minimizes some types of learning effects, and feature control groups, whose learning can be independently monitored. Furthermore, lab experiments have structural elements that are either purposefully maintained or manipulated by the researcher, which either minimizes such effects or allows for structured analysis.

External validity also suffers from several threats:

*Potential interactions with elements and context not found in the study's setting.* Just as confounding elements can destroy the ability to isolate a causal relationship, applying a result inferred under controlled conditions to a context involving uncontrolled conditions may destroy the ability of the researcher to predict the outcome. If relationships identified in one field setting have been determined to be internally valid, the likelihood that the inference will be relevant in another ecologically valid setting may increase. Lab experiments, on the other hand, often are devoid of many contextual elements found in organic institutions/markets; hence, there is a larger likelihood that these additional and different contextual elements could limit the applicability of results isolated in the lab.

*Limited variation within stimulus or response.* If the study is conducted under circumstances that expose participants to only limited levels of the stimulus, the relationship that is found may fail to represent what would happen without such restrictions. For example laboratory experiments often have limited latitude to research individual losses, high rewards or extended time horizons. Similarly, restrictions on the response to the stimulus may lead to a special relationship that would not otherwise exist. Field data often allow for a greater variation and for variation in relevant domains (e.g. losses, large rewards, longer time horizons).

*Systematic differences between the groups for which the result will be applied.* If the relationship is inferred on a narrow subset of the population, the relationship may not hold in more diverse samples. Laboratory experiments often recruit from convenience samples such as students and must comply with ethical standards such as informed consent, which may further exacerbate differences between laboratory subjects and the populations of interest. For example Roe et al. (in press) identify that participants in neuroeconomics experiments are less risk averse than recruits who reject participation.

### **Relaxing the Internal-External Validity Tradeoff with Intermediate Approaches**

Field and natural experiments have gained wider popularity in recent decades because these methods relax the inherent tension between uncontrolled field and controlled laboratory data collection approaches. For example compared to laboratory experiments, field experiments are situated in institutions and contexts that arise independent of the researcher's intervention. While researcher involvement necessarily destroys some of this setting's ecology by imposing a manipulation that affects a treatment group, substantial context remains, which addresses several of the threats to external validity that plague laboratory experiments.

Natural experiments, on the other hand, leverage events outside the researcher's and subjects' control to address several threats to internal validity, such as minimizing the chance of confounding elements and self-selection into treatment groups, while sacrificing few of the features of field data that enhance external validity, such as wider, more natural ranges of treatment effects and the presence of organically-formed context.

## **Drawbacks of Intermediate Approaches**

Field and natural experiments, while potentially relaxing the inherent tension between internal and external validity, face other limitations. One key limitation is that the topic areas and subject matter available to the researcher are often more limited than that faced by the laboratory and field data researcher.

The topics and subject matter attacked via natural experiments are limited to what natural or external forces provide the researcher. While state and county-level differences in policies provide a rich palette of topics for natural experiments, researchers must remain vigilant by ruling out systematic differences at these levels of government (e.g. rule out that subtle differences in unobserved variables between states did not lead to the difference in implemented policies).

Field experiments can only explore topic areas and subjects in contexts where the researcher can exert enough influence to manipulate the existing context. This may require the researcher to have a personal or professional affiliation with the markets or institutions of interest and may limit research to settings with certain self-selected characteristics that open the entity to manipulation by a researcher. Furthermore, the entity that allows for such manipulation may restrict the types of interventions or range of stimulus in a way that is similar to or greater than the limitations observed in laboratory experiments. For example in a lab experiment concerning price response, the researcher can freely double prices, while the manager of a chain of stores that accommodates a researcher's field experiment might only allow for a 10% increase in a key price in treatment stores.

Tied to the limitation of subject matter is the issue of replicability. A key strength of laboratory work is that, if properly described, any researcher with access to a laboratory facility

can attempt to replicate previous results. In the case of field experiments, access to the field setting is often limited to the one researcher who invested in the relationships necessary to enact the original research, which makes replication difficult if not impossible. Similar designs could be implemented by other researchers in other contexts available to them, but necessarily, the context will shift, and any differences in results could be attributed to these differences. Natural experiments also retard replication, as the events used in such studies are often unique. Often replication for a natural experiment is limited to analyses of alternate data from the same event.

### **Multiple Approaches**

While any single research approach has drawbacks, a researcher's application of several approaches to study the same phenomenon may provide an alternative way to reduce tensions between internal and external validity. For example List (2006) uses coupled laboratory and field experiments to investigate gift exchange behavior. Execution of a multi-modal research design has obvious benefits in that the second mode of research can counteract the weakness of first. For example List (2006) purposefully attempts to improve the external validity of his research by creating a manipulation in a pre-existing marketplace that was similar to the manipulation used in the lab experiment portion of the study. By involving two contexts in the grand research scheme – one fully created in the lab and one organically generated in the sports-card market and minimally altered by the research design -- List (2006) examined the robustness of gift exchange to differing contexts.

Few have employed field data collection and experimental methods in concert. However, one can see the potential importance of such an approach. Consider testing the hypothesis that risk aversion decreases with wealth. In this case it is neither possible to randomize over the

stimulus (wealth) nor cleanly observe the response (risk aversion) in the field. By using a combination of econometric techniques (to overcome selection on wealth) and experimental techniques (to overcome identification), one may establish a causal link and potentially show importance in a field context.

Several practical considerations are warranted when analyzing the efficacy of multi-modal research approaches. The first is the question of economies of scope in research: is the cost of executing multiple modes of research less when one researcher is conducting all modes or if each were executed by individual (uncoordinated) researchers? Logically, coordination costs must increase across modes, as the design of one mode must align with other modes. In any single-mode application the researcher must pay attention to construct validity (i.e. whether the constructs developed and recorded in the experimental or field setting are properly related to the underlying theoretical concept). With multiple modes, however, both empirical constructs must now align with one another as well as with the underlying theory, and this must entail some additional cost. Furthermore, multi-modal research is more likely to require a team approach that features researchers that are experts in each mode. This increases budgets accordingly and requires additional coordination of human resources. On the other hand the fixed costs of developing relevant theory are spread over several research modes.

Another practical consideration related to multi-modal research efforts is that of total cost. Even if economies of scope obtain and the cost of executing multiple modes of research is lower, in the end, there may be certain budget restraints that limit the number of modes possible. In a related issue funding agencies may have uncertainty regarding a research team's ability to execute multiple modes of research and limit budget allocations to only a single mode or may

only fund additional modes contingent on the successful execution of the first mode, which could frustrate attempts at obtaining economies of scope by requiring sequential execution.

## **Conclusions**

Economics has shifted from a field dominated by the use of uncontrolled field and market data to one in which many modes of research are used to test theory and inform policy. The popularization of laboratory experiments, natural experiments, and field experiments in economics has each generated its set of advantages and disadvantages relative to the traditional reliance upon field and market data; past reviews of each of these burgeoning areas often stress such advantages and remark on some disadvantages. In this article we take a more systematic view across this spectrum of research approaches and discuss the tradeoffs a researcher faces when choosing from this menu of research modes.

We note that the ends of the spectrum – uncontrolled field data and highly controlled laboratory experiments – yield a stark tradeoff between internal and external validity, where laboratory experiments, by the nature of researcher control, allow for more robust establishment of causality at the expense of naturally-occurring context and more limited subject populations which diminish claims to external validity. Both field and natural experiments can relax the tension between internal and external validity. In the case of field experiments, a well-designed manipulation in a naturally-occurring context can provide stronger claims to causality with limited effects on context, while in the case of natural experiments, context is left intact with the benefit of nature or other external forces allowing for greater claims of causality. However, these approaches introduce other limitations, such as a limitation of topic and subject matter focus to those areas allowed by nature or by a researcher's connections and persuasion. These

limits reduce the chance for replication, which is a key strength of laboratory work and to a lesser extent of field data.

We also note that multi-modal approaches can ease the tensions between internal and external validity. We discuss several issues that arise with multi-modal implementation, such as whether such approaches lead to economies of scope in research and whether such approaches suffer from coordination issues.

Like nearly any topic considered by economists, we find that our decisions concerning research design in economics offer no free lunches – no single approach universally solves problems of general validity without imposing other limitations. However, as researchers become more aware of and comfortable with the full menu of research approaches available, efficiencies can be identified by individual researchers on a case-by-case basis and forward the goal of cost-effective, valid economic research.

## References

- Harrison, G.W., and J.A. List. 2004. "Field Experiments." *Journal of Economic Literature* 42:1009–1055.
- Herberrich, D., S.D. Levitt, and J.A. List. 2009 "Title?" *Journal* vol:iss, pp. In this Issue.
- Just R.E., and D.R. Just. "Global Identification and Tractable Specification Possibilities for Risk Preference Estimation." *Journal of Econometrics* (special issue on risk): in press.
- List, J.A. 2006. "The Behavioralist Meets the Market: Measuring Social Preferences and Reputation Effects in Actual Transactions," *Journal of Political Economy* 114:1-37.
- Meyer, B.D. 1995. "Natural and Quasi-experiments in Economics." *Journal of Business and Economic Statistics* 13:151–161.
- Roe, B.E., T.C. Haab, D.Q. Beversdorf, H.H. Gu, and M.R. Tilley. "Risk-attitude Selection Bias in Subject Pools for Experiments Involving Neuroimaging and Blood Samples." *Journal of Economic Psychology*: in press.
- Rosenzweig, M.R., and K.L. Wolpin. 2000 "Natural 'Natural Experiments' in Economics." *Journal of Economic Literature* 38:827 – 874.
- Trochim, W.M. 2008 *Introduction to Validity*. accessed November 1, 2008. Available at: <http://www.socialresearchmethods.net/kb>.

	Relative Internal Validity	Relative External Validity	Topic and Subject Limits	Replicable?
Lab Experiments	High	Low	Long duration topics, larger stakes, losses	High
Field Experiments	Medium to High	Medium to High	Limited by researcher connections	Low to medium
Natural Experiments	Medium to High	High	Limited by occurrences of nature and policy	Low
Field/market Data	Low	High	Limited by privacy, recall and trade secrets	Low to medium

**Figure 1. Tradeoffs across research methodologies**