



Rigor in Behavior Experiments: A Basic Primer for OM Researchers

Daniel G. Bachrach, University of Alabama
Elliot Bendoly, Emory University

The study of the nuances of human behavior in operations management contexts and the behavioral reactions that accompany changes in operating policies has finally started to gain a strong headwind. This has come after several decades of operational modeling in which the behavior of the human actors, so critical to the mechanics of operating policies, has either been largely simplified or ignored. With the growth in joint work in experimental behavioral testing and improvements in behavioral codification, greater insight into the practicality of operational policies is now emerging. Yet in order to ensure such practicality, the rigor of this new joint experimentation needs to be ensured. While OM researchers have a rich history in the rigor of artificial modeling, the sparse history of behavioral experimentation in OM provides much less evidence of an understanding what "rigor" with such methods entails. The purpose of this brief primer is to touch on some of the basic tenants of rigorous behavioral experimentation, and to hopefully promote such rigor in future joint OM behavioral studies.

Four basic elements are typically viewed as critical to traditional experimental design, regardless of research context: (1) Random selection of subjects, (2) Random assignment of subjects to the different treatment conditions, (3) Experimenter manipulation of the treatments, (4) Experimenter control over the conduct of the experiment. According to established research design "random assignment of test units to treatment conditions facilitates causal interpretation by eliminating potential systematic differences across treatment conditions due to extraneous factors associated with characteristics of the test units" (Perdue and Summers 1986; Keppel 1982). The effort to sidestep extraneous effects is supposedly furthered by the direct manipulation of treatments imposed during the experiment by the researcher. However, it is worth noting that

discussions of what is referred to as "quasi-experimentation" in behavioral studies suggests that some of this rigor might be flexible in the interest of ensuring realism and robustness in result application (e.g. using real workers in action-study experimental designs) (Cook and Campbell 1979). Yet, regardless of the extent of strict control, several steps remain common in experimental behavioral analysis:

Conceptualizing the research question

Although seemingly obvious, the positioning of research questions, particularly those that are intended to "strongly suggest" causality of some form, is critical in ensuring the researchers ability to make use of any of the data subsequently collected for analysis (Keppel and Zedeck 1989). Once data is collected with the intent of assessing a firmly codified research question, it may be extremely difficult to apply it to any other purpose (given the extent of control typical in behavioral experimentation). For example, an OM researcher might be interested in asking the following question (typically taken for granted in standard modeling): "Do different levels of staffing produce differences in customer attitudes that in turn impact the difficulty or, by some alternate means, general productivity and throughput of the members of that staff?" Such a complex and suggestively causal question implies that we have an interest in at least three variable sets (i) staffing levels, (ii) customer attitudes and (iii) worker productivity/throughput. The question itself is the basis for the multiple research hypotheses that would subsequently need to be tested (as a set). Yet once fully codified and followed through experimentally, due to often specific timing of data collection phases, it would be very difficult to make an argument for the valid use of such data in answering questions regarding alternate forms of causality (e.g. how do customer attitudes effect staffing level choices?).

Operationalization and design

Given a well codified research question, the elements of the question (i.e. variables involved in the implied relationships) themselves need to be codified. The base-level independent variable in the example above appears to be staffing level – most likely operationalized as alternate numbers of bank tellers (e.g. 4 vs. 2) per shift.

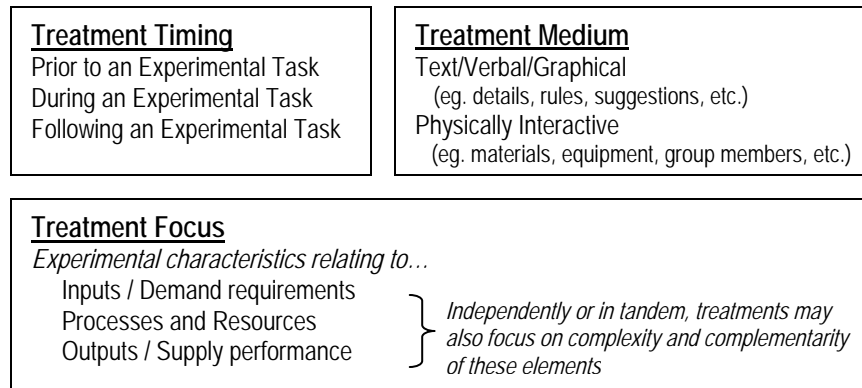


Figure 1. Dimensions Characterizing Common Experimental Treatments

What might be viewed as an intermediate outcome (a dependent variable whose repercussions impact subsequent variables) is the issue of “customer attitude”. Such a variable might be measured as a summary index of responses to a set of items in a post-hoc survey, or a statistically factor-analyzed construct of such items. The final dependent variables in this example, productivity and throughput, could be measured objectively (i.e. as time to complete work, and work accomplished per worker per unit time) and in this case potentially compared to model benchmarks (i.e. derived from models that assume constant productivity, independent of customer attitude and its potential antecedents). Ultimately, there are no true limitations on the nature of the data collected, however given the nature of OM research, the benefits of being able to collect and analyze objective data are clear, particularly if experimental behavioral research is intended as a precursor to behavior codification and subsequent use in artificial modeling. This is not to suggest that subjective scale assessments cannot prove useful toward codification as well, though OM researchers should do their best to support subjective scale interpretations with other hard numbers common to OM practice.

Another often critical use of subjective scales in experimentation however comes from the ability of researchers to ensure that the variables they are using to distinguish various scenarios “by design” (i.e. the experimental treatments) have the intended transparency and interpretability that they assume. The effects of well intended treatments are not always obvious, and their presence is often lost to the subjects they are intended to impact. Having said this, it

is worth noting that treatments used in experimentation vary greatly, characterized by issues such as timing, medium and focus. The careful specification of treatments along such dimensions is a pre-requisite to ensuring research designs that will ultimately yield interpretable results. If treatments are applied ad-hoc, with insufficient consideration or with a lack of definition that risks confounds with other treatments, the capabilities of the research design may be largely compromised (Cook and Campbell 1979).

Methodology and collecting data

Ideally subjects chosen for participation would be suitable proxies for the real-world roles in question to avoid possible added confounds and help ensure robust applicability of results. Action studies with the participation of real firms and employees and/or customers/partners would be ideal. In the present example, a researcher might find it sufficient to get authorization from bank managers to passively monitor flows under different staffing scenarios and ask customers post-hoc questions regarding attitudes (so as not to influence behavior inadvertently). Enticement might be through entry into lottery or alternate reimbursement schemes. In such scenarios, while the researcher does not have express control over staffing {and hence this aspect of study would not fall under the strict rubric of controlled experimentation} they would have control over the periods (ie. high versus low staffing, peak vs. low traffic) selected for examination and analysis. From a “general” research design perspective there may be nothing technically invalid in such an intelligent

selection (provided the researcher is honestly selecting data scenarios based only on criteria of condition representation, not on data-driven theory support). However such a design is not strictly consistent with the traditional interpretation of “controlled experimentation” outlined earlier. Criticisms of such a design might stem from issues with insufficient control over externalities that might confound data and analysis.

If concerns over control in design are at issue it is possible to substitute this process with a more traditional experimental design where the researcher has direct control over the independent variables of interest. In the best of all worlds managers at the context venue (here a bank) allow the researcher to manipulate treatment level populations (here staffing levels for example) with the hope of promoting unbiased and ostensibly random assignment of subjects to such treatments. Such is the nature of a truly rigorous action-style experimental study. In a less ideal but still potentially insightful scenario, the experimenter creates an off-hour fabricated setting such that manipulations do not threaten real world performance of the firm the individuals studied work for. Increasingly less realistic (yet more available and hence increasingly common in research) are experiments that use fabricated settings as well as alternate individuals (e.g. students) who lack real world experience with the context and associated tasks. While researchers may be able to illustrate an extensive level of control over such experiments, the tradeoff comes in criticisms of external validity (i.e. how can the observations be justifiably extended to real-world settings).

Validity testing and interpretability

Analyzing the data begins with scoring responses to any subjective measures collected in the study and combining those with any objective measures collected for each individual (or other unit of analysis that may happen to apply). These objective measures might include things like time spent in line, arrival time, task time, etc. In contrast, subjective measures might include things like perceptions of adequate staffing, type of service requested, post task (e.g. end-of-day) perceptions of the tellers, etc. If other experimental treatments are used (exit

signage describing peak hours and reduced fees at off-peak times – something clients are exposed to only as they leave the work environment), measures to check for their effectiveness should be consolidated at this point as well. In the event of multiple simultaneous treatments (i.e. more than one), the evaluation of these “treatment” checks is essential in order to demonstrate the validity of the experiment carried out. Without checks to validate such roles, the conclusions drawn with respect to the impact of the treatment classes acting on key dependent variables may quickly become suspect. As a result the credibility of behavioral experiments hinge on such validation, particularly when results are intended to be extrapolated towards practical application or subsequent theory development.

At least three classifications of treatment checks can provide meaningful support for researchers (c.f. Bendoly and Swink 2007, Bendoly et al. 2007). Those checks that serve to assess the ability of the treatment to characterize differing levels of an intended construct (ie. *manipulation checks*) focus on the convergent validity of the treatment. Manipulation checks are often best conducted through the use of well-developed or established multi-item scales indicative of each treatment, and the collection of subject responses to these items following soon after the treatment application. Comparative statistics (eg. t-Tests, ANOVA, etc.) are often used to test delineations of treatment levels and thus support convergent validity.

Other checks that serve to ensure that individual treatments do not confound other theoretically ‘independent’ issues of interest focus on discriminant validity. These secondary checks are often referred to as *confounding checks* (Wetzel 1977), and are often tested through comparative statistics as well – in this case testing whether the treatment levels inadvertently impact perceptions of other supposedly independently controlled issues. Both confounding and manipulation checks are particularly helpful in the “pre-test” or pilot phases of studies to ensure the main experiment, though should be included as part of the main experimental analysis as well.

Hawthorne checks (Adair 1984; Parsons 1992) against extraneous perceptual effects of

treatments constitutes a third validity test. Such checks are often conducted using supplemental measures not viewed as critical to the research questions studied but thought to be nevertheless related to the context studied. Successful results of such checks should suggest no impacts from any of the treatments on supplemental measures otherwise assumed to remain independent of the study. In this example, such supplemental measures might include customer perceptions of the convenience of the bank's "location". Perceptions of the availability of seating (or parking in a more realistic setting) would not be a reasonable measure for use in Hawthorne checks since line length and its relationship to staffing and throughput can reasonably be viewed as intertwined with such measures. Therefore successful validity checks of this nature require both an appropriate selection of supplemental measures as well as results that suggest they are not impacted by the design's treatments. If impacts are found, then the focus and isolation of the treatments can be called into question – and thus the clarity of the relationships analyzed.

It should be emphasized that to date, the vast majority of OM behavioral studies have failed to provide for any of the above checks.

Effect and relationship testing

If the treatments are found to be valid and all other elements of experimental rigor have been followed, researchers are then free to follow any series of analytical procedures deemed suitable in assessing their stated hypotheses. From ANCOVA to hierarchical regression to SEM, a wide gambit of methods can be applied. Rather than review such methods (which are specific to the nature of the relationships and data collected), it is sufficient to recognize that each has their own set of established criteria for analytical rigor that should be followed. Researchers need to not only follow such criteria but demonstrate an understanding of how the

tests they use align logically with the steps taken up to this point. If executed appropriately, the final codified model estimates can provide highly valuable new "behavioral" elements of existing artificial models, and thus extend the applicability of such models in practice.

References

- Adair, J.G. (1984). The Hawthorne Effect: A reconsideration of the methodological artifact. *Journal of Applied Psychology* vol. 69, no. 2, p. 334-345.
- Bendoly, E. and M. Swink (2007 forthcoming) Moderating effects of ERP information access on project management behavior, performance and perceptions, *Journal of Operations Management*
- Bendoly, E., D.G. Bachrach and B. Powell (2007 forthcoming). The role of operational interdependence and supervisory experience on management assessments of ERP systems, *Production and Operations Management*
- Cook, T.D. and D.T. Campbell (1979). *Quasi-Experimentation: Design and Analysis Issues for Field Settings*, Houghton-Mifflin, Boston.
- Keppel, G. (1982). *Design and Analysis: A Researcher's Handbook*, 2nd ed Englewood Cliffs, NJ: Prentice-Hall, Inc.
- Keppel, G. and S. Zedeck (1989). *Data analysis for research designs*. New York: W.H. Freeman and Company.
- Parsons, H.M. (1992). Hawthorne: An early OBM experiment, *Journal of Organizational Behavior Management*, 12, 1, 27-44.
- Perdue, B.C. and J.O. Summers (1986). Checking the success of manipulations in marketing experiments, *Journal of Marketing Research*, 23, 317-326.
- Wetzel, C.G. (1977). Manipulation checks: A reply to Kidd, *Representative Research in Social Psychology*, 8, 2, 88-93.