



Exploring Modeling Aspects of Design Experiments

by Finbarr C. Sloane and Stephen Gorard

In this article the authors use the process of model building (model formulation, fit, and validation) in applied settings to raise pertinent questions about design experiment (DE) methodology. We argue that the DE work presented in this issue highlights features of model formulation and local validation, but does not discuss model fitting or broader models of validation. This article marks out key areas for the DE community to address and concludes by positing that the concept of artifact failure in design research may be a more appropriate area of concern when designing an artifact (whether the artifact is a learning process or a software product). DE research is relatively new as an educational research method (Brown, 1992; Collins, 1992). We believe that DE researchers and the more general research methodology communities must work together to fully evaluate and reap the potential rewards of this developing research method.

A standard critique of design studies as presented in this issue might include, for example, concerns with external validity, measurement, the lack of control groups, and other sources of possible error (Shavelson, Phillips, Towne, & Feuer, this issue). Further complicating the inferential picture, design research data are drawn from small samples (convenience or otherwise) of teachers, of students, of settings, and of content. Additionally, design-study data are likely to be aggregated over time, over people at different organizational levels, and, in some cases, across sites. Drawing valid single-level or cross-level inferences from aggregated data requires that design experiment (DE) researchers begin to deal more explicitly with the possibilities for error in their work. Note that these observations apply equally to traditional clinical trials, quasi-experiments, or efforts to produce “paradigm cases” in design studies (Cobb, Confrey, diSessa, Lehrer, & Schauble, this issue).

We do not diminish the importance of a critique based on issues of validity and measurement error, and so on. We believe that for our purposes it is more advantageous to focus, instead, on the general features of model building, including model formulation, model fit, and model validation. We contrast these model features with the critical characteristics of DE research presented in this issue. The process of model building, we will illustrate, requires DE researchers to address issues associated with model fit that go beyond local model validity. However we view the intellectual struggles, presented by DE researchers in this

issue, with model formulation and local validity as a welcome contribution to research methodology.

Model Building in Applied Settings

“All models are wrong, but some are useful.”—George Box, 1978

There are three main stages in model building: (a) model formulation (or model specification), (b) estimation or fit, and (c) model validation. Introductory statistics courses usually emphasize estimation and, to a much lesser extent, validation. Model building is largely ignored in the early quantitative training of educational researchers. This is unfortunate, for it encourages novice researchers to think that all of statistics centers on problems of estimation. In reality, however, model formulation is often the most important and the most difficult stage of the research process. This centrality is reflected in the first rule of applied mathematics (often attributed to John Tukey): “An approximate answer to the right question is worth a great deal more than a precise answer to the wrong question.” The challenge for any model builder is to construct or select a model of the appropriate form and complexity. This observation holds, equally, for traditional quantitative research, DE research, and engineering. We see DE researchers as primarily involved in early stage model building with emphasis on model formulation. This is, we believe, one of the most useful influences of the DE work presented in this issue because these researchers bring the model formulation stage to the forefront.

Model Formulation Objectives

There are various objectives in model formulation (see, e.g., Cox & Snell, 1981; Gilchrist, 1984):

1. To provide a parsimonious description on one of more sets of data. By parsimonious, we mean that the model should be simple, but also complex enough to describe important features of the target process and the data that purport to represent it. DE researchers address this issue by using videotape and narrative accounts to reflect the complexity of the settings they are modeling.
2. To provide a basis for comparing several different sets of data.
3. To confirm or refute a theoretical relationship suggested a priori.
4. To describe and understand the properties of the random or residual variation, often called the error component. When based on randomly drawn samples, this enables the quantitative researcher to make inferences from a sample to the corresponding population, to assess the precision of parameter estimates, and to assess the uncertainty in any conclusions. Other than McCandliss, Kalchman, and Bryant (this issue), who use an approximation to Brown’s (1992) approach that includes hypothesis testing, DE researchers in this issue do not convincingly address possible sources of

error in their models, or the uncertainty of conclusions they draw.

5. To provide predictions which can act as a “yardstick” even when the model is known not to hold for some reason.
6. To provide empirical insight into the underlying process (sometimes by perturbing the model).

One should notice that this list does not include getting the best fit to the observed data. The purpose of model building is to construct a model that is consistent not only with the data but also with existing knowledge and assumptions about the processes that produces the data. In fact, it may be useful to construct more than one model using a variety of plausible assumptions about the “true” model and about what the future may hold.

Model Formulation Processes

General principles for model formulation are neglected in most introductory statistics textbooks. However, six are offered here. The modeler should

1. Consult, collaborate, and discuss with appropriate experts on the given topic, ask lots of questions, and most importantly, listen.
2. Incorporate background theory, not only to suggest which variables to include and in what form, but also to indicate constraints on the variables and known limiting behavior.
3. Collect and then examine the data to assess their more important features.
4. Incorporate information from other similar data sets.
5. Check that a model formulated on empirical or theoretical (or both) grounds is consistent with any qualitative knowledge of the system. Moreover, the selected model must be capable of reproducing this qualitative knowledge, and of reproducing the main characteristics of the data.
6. Remember that all models are approximate and tentative, to start with; and be prepared to modify a model during the analysis or as further data are collected and examined.

All of these principles are consistent with the steps taken by DE researchers in this issue. As a point of emphasis, we note that at all stages of model formulation it is helpful to distinguish further between (a) what is known with near certainty, (b) what is reasonable to assume, and (c) what is unknown.

With regard to Principles 2–4, it is worth noting that the extent to which model structure should be based on background theory or on observed data is the subject of some controversy. For example, in time-series analysis, an econometrician will tend to rely more on economic theory, while the statistician will tend to rely more on the properties of the data. To some extent this reflects different objectives, but it also indicates that model formulation, and consequently model building, depends partly on the knowledge, prejudices, and experiences of the researcher. We see these differences in the work presented in this issue. The Design-Based Research Collective (DBRC) derive their knowledge in part from the affordances of technology and from general principles of cognition. In contrast Cobb et al., Lobato, and McCandliss et al. (this issue) make content knowledge (e.g., dyslexia, mathematics) and cognitive processing the dominant lenses for their design work.

In general, some researchers wrongly ignore theoretical reasoning while others, we add, may place too much faith in it. For example, there are some areas (e.g., economics) where theories

conflict, and in these circumstances it is essential to let the data “speak.” Clearly, the theories we hold, and the training we have received, critically affect the data we collect and the lenses we choose in looking at such data. As one might expect, a combination of theory and empiricism is generally the most fruitful.

Model Estimation and Model Validation

In statistical modeling, the estimation stage consists of finding point and interval estimates of the model parameters. This work is detailed in almost all statistics texts and is not discussed further. Computer packages make it relatively easy to fit most standard models. All researchers know that this broad availability does little to help us formulate *sensible* models. In fact, there are many examples of “suboptimal” use of such complex techniques, some bordering on statistical fantasy (see the commentary by Maruyama, 1998, p. 275). DE researchers in this issue do not describe in what ways their final models may be suboptimal, and we believe this to be a serious omission.

When a model has been fit to a set of data, the underlying assumptions need to be checked and the model modified, as necessary. It is important to distinguish between gross violation of the model assumptions and minor departures. Diagnostic checks on a single data set, while valuable, can be overdone, particularly as there are philosophical problems in constructing and validating a model on the same set of data. It is far more important in the long run to see if a fitted model generalizes to other data sets rather than question the fine detail of each fit. This is true for all research models including those built by DE researchers. DBRC (this issue) echoes this commentary when it discusses the need for design principles that can be useful in more than one context and the difficulty in generating such principles. Thus, we would like to see model validation expanded to include checking the model on further data sets where possible and, where reasonable, returning to the original Brown concept of working toward, or iterating between, more “definitive” trials (see McCandliss et al., this issue).

Tukey and Box Revisited: Have We Asked the Right Question?

Although we believe that the steps previously discussed help increase the possibility for scientific inference in education, we now ask, “are these steps the one right way to design and evaluate educational interventions?” Lagemann (2002) challenged the research community to conduct research that resides in and better supports classroom practice. DE researchers have responded to this request. But, the quality of inferences that can be drawn from interventionist, real-time DE studies (not to mention the sheer quantity of data gathered) clearly presents significant inferential problems (Shavelson et al., this issue).

Nevertheless, the force of design research resides in an idea that unifies all of engineering—the concept of *failure*. From the simplest paper clip to the Space Shuttle, inventions are successful only to the extent that their developers properly anticipate how a device can fail to perform as intended (Petroski, 1996). The good scientist (or engineer) recognizes sources of error: errors of model formulation, of model specification, and model validation (and of course errors of measurement). When a model is built, we acknowledge how much error we have encountered

and, as best we can, where we have encountered it. Further, we look to justify our findings by indicating the degree to which we have theoretical fit. Designers, on the other hand, must also focus on failure (and sometimes gross failure at that). They too have these problems of error. But their end goals are different. Their goals reside in understanding failure, building better practical theory (“humble theory” in the words of Cobb et al., this issue), and building things that work (whether they are processes or products). This is seen most clearly in Japanese Lesson Study (Lewis, 2002).

Most engineers develop failure criteria, which they make explicit from the outset. These criteria provide limits that cannot be exceeded as the design develops. However, failure manifests itself differently in different branches of engineering. Some problems of engineering design do not lend themselves to analytic failure criteria, but rather to models of trial and error or to build-and-measure techniques. In the design of computer programs, for example, the software is first “alpha tested” by its designers and then “beta tested” by real users in real settings. These users often uncover bugs that were generated in the design, or in its modification. Furthermore, these users also show how the program might fail to perform as intended. No matter the method used to test a design, the central underlying principle of this work is to obviate failure. This is a very different model than the one followed by most social scientists—where the end goal is to produce unbiased estimators in support of robust theory. At issue here is a central question, and one that space requirements will not allow us to elaborate. The question put simply is this: Can and should educational research be a social science, an engineering science, or both? In this special issue, DE researchers provide their answer, Shavelson et al. offer their criticism. We take the middle ground, as our expressed goal is to generate conversations, not answers.

Conclusion

All research methods share dilemmas and choices associated with model formulation, model fitting, and model validation. These difficulties subsume others, for example, operational bias, “experimenter” effects, measurement error, and so forth. We posit that, by contrast, the central feature of design is to obviate failure. This feature is not shared or even framed in the same way by social scientists. Resolving the role and place of error versus that of failure will be central in bringing social scientists and design researchers together in ways that foster conversation, debate, and we believe agreement—to the benefit of all concerned.

NOTE

The views expressed in this article do not necessarily reflect the views of the National Science Foundation. We would like to thank Eric Hamilton, Larry Hedges, Anthony E. Kelly, Therese Pigott, and the anonymous reviewers for their encouragement, commentary, and feedback.

AUTHORS

FINBARR C. SLOANE is a program director at the National Science Foundation, Division of Research, Evaluation, and Communication, Room 855S, 4201 Wilson Boulevard, Arlington, VA 22230; fsloane@nsf.gov. His research interests include student learning of mathematics, educational policy, and statistical modeling.

STEPHEN GORARD is a professor at Cardiff University, School of Social Sciences, Glamorgan Building, King Edward VII Avenue, Cardiff, CF10 3WT; gorard@cardiff.ac.uk. His research interests include research capacity, indicators of equity, and patterns of lifelong learning.

Manuscript received June 15, 2002
Revisions received November 7, 2002
Accepted November 7, 2002