MULTILEVEL MODELING OF EDUCATIONAL INTERVENTIONS: EDUCATIONAL THEORY AND STATISTICAL CONSEQUENCES

Finbarr C. Sloane
School of Education, University of Colorado at Boulder, United States of America
Barry.Sloane@Colorado.edu

In this paper issues of educational context are described. The implications of these issues for the design of educational research are then articulated. These context variables (e.g., that students are instructed in clusters, and that teachers require ongoing training to support and implement new and innovative curricula) are then used as a lens to examine the ASA's 2007 report Using Statistics Effectively in Mathematics Education. A meta-model is offered to better address some of these concerns of educational context not fully articulated in the ASA report. The goal of the paper is to: (1) describe the component parts of this meta- model, and (2) generate the opportunity for richer conversation about the role and value of experimental statistics in education research.

BACKGROUND

Following the unifying conference theme: "Data and context in statistics education: toward an evidence-based society," the expressed goal of this paper is to map the phases of research used by statisticians to describe and analyze experimental medical research so as to attend to the context of educational research, particularly educational research that employs curricula treatments to improve student learning (American Statistical Association, 2007). In an earlier paper the author addressed the potential matching and mismatching between medical research and educational research and will not revisit this discussion here (Sloane, 2008a). In this paper I have chosen to intellectually explore the context of a particular form of educational research (experimental research in field settings) to articulate for education research what the late George Box might have described as the evolution of a program of research.

Box, Hunter and Hunter (1978) describe the central paradox of research design by noting that "... the best time to design an experiment is after it is finished, the converse of which is that the worst time is at the beginning, when least is known. If the entire experiment was designed at the outset, the following would have to be assumed as known: (1) which variables were the most important, (2) over what ranges the variables should be studied, (3) in what metrics the variables and responses should be considered..., (4) what multivariable transformations should be made... The experimenter is least able to answer such questions at the outset of an investigation but gradually becomes more able to do so as the program evolves" (p. 303). Consequently, a meta-model for a program of research in education must minimally afford the research community the opportunity to answer these, and of course, other questions. I present here one such model highlighting critical contextual issues for education research. In doing so, I discuss the distinctive characteristics of efficacy and effectiveness trials, dosage and double dosage concerns, fixed and random effects, along with their concomitant multilevel structure.

EFFICACY AND EFFECTIVENESS TRIALS FOR EDUCATION RESEARCH

Before efficacy or effectiveness research can be conducted in an educational setting it is assumed that an educational treatment of potential value has been developed. Often such development occurs in an engineering design space. In educational research this development work is most often conducted by employing qualitative methodologies. The inferred insights seem to draw on a process of elimination much like a doctor investigating allergic reactions in a patient through a process of oblation. Sloane (2010) has written about this process and draws on the idea of fractional factorial designs used in engineering to improve processes as a way to shorten the design cycle and improve these inferences. This kind of work is embedded in the early phases of the 10 phase model described later in this paper. However, for the purposes of our

In C. Reading (Ed.), *Data and context in statistics education: Towards an evidence-based society. Proceedings of the Eighth International Conference on Teaching Statistics (ICOTS8, July, 2010), Ljubljana, Slovenia.* Voorburg, The Netherlands: International Statistical Institute. www.stat.auckland.ac.nz/~iase/publications.php [© 2010 ISI/IASE]

discussion of efficacy and effectiveness trials, we assume that the educational treatment to be evaluated is already in place.

Efficacy trials should assess the value or worth of an education technology, treatment, or program. To do so they must provide tests of whether a technology, treatment, instructional strategy, or curricular program does more harm than good when delivered under optimal circumstances. By contrast, effectiveness trials provide tests of whether the formally tested efficacious technology, treatment, instructional strategy, or curricular program, again does more harm than good when it is delivered under real world conditions. *Efficacy trials must be considered a necessary condition for effectiveness trials to be viable, an issue often overlooked in educational research where efficacy and effectiveness trials are regularly confounded*. In contrast, this is explicit in the Phases of the randomized trials model described by the American Statistical Association (2007) especially Phase III trials. As Sloane (2008a) notes, Phase III trials are used to confirm and test the efficacy of treatment effects. They occur in multi-institutional settings, with careful standardization in the procedures being used. Generally, they are conducted with samples drawn from well defined populations. Further, they require large sample sizes (with appropriate power analyses conducted a-priori), and in many cases some subset analyses are planned for and executed. The studies have well defined endpoints and require randomized comparisons. They include serious study controls (e.g., double blinding, where neither the doctor nor the patient is aware of who is actually receiving the treatment). Moreover, they are monitored with extreme care. The reason that the Phase III clinical trial has been initiated is that the superiority of one treatment over the other has not yet been firmly established. Once efficacy is established, Phase-IV effectiveness studies are engaged to assess the effectiveness of the intervention in real world settings. Here is where the shift from internal to external validity is addressed in the medical research model. This shift from efficacy to effectiveness is formally acknowledged in the models set out by two divisions of the U.S. National Institutes for Health: The National Cancer Institute (NCI) and The National Heart Lung and Blood Institute (NHLBI). In contrast to the model rendered by the ASA (2007), both institutes support a five Phase continuum of research (where the pre-Phase-I stage devoted to basic research is considered as a formal phase of the research program).

The NCI cancer control research phases include:

a. Hypothesis development (P-I);
b. Methods development to ensure that accurate and valid procedures are available before a study actually begins (P-II);
c. Controlled intervention trials (P-III), where hypotheses developed in P-I are investigated with methodology validated in P-II). Often the case control methodology is employed at P-III;
d. Defined population studies (P-IV) to measure the efficacy of an intervention in a sizable, distinct, and well described population;
e. Demonstration and Implementation studies (P-V).

The NHLBI research spectrum includes the following phases:

a. Basic research (P-I). Research that seeks new knowledge about normal and abnormal function of the heart, lungs and blood and the etiology of their diseases;
b. Applied research and development (P-II). Research that asks what new ways can the results found in P-II be used in to achieve practical goals;
c. Clinical Trials. Conducted with large samples drawn from well specified populations to determine the efficacy and safety of the interventions;
d. Prototype Studies. Small-scale tests of refined programs using components of Phase III research to be efficacious; Further development of methods for future research are also conducted here;
e. Demonstration and Education Research. Tests of intervention effectiveness.

What becomes clear when we look carefully at these research phases is that, along with distinguishing between efficacy and effectiveness trials, the phases of research are fundamentally nested in nature. That is, research at one level builds on that at another level or phase. *No such model exists for educational research*. Were it to exist, such a model, if nested appropriately, would optimize the possibility for research knowledge in education to build and accumulate more consistently over time. The expressed goal of this paper is to offer such a conceptualization, for without such a framing national and international goals for an improved research infrastructure in education are unlikely to be attainable (Sloane, 2008b). Such a framework (or body of work) must overtly acknowledge the need for research that can and does address both internal and external validity (Mosteller and Boruch, 2002; Shadish, Cook and Campbell, 2002; Sloane, 2008a) in a nation's research portfolio. In the terms of the medical research model described here this portfolio of work must include efficacy as well as effectiveness studies before definitive statements can be made about 'what works' in the very real and changing world of schools.

THE META-MODEL: PHASES OF RESEARCH IN EDUCATION

A pharmaceutical company tests hundreds of new medications trying to find one that will be both safe and superior to the standard treatment for a specific disease. People, obviously, vary in responses to a medication. Following the medical model, testing is then conducted in phases (or stages) to assess which medications will add value to people's lives. Most medications are eliminated at the initial stage based on employing a small number of subjects. However, if a medication looks promising it is re-examined at a later phase of research where a more elaborate and severe test of its efficacy is made. The central problems then are to a) generate an appropriate set of hierarchical phases for research; and b) carefully delineate the sizes and severities of experiments at the successive phases. This is done so that new medications considered 'good' are unlikely to be discarded. It also serves to ensure that 'poorer' medications do not receive expensive and resource intensive investigations. Without a shared set of research phases being in place the task becomes impractical if not impossible.

In education we are also interested in developing high quality interventions and testing these interventions so they can be either discarded or deployed in the cauldron of schooling. No clear shared set of hierarchical phases currently exist to support this decision making process for education researchers. I argue that without such an organizing structure, supported by appropriate training (i.e., intellectual capital) and quality measurement, it is highly unlikely that knowledge drawn from education research can accumulate.

It is clear that a single paper such as this one will not suffice to fix the perceived problem of knowledge accumulation in education. So instead of offering the perfect answer to this perfect storm I offer instead a working meta-model. In generating this ten-phase model I draw heavily on components of each the models described earlier from the ASA, the NCI and the NHLBI to generate one that, I believe, better fits the needs of the education research community. I invoke George Box's now famous insight that "essentially, all models are wrong, but some are useful' (Box & Draper, p. 424, 1978) to indicate my hope that the working model offered here proves useful in developing "an evidence-based society" of education researchers dedicated to testing the actual value of innovations in curriculum and needed improvements in classroom practice.

TEN PHASES: A BRIEF DESCRIPTION

The ten phases are presented here under the headers: basic research, hypothesis and measurement development, pilot applied research, prototyping-A and B, efficacy trials, effectiveness trials, implementation trials, scaling research, and sustainability studies. For each proposed phase I will briefly describe the types of research question being investigated, the methods likely to be used.

Phase I: Basic Research

Basic research is the bedrock of scientific investigation across all disciplines. In fact, more rigorous research designs and complex theories are often (if not always) based on it. The researchers' or practitioners' intuition leads him or her to some conclusion based on limited data, with myriad alternative hypothesis, and with great opportunities to be wrong. The point, however, is that they also have the possibility to be right; and in either case provide the fodder, in the form of hypotheses, for more rigorous inquiries.

Phase II: Hypothesis Development & Measurement

Phase II involves developing the prenatal hypothesis of phase I by attaching to those early hypotheses very general conditions for success (design research), measures with which to gauge the complexity of the hypotheses and conditions under which they hold, and methods by which to scale those hypotheses if they are found useful. Phase II studies are conducted by many researchers, requiring a large amount of time to develop appropriate measures, methods and estimators to study the hypotheses with greater clarity, simplicity, and rigor.

Phase III: Pilot Applied Research

Phase III tests the refined hunches of phase I, distilled into the hypotheses of phase II, on small samples. Here we seek to find if the results found in phase II hold under mildly experimental conditions. That is, are the hypotheses sufficiently developed to be tested in more precise ways?

Phase IV: Prototyping-A and Phase V: Prototyping-B

I cluster these two phases because, although not the same, they are more interrelated than many of the other phases. Both involve small scale tests of phase III hypotheses that have been vetted against intuition (phase I), and the development of measures, methods (phase II), and basic experimental conditions (phase III).

The fundamental difference between the two phases is the type of insight that is sought. In phase IV researchers seek insights about individual students, using them as the unit of analysis. Phase V, however, involves seeking insights at the classroom level. The combination of phases IV and V, then, provide researchers an opportunity to investigate how a treatment, once appropriately measured and hypothesized, affects a sample across its naturally occurring levels (Raudenbush & Bryk, 2002). Moreover, dosage and double dosage issues will need to be explored. Sloane (2008a) noted that to deliver new curricular in schools requires in depth training for teachers, this training tends to be much more detailed than that involved in assisting someone to follow a well articulated protocol. As such, a double-dosage issue presents itself raising questions about whether the treatment is the new curriculum, the training program, or both. Put simply, the researcher must explore carefully how much training is actually needed for teachers to faithfully implement the treatment as originally conceived and tested so that it can be delivered with high degrees of fidelity and positively impact student learning. This is a non-trivial matter that raises salient questions regarding nested data structures, along with the possibility of seeing random as well as fixed effects when modeling the data produced by such a nested series of interventions (teacher training and curriculum implementation).

Phase VI: Efficacy Trials

An intervention or hypothesis under investigation at phase VI has been carefully measured, translated into methodologies and analytic methods, and now requires prime facie evidence. One does this with a randomized control trial to eliminate potential threats to internal validity. Of note is that this phase does not guarantee external validity, which is saved for phase VI research.

Phase VII: Effectiveness Trials

The efficacious treatment rising from phase VI research still may not work in the real world. In fact, the experimental act of randomizing may not have accounted for a variety of naturally occurring problems simply because the act of randomizing was ultimately experimental. It does ensure an effect is there, yes; but it does not ensure that the effect persists when experimental conditions are removed. In essence, internal validity does not ensure external validity. Phase VII research seeks to test just that: does an effect under experimental conditions (internal validity) persist in larger population under real world conditions (external validity)?

Phase VIII: Implementation Trials

An effective and efficacious treatment, one protected against the wiles of external and internal validity, now faces the real world *in situ*. This is when a treatment is given to larger populations, and allowed to bend to the variety of realities that individuals and groups thereof face on a daily basis. The myriad threats to internal and external validity, ruled out in the previous two phases, may find an alley in some previously unanticipated threat or event that requires researchers to return to an earlier point in the progression of phases. Effects may vary randomly or in a fixed manner across individual and groups, interactions between groups and individuals may change the nature of the treatment or the interpretation of its effects. The variety of plausible problems is challenging to even the most closed mind, and ultimately require that analyses in this phase of research be conducted with great respect for the complexity of students, classrooms, and the social network within which they exist.

Phase IX: Sustainability Research

A treatment that works for students and teachers, a population, in the real world, in all of its complexity is not out of the proverbial woods yet. Over time, the effect of the treatment may dwindle or disappear, increase or intensify in neither sought nor invited ways. What is required for a treatment to function over time? Does it ultimately harm some group of individuals, students or teachers as they grow older? Or as the treatment grows older? Phase IX research seeks to investigate these questions. The usefulness of this phase of research is well established in medical research.

Phase X: Scaling Studies

Finally a treatment, housebroken with checks against external and internal validity, is ready to be given en masse. The treatment may still deteriorate or fail to have an effect (although with carefully completed, previous phases this is unlikely), or worse yet, hurt students. With scaling studies researchers will be able to investigate, with large-scale quasi-experiments, the effect of a treatment as it diffuses throughout a population. They will be able to vet the experimental or quasi-experimental findings of previous phases against the real world result in application.

CONCLUSION: DRAWING CAUSAL INFERENCES FROM EDUCATION RESEARCH

Drawing valid causal inferences for the effects of interventions in education requires that the researcher attend carefully to the many issues associated with the context of education along with concerns regarding internal and external validity (Shadish, Cook & Campbell, 2002). The recommendations by bodies like the US National Mathematics Advisory Panel for example, tend to emphasize one form of validity only – internal validity. The push to deliver increased numbers of randomized trials in education research, with an explicit focus on internal validity alone, will likely result in studies with limited external validity. However, such research, while incomplete, is seen as a necessary condition in helping policy makers address “What Works” questions.

It is also true, however, that studies with high internal validity are but one step in a very long program of needed research (Box, Hunter & Hunter, 1978). It is my hope that the meta-model explored here, re-mapping the ASA’s rendering of the effective use of statistics in mathematics education research, sheds more light on the contexts where education research is conducted. In consequence, I also hope that this model increases the potential for the improved delivery of appropriate evidence from which more informed education based decisions can be made.

REFERENCES

American Statistical Association. (2007). *Using statistics effectively in mathematics education research***.** Alexandria, Virginia. Author.

Box, G. E., Hunter W. G., & Hunter J. S. (1978). *Statistics for experimenters: An introduction to design, data analysis, and model building*. New York: Wiley.

Box, G. E., Box, G. E., & Draper, N. R. (1987). *Empirical Model-Building and Response Surfaces*. New York: John Wiley & Sons.

Mosteller, F., & Boruch, R. (2002). *Evidence Matters: Randomized trials in education research.* Washington, D.C.: The Brookings Institution.

National Mathematics Advisory Panel (2008). *Foundations for success: The final report of the National Mathematics Advisory Panel.* Washington DC: U.S. Department of Education.

Raudenbush, S., & Bryk, A. (2002). *Hierarchical linear models: Applications and data analysis methods.* 2nd edition. Newbury Park, CA: Sage.

Shadish, W., Cook, T., & Campbell, D. (2002). *Experimental and quasi-experimental designs for generalized causal inference.* Boston: Houghton Mifflin.

Sloane, F. (2008a). Through the looking glass: Experiments, quasi-experiments, and the medical model. *Educational Researcher, 37*(1), 41-46.

Sloane, F. (2008b). Randomized trials in mathematics education: Recalibrating the high water mark for research in education. *Educational Researcher, 37*(10), 624-630.

Sloane, F. (2010). *Research design in the service of design researchers.* [Currently under review].