23 Longitudinal Analysis and Interrupted Time Series Designs

Opportunities for the Practice of Design Research

Finbarr C. Sloane and Brandon Helding
Arizona State University

Anthony E. Kelly
George Mason University

Introduction

Design researchers are interested in devising and iterating innovative interventions to support student (and teacher) learning, particularly in situations where the extant research provides little guidance. In the practice of design research, many students and teachers participate in multiple forms or iterates of the designed artifact or emerging practice. Multiple iterates make it difficult to know which version of the artifact, or changed practice, caused the observed change, or learning, on the part of the participants. Logically, there is no counterfactual (e.g., control condition). From the perspective of data modeling and analysis, one option is to treat students as their own control. That is, researchers can look at student trajectories over time and ask whether the growth trajectory changes significantly with parallel changes that are occurring in the artifact or practice (that is being re-designed). Taking a longitudinal perspective can provide design researchers with insight and sources of evidence to support causal claims they could not otherwise entertain. This chapter will describe the use of an “interrupted time series research design” as a plan for the conduct of design research. This type of design offers strength and direction regarding the number of subjects and the spacing of time-points needed to buttress the design researchers’ efforts to warrant their claims for the effectiveness of their design interventions.

We begin by describing a simplified design research study and suggest how its argumentative grammar could be structured by re-conceptualizing it, methodologically. Next, we outline recent changes in the analysis of longitudinal data. Then, we describe a research design framework for such analysis: the interrupted time series design, and show that this framework maps nicely to needs of design researchers. In sum, we show that it is possible to marry the many advances in statistical modeling and quantitative research methods to the emerging procedures, practices, and goals of design research.

The Simulation

Here, we simulate a design research study in which we envisage a researcher who is interested in understanding how students’ comprehension of ideas in mathematics (in an area with little prior research) can be advanced by the design and re-design of some artifact (e.g., instructional software, curricular intervention, or pedagogical technique). The goal of the researcher is to determine the extent to which it is possible to advance
the students’ comprehension as measured by a disciplinary description of the content, and, simultaneously learn from and provide feedback to the teacher:

1. The researcher should establish a baseline of competence or mastery in the domain for the students. This step requires one or more measures, which should be reliable and valid. These tests may be given once, or a number of times in order to establish the character of the trend in understanding.

2. The design researcher then intervenes with an early version of the artifact and determines the impact of this “perturbation” on the classroom system (i.e., how do the students and teacher respond?). This intervention is video-taped and otherwise documented.

3. As in Cobb et al. (2003), the design researcher then retrospectively analyzes the learning events with the teacher and other researchers. He or she may administer an additional test or other form of assessment. On the basis of this analysis, the design researcher then decides what changes, if any, to make to the artifact.

4. The design researcher may intervene with a modification to the artifact. The study then advances, repeating Steps 2 and 3 (with the evolving artifact) until some stopping rule is triggered (e.g., end of semester, unit, and resource limitations).

In Figure 23.1 we describe the general process associated with the simulation. First, we see that the testing of student knowledge without intervention shows little by way of student development. Next, we see the effects of the first designed iteration across four measurements. Finally, we see the effects of the second major design iterate. In the context of learning rational numbers, we might consider phases B and C as mapping to Vergnaud’s (1988) conceptualization of student understanding of additive and multiplicative mathematical structures. Explicitly, the goal of the design intervention might be: (a) to improve student knowledge of additive structure, and (b) to develop multiplicative insight and models on the part of students as they learn rational numbers. Design iterate number II, which occurs during measurement phase C, serves to develop a qualitative change in student understanding of rational numbers. It is clear that not all students will follow this idealized trajectory. However, the figure serves to highlight the explicit goals of the design researcher in terms of student learning of mathematics.

The products of the design research include documentation of the changes in student learning from the baseline, changes in teacher learning or behaviors, early and later versions of the artifact, some humble theorizing about learning and teaching this
Modeling Change: Recent Advances

Over the past 25 years, longitudinal modeling has become increasingly popular in the social sciences (Raudenbush & Bryk, 2002) because the tools to support these analyses have improved significantly (Hedecker, 2004). For example, the hierarchical linear model, a tool described by Sloane (this volume), allows for irregularly spaced measurements across time, time-varying and time-invariant covariates, accommodations of person specific deviations from the average time trend, and the estimation of population variance associated with these individual effects (Bryk & Raudenbush, 1992). However, longitudinal data analysis requires the researcher to be explicit about the outcomes of the designed intervention and how they will be measured. Sloane and Kelly address this issue of measurement in design research (this volume). We do not take up this issue in this chapter other than to note the need for high quality measurement tools that work well at single and multiple time-points (for greater specificity, see Singer and Willett, 2003).

In longitudinal models, subjects are measured on a number of occasions (three or more), and the researcher is interested in the shape of the learner’s development (or growth) over time and what predicts differences across learners in their respective growth curves. Specifically, the researcher is interested in answers to two separate, but entwined questions (Singer & Willett, 2003): (a) how does the measured outcome change over time for individuals (or groups), and (b) can we predict or model the character of these differences that occur over time? These two questions are central to design researchers who iteratively build artifacts to positively affect student learning. Moreover, they sit at the core of every study of change. The first question is descriptive in its nature, asking us to characterize each subject’s pattern of change over time. Is the change linear or non-linear? Is the pattern of change consistent or not? In our simulation this question asks if students improve their capacity to function as mathematical learners moving from additive to multiplicative insight with respect to their knowledge of rational numbers. The second question is relational and predictive. Here we focus attention on the association between independent variables and patterns of change in the sampled data. Does participation in the design study change one’s pattern of learning? Following our simulation, we use the baseline data to simulate what learning could have been without the designed intervention. We can now examine the data patterns to see if a significant change in the learning trajectory has occurred for students who participated in the design innovation. Additionally, we can ask if all subjects experience the same pattern of change. For example, do males and females share the same pattern of change?

The first question requires the formulation of a within-person model (the intra-individual growth model); the second question requires a between-person model (the inter-individual model). These questions are also important for design researchers. However, they are framed as what changes occur within, or during, a design iterate, and what predicts differences between these changes across design iterates (see Cobb et al., 2003; Design-Based Research Collective, 2003; Fishman et al., 2004).

In an effort to deal with these questions, the statistical modeling of change has improved significantly in the past 25 years (Hedecker, 2004). These relatively new modeling tools have not influenced the conduct of design research, perhaps because they are quite technical in nature. Whatever the reason, this result is unfortunate as the
iterative structure of design research leads itself nicely to the study of student growth over time. The design researcher can pose the following set of questions: what growth occurs in student learning during a design iterate of an intervention? Is this growth the same for all participants? Is the growth due to the innovation? Does the growth trajectory change for participants who stay in the study during design iterate two? Following our simulation, we ask if students draw on multiplicative models when responding to items used to measure rational number knowledge. Does the trajectory of learning increase or decrease as a consequence of changes in the learning artifact? Again, are the growth trajectories the same across groups of participants?

Clearly, the designed artifact changes carefully with each design iteration and many of the research questions will themselves recur in parallel with the number of artifact iterations. Consequently, we argue for a one-to-one correspondence between the style of analysis, and the design research paradigm as presented by members of that research community (Cobb et al., 2003; Design Based Research Collective, 2003; Fishman et al., 2004).

Longitudinal or Multiwave Analysis

We define longitudinal studies as studies where subjects are measured repeatedly, and where the research interest focuses on characterizing subject growth across time. Traditional analysis of variance methods for such repeated measures models are described by Bock (1975). These traditional methods are of limited use to design researchers because of restrictive assumptions concerning missing data across time, and the variance-covariance structure of the repeated measures. The univariate “mixed-model” analysis of variance assumes that the variances and covariances of the dependent variable are equal across time. This is rarely sustainable in practice. The multivariate analysis of variance for repeated measures is also quite restrictive. Models of this variety force the researcher to omit from the analysis subjects without complete data across all time-points. In general, these two procedures focus our attention on the estimation of group trends across time and provide little by way of assistance to our understanding of specific individual’s change over time. For these reasons, hierarchical linear models (HLMs; Bryk & Raudenbush, 1992) have become the method of choice for quantitative growth modeling of longitudinal data.

Several features make HLMs especially useful to the longitudinal researcher. First, subjects are not assumed to be measured on, or at, the same number of time-points. Consequently, subjects with incomplete data across time are included in the analysis. The ability to include subjects with incomplete data across time is an important advantage relative to procedures that require complete data for all children across all time-points because we can include all the collected data, increasing the statistical power of the analysis. Further, complete case analysis suffers from biases to the extent that children with complete data are not representative of the larger population of children. That is, many children miss days during the school year—this is particularly true of lower SES children—and when we exclude these children from our analysis our results reflect only those students who happened to always attend when data were collected. This bias would reflect very unrealistic school and design research settings. Second, because time is treated as a continuous variable in HLMs, individuals do not have to be measured at the same time-points. In general, this is useful for analysis of longitudinal studies in which follow-up times are not uniform across all participants. This is particularly useful to design researchers as it is unlikely that student interviews or testing for example will occur on the same day for all children. Third, time-invariant and
time-varying covariates can be included in the longitudinal model, providing a conceptually rich framework for analysis, a framework that better maps to the realities of schools and classrooms. Put simply, changes in student learning may be due to characteristics of the individual that are stable over time (e.g., student gender) as well as characteristics that change across time (e.g., individual interactions with the design innovation). Finally, whereas traditional approaches estimate average change in a population (across time), HLMs can also estimate individual change for each subject. These estimates of individual change are particularly useful when proportions (or groups) of students exhibit change that differs from the average trend. That is, HLMs afford richer insight when we expect different groups of students to grow at differing rates because these differences can be modeled. The HLM modeling technique affords the longitudinal researcher the opportunity to (Bryk & Raudenbush, 1992):

- Specify the structure of the mean growth trajectory.
- Model the extent and character of individual variation around the mean growth.
- Estimate the reliability of the measures for studying both status and growth.
- Estimate the correlation between entry status and growth rate.
- Model correlates of both status and growth.

In the next section we describe the general structure of the hierarchical linear model when used to investigate change. It should come as no surprise that its features parallel the two basic questions we have posed about change, the former descriptive and the latter predictive.

A General Two-Level Growth Model

Many individual change phenomena can be represented through a two-level HLM. At Level 1, each subject’s development is represented by an individual growth trajectory that depends on a unique set of parameters and some error. As a set, these individual parameters become the outcome variables at Level 2, where their variability can (possibly) be accounted for by a set of between-person characteristics. Formally, the repeated measures on each subject are considered nested within each individual. As a consequence, this model is less restrictive than the multivariate repeated measures model, allowing for uneven spacing of measures and missing data in the Level 1 model.

We assume the \( Y_{it} \), the observed status at time \( t \) for individual \( i \), is a function of a systematic growth trajectory (or growth curve) plus random error. It is convenient to assume that systematic growth over time can be represented as a polynomial of degree \( P \). Then, the Level 1 model is:

\[
Y_{it} (\text{Subject } i \text{'s response at time } t) = f(\text{Growth parameters} + \text{error})
\]  

(1)

In the Level 2 model, we ask if these growth parameters vary across subjects. We represent this parameter variation in the between-subject model:

\[
\text{Growth curve of subject} = f(\text{Subjects' background characteristics} + \text{error})
\]  

(2)

We use these two equations to structure our longitudinal investigations. The first equation forms the basis for our descriptive questions. The second equation supports our predictive explorations of the variation that occurs across individuals in the first equation. Consequently, the two models map in one-to-one correspondence with the two
questions posed earlier: (a) how does the measured outcome change over time, and (b) can we predict differences in these changes? The shape of the growth curve is central to our understanding, and estimation, of the designed innovation.

In sum, modeling techniques for longitudinal data have changed dramatically over the past 25 years, affording the educational design researcher more opportunity to map to the practical (and often) critical realities of classrooms and schools. These advances afford design researchers access to a set of analytic tools they normally do not use but which are now robust to the daily life of the learning settings in which design researchers work.

A Research Design for Design Research: The Interrupted Time Series Model

To properly discuss interrupted time series designs (ITSDs), it is first necessary to define them in some functional way. According to Cook and Campbell (1979), ITSDs involve multiple observations over time on the same units (e.g., particular individuals) or on different, but similar, individuals (e.g., same community or worksite); and require knowing when an intervention took place in order to compare before and after treatment. We draw on this definition and address each element of the definition separately. The elements include: methodological variations of interrupted time series designs, and considerations that need to be met when using ITSDs and its variants.

Observations

A crucial element of ITSDs is the number of time-points over which data are collected. The number of time-points needed varies based on the type of inferences and analyses intended. Time-points typically consist of observations on some measurement device (e.g., interview, assessment, or test) that occur, on a number of occasions before, during and after the treatment or design iterate. Whatever measure is used at each time-point, validity of the measuring tool is a required condition. Because validity is a very broad topic, it is not included in this discussion (see Nitko & Brookhart, 2007). Likewise, reliability of whatever measure is employed is critically important to the ITSD; although it, too, is not discussed in depth here (see Brennan, 2001).

The frequency and quantity of observations is largely based on other design characteristics, including analytic design features, expected main effects, and a desired model (Shadish et al., 2002). Classic statistical analyses (e.g., t- or F-tests) require that data at one time point be independent of data collected at any other time point. But, because ITSDs, much like design research, use multiple observations over the same or similar participants, its observations are necessarily autocorrelated. When traditional techniques are used to estimate the size of an autocorrelation, the task is laborious and large samples of time-points are needed (Box et al., 1994; Shadish et al., 2002; Velicer & Harrop, 1983). However, hierarchical or multilevel modeling tools allow one to adjust for autocorrelated errors without the need for this additional data collection—sometimes in the hundreds. In fact, one can get by with as few as three measures per design iterate (Hedeker, 2004). Additional design features can be integrated into an ITSD to reduce the number of observations needed, but trade-offs between the quality of inferences desired and the number of observations are inevitable. The combination of the ITSD with modern multilevel statistical tools offers the design researcher a flexible analytic environment in which to work.
**Same or Similar Units**

Interrupted time series designs include multiple observations over time on the same or similar units. These “units” are embodied in the social sciences as people or groups of people. For the learning scientist, these units include individuals, dyads, groups, and classrooms of learners. One requirement of ITSDs is that the same units be employed over time. This requirement is not problematic as many design researchers use the same students over many cycles of the design process. And, fortunately, this is an advantage from the ITSD perspective.

**Knowledge and Timing of the Intervention**

Interrupted time series designs can provide a reasonably strong “causal” frame even in quasi-experimental settings. Consequently, this discussion will focus on ITSDs when randomized control-treatment groups are not possible and multiple observations are used as the quasi-experimental alternative. One of the primary distinguishing elements of the ITSD, apart from more general time series designs, is knowledge of the intervention.

The onus of the intervention is yoked to the researchers or those implementing the research. This feature is again advantageous from a design research perspective. This is especially true when changes in the intervention itself are held constant during different intervals (or cycles) of the time series, the fidelity and similarity of each treatment (within cycle) is then vital to comprehensible results. Given that design researchers develop the innovation, this feature of ITSD aligns nicely with the design researcher’s intent.

The three design elements—multiple observations, timeliness of intervention, and similar units of analysis—work in concert to provide for strong inferential possibilities. Interrupted time series designs, like all quasi-experimental designs, are fraught with plausible alternative causes, but accurate knowledge of intervention with multiple observations across similar units serve as design elements that can be optimized to reduce the likelihood that alternative hypotheses come into play, and in doing so they increase the resulting accuracy of the final inferences.

**Methodological Variations**

To address threats to internal validity listed by Shadish et al. (2002), we suggest possible alterations to the ITSD. Variations on the ITSD include adding a control group, observing multiple variables, removing treatment, lapsed treatment, and switching control and treatment groups (ibid.). We discuss a number of these alternatives that we think may be appropriate to the design researcher.

**Including a Control Group**

The classic variation on ITSDs includes a no-treatment control group (Shadish et al., 2002). Randomization to the conditions is preferable because it increases the internal validity of the study, but is often not possible. In many design studies no control group is included, and as noted earlier, ITSD allows the individual students to serve as their own control as long as good baseline data are available. The inclusion of a different group that is not randomly selected and not given an intervention allows for comparison over time with a treatment group and strengthens resulting inferences. Because the comparison group is not randomly selected, trade-offs between inferential ability and
experimental design are inevitable. Systematic differences between groups confound inferences, but insofar as the groups are similar, and systematic differences do not exist, the effect demonstrated through observation is the result of the designed artifacts (rather than latent or unknown causes).

Observing Multiple Variables

Shadish et al. (2002) also suggest adding measures that focus on another variable that is conceptually related to the primary dependent variable and that is equally sensitive to threats of internal validity. Consequently, any documented change in the main dependent variable that does not occur in the secondary dependent variable is unlikely to be caused by a threat to validity that both variables share and increases the internal validity of the study.

Removing the Learning Artifact

Another concession to the threats of internal validity is adding a treatment over some number of intervals, and then removing the treatment. The ITSD constitutes, in this case, essentially two interrupted time series experiments, one that tests the addition of the treatment, and one that tests its removal (Shadish et al., 2002). Problems associated with plausible alternatives still exist, especially when a treatment’s effect is long-lasting or permanent. For instance, if a treatment’s effects are permanent, removing that treatment will not yield any new results. Furthermore, given the sensitivity of children as research subjects, it is not always ethical to remove a treatment that proves effective. As such, removing a treatment in ITSDs is not always possible in practice, but can, and does, strengthen the inferential space when this level of control is available and appropriately used by design researchers.

Considerations

We believe that the interrupted time series research design, aligned with improvements in the statistical modeling of longitudinal data, provides the design researcher with a quantitative methodology to investigate the effect of design-based innovations. However, this belief is not without caveat. Assumed throughout this discussion of ITSDs and previously mentioned, is the a priori requirement of a reliable and valid measure at each observation. This point should be emphasized. Without a reliable measure each observation may produce slightly different results and mimic a successful treatment. Moreover, effective results can also be missed due to lack of reliable measurement. These are not trivial insights and the issues of measurement in design research are deserving of further investigation (Sloane & Kelly, this volume, address some of these concerns and highlight others). Additionally, without a valid measure each observation may not reflect the intended phenomenon and produce spurious or falsely negative results. A reliable and valid measure is critical to effectively implement an ITSD in a design research setting.

All treatments do not behave equally. Some treatments are slow acting. That is, they take time to diffuse sufficiently throughout the sample to produce results (Shadish et al., 2002). Step functions are one way of analytically negotiating slow-acting treatments (Holder & Wagenaar, 1994; Shadish et al., 2002). Other treatments have delayed effects that, without long observational periods, can be missed. Some treatments may be more abrupt than expected. Innovations lasting longer than necessary can produce
unintended consequences that can overshadow the effect of one properly timed. The 
operative point is that treatments can have unpredictable effects over time based on 
idiosyncrasies within samples (and the treatment itself), and we need to exercise due 
caution in our inferences. In sum, the interrupted time series research design offers the 
potential strong warrant when properly implemented and analyzed.

Some Lessons for Design Researchers

Design researchers are invariably interested in student learning. They spend their 
research careers devising and iterating innovative ways of supporting student learning. 
Whether framed explicitly or implicitly, fundamental questions in design research cen-
ter upon issues of individual and group learning. If we are to improve educational 
practice in our schools by way of improved instruction, or by the deployment of innova-
tive technologies, then the accurate measurement of individual learning can provide one 
yardstick by which the effectiveness of pedagogy and the effectiveness of innovation can 
be judged. Similarly, valid and reliable statistical tools are required for the measurement 
and modeling student learning in designed settings. Design researchers employ iterative 
design as a central tenet of their work. Moreover, much design research is conducted 
with students in what might be considered a case setting. As such, many students 
participate in multiple forms or iterates of the designed artifact. This makes it difficult 
to know which version of the artifact caused the wanted change or learning on the 
part of the participants. One way to work around this lack of control is to treat students 
as their own control. That is, researchers can look at student trajectories over time 
and ask whether the growth trajectory changes significantly with parallel changes that 
are occurring simultaneously in the artifact under study. Taking a longitudinal perspec-
tive can provide design researchers in education insight they could not otherwise 
entertain.

The ITSD is a mode of investigation that allows us to explicitly consider change over 
time when the intervention is well understood, or in development (but held constant 
within a design iterate), when multiple observations are possible during the testing of a 
design iterate, and when the same or similar units of analysis can be measured at each 
observation. Having access to a control group can yield an experimental design that 
adores both to traditional control treatment and interrupted time series research 
designs. We believe that the interrupted time series design, when aligned with appropri-
ate analytic techniques (e.g., HLMs or other multilevel modeling techniques) provide 
the design researcher with a stronger arsenal of research tools that go beyond case study 
methods. The employment of this new tool kit will impose measurement precision on 
the design researcher.

Conclusions

In this short chapter, we sought to indicate by use of a simulated example how it is 
possible to marry the many advances in statistical and research methods to the emerging 
procedures, practices, and goals of design research. We hope that this work can serve to 
inspire conversations among different genres of research to the mutual benefit of all in 
the service of advancing educational research methods and the tenor of our scientific 
claims, generally. We hope that this chapter has helped increase the awareness and 
understanding of longitudinal methods and longitudinal designs and their potential for 
analyzing the learning outcomes inherent in design-based research.
Acknowledgments

The first author would like to thank Dan Battey (Mary Lou Fulton College of Education, Arizona State University) for his having to listen to the many drafts of this paper that occurred orally. Additionally, the first author acknowledges funding from the National Science Foundation, NSF 0634103 (Moving from Arithmetic to Algebra) and NSF 0412537 (i.e., MSP: Project Pathways). Opinions expressed here are those of the author and not necessarily those of the Foundation.

Note

1 We suggest that the design researcher collect measures on multiple occasions within a design iterate. Here, we define a design iterate as a period where the innovation is reasonably stable. That is, it is a period where the researcher (and his or her team) is involved in thinking about changes to the design but has not as yet enacted those changes. The length of this “period” is not fixed across iterations.

References


