The Value of Modularity in Instructional Design: Implications for Improved Validity in the Evaluation of New Techniques in Distance Learning

Kenneth A. Yates
David F. Feldon
University of Southern California

In the development of distance learning, advances in cognitive science merge with new technology to deliver instruction worldwide. However, one major difficulty in evaluating the efficacy of these tools is determining which elements of instruction truly lead to observed changes in student performance. As content, pedagogical methods, and media are intertwined, identifying the "active ingredients" is an essential element of facilitating training that is of high quality and minimizes development costs (Clark & Estes, 2000). To accurately evaluate applications in the field, researchers must be able to identify specific instructional components, make decisions on who and what will be subject to treatment, and accurately draw inferences regarding causal interactions without the control offered by a laboratory. The purpose of this paper is to briefly review current use of various research methods for evaluating instructional technologies, discuss previous solutions to balancing the conflicting demands of internal and external validity, and then to propose a new research design that achieves this goal in a manner compatible with many instructional technology applications.

Current Methodologies in Instructional Technology Research

As educational researchers have grappled with the challenges of obtaining robust findings that successfully transfer from the laboratory to the field, use of experimental and quasi-experimental designs has decreased in favor of contextually embedded case studies that are descriptive in nature (Winn, 2002). Emphasizing the post hoc measurement of past learning outcomes or present conditions, it has been asserted that descriptive research may often be the only feasible way to study relationships between educationally relevant variables. Describing and interpreting "what is" makes descriptive research particularly appropriate to study questions, where significant variables cannot be manipulated in an authentic setting without being detrimental or threatening to human subjects (Best & Kahn, 2003).

The movement toward qualitative studies can be examined from the perspective of new entrants to research in instructional technology. Caffarella (2000), analyzed the content of doctoral dissertations in educational technology from 1977 through 1998 and found a clear shift in the selection of research methodologies. Over that time span, he observed a reduction in the number of experimental studies and an increase in the number of qualitative studies. Noting that his selection method most likely underestimated the actual percentages and that doctoral studies tend to be concentrated in a few institutions and professors, he concluded that "the balance between experimental and qualitative studies will most likely continue to show a change" (p. 20).

In their review of distance education research from 1990 to 1999, Berge and Mrozowski (2001) evaluated 890 research articles and dissertations and concluded that approximately 75% involved descriptive research, while only 6% employed a valid experimental approach. Further, their analysis of the data indicated that pedagogical methods dominated the research and that the trend continues to increase. Similarly, Ross and Morrison's (2003) analysis of 424 articles published by *Educational Technology Research & Development*, and its predecessor publications from 1953 through 2001, found a decline in the use of true-experimental studies from their height (77%) during the period 1983-1992, to 53% during 1993-2001, while descriptive studies increased from 13% to 45% during the same time periods. They suggested that this demonstrates the growing influence of alternative designs, such as case studies.

Their analysis of the types of stimuli and assessment used in these studies was also noteworthy, because it revealed a significant decrease in the use of materials that were developed only for the purposes of conducting the reported study, such as nonsense words and fictitious content, in favor of authentic curricular materials from 1993-2001. They suggest that this trend is due to the increased interest in external validity and an increased concern about the applicability (generalizability) of laboratory-controlled findings for real-world settings.

An Examination of Internal and External Validity Considerations

Ironically, the pursuit of external validity through non-experimental methods inherently limits the generalizability of the findings. While case studies and other qualitative methods are very helpful for identifying potential pivotal variables for further study and capturing the experiential elements of individual experiences

regarding the technology in question, they are inherently not generalizable beyond the time- and activity-specific setting in which the data was gathered (Stake, 1998). Likewise, the limitations of descriptive studies that utilize quantitative correlational methods prevent researchers from validly concluding that use of a particular technology of interest caused any reported outcome (Pedhazur & Schmelkin, 1991).

Addressing the relationship between external validity and generalizability, Banaji and Crowder (1989) caution against conflating the two:

> No one would deny that, other things being equal…ecologically valid methods…used to achieve generalizable results is the best situation in which to find oneself. Nor could it possibly be denied that the combination of contrived, artificial methods and conclusions with no external validity produces a sorry state…. The multiplicity of uncontrolled factors in naturalistic contexts actually prohibits generalizability to other situations with different parameters. The implication that tests in the real world permit greater generalizability is false once the immense variability from one real-world situation to another is recognized. (pp. 1188-1189).

In fact, they suggest that when a tradeoff must be made between ecological validity and generalizability, it is generalizability that ought to win out, in order to advance our understanding of causal mechanisms that can be harnessed to develop technologies. Ecological factors, they reason, can validate the generalizability of a causal mechanism, but the uniquenesses of particular settings can obscure the accurate identification of causal principles. As Morton (1991) noted in his commentary on their article, "generalizations that do not extend outside the restricted environment in which they were bred are not of much use, irrespective of their beauty" (p. 33).

Especially in educational technology research, investigators must be cautious of sacrificing internal validity, because such a strategy can lead to confounded results, such as those identified by Clark (1983; 1985) in his discussion of media selection and instructional strategies. Clark's analysis of effect sizes in media research revealed that the most common sources of invalid causal inference in technology studies were the uncontrolled effects of the differences in the method or content compared and the novelty effect for the use of a newer media. Because the demands of authentic settings often require the use of imperfectly matched instructional materials (e.g. lecture vs. interactive simulation) or blatant introduction of new apparatus, it is often extraordinarily difficult to design research that meets the demands of Banaji and Crowder's (1991, p. 78) "best possible situation."

Campbell and Stanley (1963) identified eight common threats to internal validity that manifest as a design's inability to control the influence of extraneous variables (see Table 1). These threats are best controlled in the laboratory, where "true experiments" provide equivalence of the subject groups and tight controls over variables that cannot be managed in the field. However, to a great extent, carefully considered quasi-experimental designs can also control these potential confounds. As discussed above, the traditional disadvantage is that the controlled laboratory environment often reduces the robustness of findings by limiting external validity. However, the tension between laboratory-based internal validity and field-based external validity can be alleviated by creatively combining research designs to measure the effects of instructional technology innovations (Clark & Snow, 1975).

Table 1: *Campbell & Stanley's (1963) Threats to Internal Validity*

History	Events, other than the experimental treatments, influence results.
Maturation	During the study, psychological changes occur within subjects
Testing	Exposure to a pretest or intervening assessment influences performance on a posttest.
Instrumentation	Testing instruments or conditions are inconsistent; or pretest and posttest are not equivalent, creating an illusory change in performance.
Statistical Regression	Scores of subjects that are very high or very low tend to regress towards the mean during retesting.
Selection	Systematic differences exist in subjects' characteristics between treatment groups.
Experimental Mortality	Subject attrition may bias the results.
Diffusion of Treatments	Implementation of one condition influences subjects in another condition.

Past Attempts to Satisfy Both Types of Validity

Building upon classic research methodology, Clark and Snow (1975) offered one of the original treatises in proposing alternative research designs specifically for instructional technology. Common design problems in the studies they reviewed included reliance on pre-experimental research designs, which lacked random assignments to conditions, control groups, and/or equivalence among the subjects. Consequently, it was impossible to draw valid conclusions regarding the causal relationships between subjects' characteristics, technology design, and observed

outcomes.

Unfortunately these shortcomings are still in evidence today (Clark & Estes, 1998, 1999). Much of the activity in educational technology, Clark agues, has been the practice of craft, not technology. Whereas technology is the application of scientific principles to solving real-world problems and the generalizability of solutions, craft is characterized more by situated trial and error and solutions that are indeterminate, non-transferable, and unconnected to a systematic knowledge base. Further, the result of this confusion in the practice of educational technology is reflected in research studies that report "no significant difference," that fail to isolate the "active ingredient" for effect attribution, and that are not generalizable to different contexts.

Various alternative design solutions for balancing the conflicting demands of internal and external validity have been proposed. Two designs that have recently come to the forefront are the randomized field experiment and the design experiment. The randomized field experiment (Ross & Morrison, 2003) requires that subjects in the treatment group be randomly selected to eliminate selection as a threat to internal validity. The advantage of this type of experiment is its high external validity; however, internal validity with respect to history and the overall complexity of the experiment allows for confounding variables.

Design experiments (Winn, 2003) are also conducted mostly in field environments. In this design, the treatment, such as an instructional tool or strategy, is applied in an educational setting, and data is gathered. Based on an analysis of the data, the treatment is revised and applied with additional data collection. This iterative intervention continues over time until the treatment is proved consistently effective. It is the iterative nature of design experiments that controls for spurious variables, in that the intervention can be adapted to correct a problem that the analysis of the data reveals. Design experiments should be replicated to establish validity; however, "the techniques for gathering and analyzing data in design experiments are typically less prescribed and less procedural than those used in experimental studies" (p. 370).

A New Design

Although conducted in real world settings, randomized field experiments and design experiments radically tip the balance of the evidence collected in favor of external validity. Due to the threats to internal validity inherent in both designs through history, maturation, and diffusion of treatment, the results of studies using these and similar designs may only be useful to generally interpret learning outcomes, rather than assist researchers to understand the specific mechanisms underlying and supporting student learning. It is critical, therefore, to continue to refine experimental designs in educational technology research to improve the yield of experiments in complex field settings, the causal inferences they provide, and the generalizability of these inferences to constructs over a variety of populations, settings, treatments, and outcomes (Shadish, Cook, & Campbell, 2002).

In the case of distance learning, the balance between internal and external validity must be informed by the realities of its implementation. In addition to taking place over a geographically distributed area at varying times, course designs are often modular in nature. That is, instructional content is divided into sequential modules that deliver the curricular content in a logical order (Khan, 2001). Because some content is often dependent on the mastery of material presented in a previous module, consideration of such features is vital during the experimental design process to preserve both internal validity (i.e. eliminate history and maturation threats) and external validity (i.e. maintain authentic delivery).

Built specifically with sequential module-based courses in mind, the Strand of Pearls (SOP; see Figure 1) design consists of four conditions into which distance education subjects are randomly placed. Each condition presents the course material in the same sequence and with the same timing. What differs among the conditions for each module is solely the inclusion or exclusion of the targeted feature to be tested. The first condition entails a standard sequence of instructional modules that do not include the feature to be tested as an equivalent control group against which treatment modules are compared (Group A in Figure 1). As with a conventional untreated control group design, the second condition into which subjects may be assigned delivers each module in the same sequence as the first group, though every module includes the experimental feature to be evaluated (Group D in Figure 1). The third and fourth conditions represent systematic alternating sequences of the experimental and the control versions of the modules (Groups B and C in Figure 1). The difference between the third and fourth conditions lies in the sequencing of the experimental and control modules. In Group B, the first quarter of the modules in the course are control versions, the second quarter are experimental versions, and so on. The order for Group C is a reverse of the B, so that the first quarter of the modules use the experimental version, the second quarter use the control modules, etc. In this way, half of the subjects across all conditions are using the experimental modules at any point in the course.

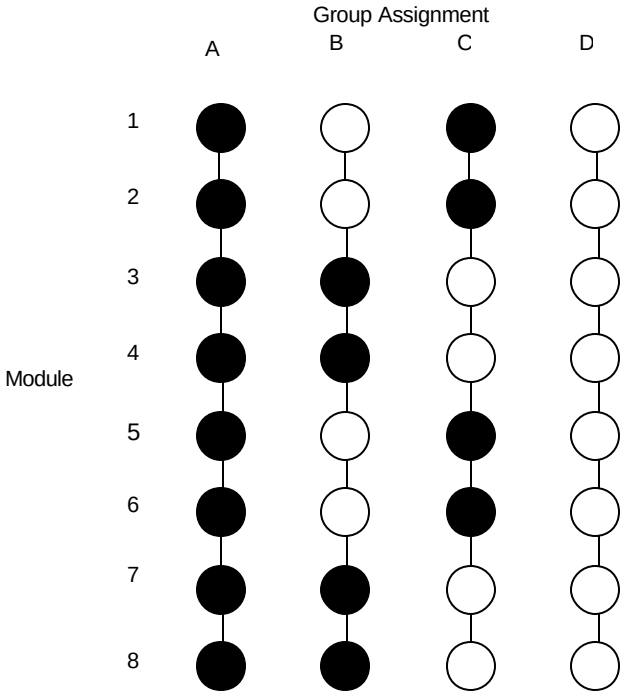

Figure 1.

In each module where it is present, the feature is implemented in precisely the same way in order to maintain the comparability of effects both longitudinally within groups and cross-sectionally across modules. By alternating the availability of the feature across two of the four groups, it becomes possible to detect any confounds or threats to internal validity due to time enrolled in the course (history, maturation), widespread environmental influences (history), and aberrant results due to an interaction between the feature tested and module-specific content.

In order to achieve external validity, the features of instructional delivery must be consistent between actual and experimental use. Such elements include the settings in which subjects participate in the course, the technological impediments that might impact the use of technological features, the presentation sequence and content of the curriculum, and the assessment mechanisms. Additionally, subjects' attention and actions must not be unduly influenced by awareness of the fact that they are participating in an experiment (i.e. the Hawthorne effect). Each of these elements is satisfied, because the study occurs through the vehicle of the course delivery itself. As such, all of the user experiences are authentic and limit the visibility of the research being conducted.

Further, generalizability is also preserved, because the sole consistent difference between a particular module used by the experimental groups and control groups is the availability of the instructional tool being tested. Thus, the findings will predominantly reflect only that manipulation. Typically, one of the major confounds in field studies is the environment in which each subject is participating. In classrooms during live instruction or in other scenarios in which subjects are in a particular physical setting, there are many uncontrolled commonalities (e.g. culture, socio-economic status, teacher bias, etc.). However, in distance learning scenarios, such environmental influences are limited, because the variance introduced by environmental effects on a single subject will not have a statistically significant impact on the overall results, if sample sizes are sufficient to provide appropriate statistical power. Further, it can be argued that those environmental variables consistent enough across users to generate an overall effect are common to the larger population of distance learners.

An additional evaluative approach to enhancing generalizability leverages the sequential and modular design of most distance learning courses. By transparently evaluating subjects' knowledge levels at frequent intervals, using pre- and post-test assessments between each module, longitudinal data is acquired for each subject that can be used to detect statistical outliers within individuals over time, in addition to particular subjects relative to the sample population. Thus, if a subject generates aberrant data for a particular module (possibly for reasons of environmental interference that would not be representative of the general population), that data can be withheld from overall analyses to increase the generalizability of the findings.

The advantage of this design over other research paradigms is that it combines the external validity of a staged innovation design (Clark & Snow, 1975) with the internal validity of an equivalent time samples design (Best & Kahn, 2003). The strength of the staged innovation design is that the "control" and "experimental" groups are created by time-shifting the treatment; however, this is also its weakness for internal validity, especially with respect to history. The traditional equivalent time samples designs, on the other hand, minimize the effect of history; however, they increase the strength of potential confounds, such as maturation, unstable instrumentation, testing, and experimental mortality. In the proposed design, the four conditions, considered collectively, serve as counter-balanced controls to evaluate not only the net effect of the experimental feature, but also the intraindividual effects on the rate and outcome of skill development related to the module design being tested.

An Example of Application

To illustrate the advantages of the Strand of Pearls design, we will consider the case of a distance learning course in U.S. law for foreign students taken in their home country. Historically, such students have either received conversational English language instruction prior to the course or enroll in a conversational English language course concurrent with the law course. However, recent studies have indicated that these approaches do not adequately prepare these students to succeed in advanced courses (Brostoff, Sinsheimer, & Ford, 2001; Feak & Reinhart, 2002; Hanigsberg, 1994). As the language of law and legal studies is complex and difficult, these students might benefit from language instruction embedded in their law courses to master the content objectives and language learning goals simultaneously.

In such courses, students must be able to read, summarize, and analyze previously adjudicated cases from the U.S. Supreme Court and various Courts of Appeal to learn how to identify the legal principle at issue, the facts of the case under consideration, and the reasoning the court used to reach its finding. Both the density of the language, often comprised of long sentences containing multiple conditional clauses and the organizational intricacy of the case often defeat students' attempts to extract the necessary information.

One method to enhance the readability, comprehension, and retention of complex text is the use of textual cues (Kerper, 2001). Textual cues are added to the reading text by reformatting the page to create additional white space, allowing the placement of "textual cues" in the form of headings or questions. These cues direct the reader's attention to a particularly important fact or issue that is essential for understanding the reading assignment.
In our example, the purpose of our study is to determine the effect of the treatment, in this case, the use of textual cues in reading assignments, on students' achievement of the course objectives, with the corresponding null hypothesis that the treatment has no effect on student achievement. As each course module contains the text of one complete case taken from the body of U.S. case law, the achievement of the individual module's learning objectives, and thus the overall course goals, is dependent on the understanding of each module's case. As such, the course is an ideal mechanism for measuring the effectiveness of this proposal and for testing our hypothesis by helping to isolate the respective influences of language and skill proficiency on outcome assessments.

The course consists of 8 ordered modules of computer-based instruction. Each module incorporates multimedia, the court case, exercises, and self-assessments. Two versions of each court case are prepared, one version that incorporates textual cues, and another version in which the text remains in its original form. The study's control and experimental conditions are solely determined by which version of the case text is present in the module. Except for this intervention, all other module components are identical in content, method, and instructional sequence in the control and experimental conditions.

The course is delivered to students via an Internet Web site. Prior to commencing the course, participants complete a questionnaire to collect background data with respect to age, level of education, experience, and English language fluency. In addition, pretests to measure knowledge of U.S. law and to assess English language skills are administered to the participants.

Following the SOP design, participants in the study are randomly assigned to one of the four conditions. The control condition (Group A in Figure 1) provides instruction in law only and does not include the textual treatment, and the experimental condition (Group D in Figure 1) includes the textual treatment. As described above, the third and fourth conditions (Groups B and C in Figure 1) alternate between the language treatment-included (TI) and treatment-excluded (TX) modules in two-module segments (i.e., TI-TX-TI-TX and TX-TI-TX-TI sequences). After each module, a combined posttest of the previous module and pre-test of the subsequent model is completed by each participant. With the use of module-specific and overall pre- and posttests, a rich data pool can be analyzed with great confidence in both internal and external validity.

Longitudinal Analyses Across Modules. For instance, it is conceivable that a direct comparison of the achievement data for Group A and Group D following completion of all 8 modules might reveal no significant difference in the effect of the experimental treatment, initially indicating a failure to reject the null hypothesis.

However, a longitudinal analysis of the posttest data for each module might demonstrate a significant effect early in the course sequence that diminished as the course progressed. As initial successes could be expected to beneficially impact academic motivation and early course attrition, such data could be used to justify the use of the treatment only in initial modules—a finding that would have been overlooked in a conventional summative comparison.

An inter-module analysis of the data for Groups B and C is also of interest. Following completion of Modules 1 and 2, Groups B and C might demonstrate significant effect differences. If the differences in these same groups following completion of Modules 3 and 4 were *similar* but *reversed*, and this trend continued in an alternating dyad sequence, the strength of the hypothesis in favor of a positive treatment effect would be increased.

Another inter-module analysis might examine the data within any of the four conditions following the completion of each module. This analysis would focus on the trends established by the data. For example, the slope of a graph representing post-module achievement scores for Group A might be compared with the slope for Group B to examine the achievement rate over time and any possible cumulative effects of the treatment, whether positive or negative.

Cross-Sectional Analyses Within Modules. Controlling for variables resulting from the background data, the SOP design enables cross-sectional analyses within modules to detect indications of threats to internal validity and anomalous results. Modular anomalies can be isolated and examined further to determine possible causes. For example, if the treatment were to be effective only with certain modules and not others, additional qualitative analyses could be performed to develop possible explanations, such as unforeseen interactions between the content specific to those modules and the treatment (e.g. assessment for a particular module emphasizes non-linguistic information, thereby reducing the effectiveness of the language-oriented treatment). Further, other threats to internal validity, such as maturation or interference from history, might be identified.

Because the course serves both instructional and experimental purposes, external validity is maximized, in that the content, modular sequence, technology, and assessments are consistent in both actual and experimental use. Additionally, participants take part in the course on an individual basis at a distance via the Internet and therefore, have no contact with one another, which protects against the diffusion of treatment threat. Further, subjects need not be aware of, or influenced by, the experimental nature of the course.

Conclusion

Distance learning courses provide a unique opportunity for research, where the internal validity of the laboratory and the external validity and generalizability of the field can be called upon to support the conclusions of instructional treatment studies. The inherently useful state of modularity common to these courses provides a practical tool for isolating treatments under randomized and controlled conditions. In particular, the Strand of Pearls design facilitates a variety of cross-sectional and longitudinal analyses that aid researchers in establishing robust convergent evidence for the effectiveness of particular treatments. Through such efforts, it is our hope that the "active ingredients" of effective learning can be identified and leveraged to improve both the scientific understanding of human learning and the quality of instruction available to learners across the globe.

References

Banaji, M. R. & Crowder, R. C. (1989). The bankruptcy of everyday memory. American Psychologist, 44, 1185-1193.

Banaji, M. R. & Crowder, R. C. (1991). Some everyday thoughts on ecologically valid methods. American Psychologist, 46(1), 78-79.

Berge, Z. L., & Mrozowski, S. (2001). Review of research in distance education, 1990 to 1999. The American Journal of Distance Education, 15(3), 5-20.

Best, J. W., & Kahn, J. V. (2003). Research in education (9th ed.). Boston, MA: Pearson Education Company.

Brostoff, T., Sinsheimer, A., & Ford, M. (2001). Practice and procedure: English for lawyers: A preparatory course for international lawyers. The Journal of Legal Writing, 7, 137-154.

Caffarella, E. P. (2000). Doctoral dissertation research in educational technology: The themes and trends from 1977 through 1998. Educational Media and Technology Yearbook, 25, 14-25.

Campbell, D. T., & Stanley, J. C. (1963). Experimental and quasi-experimental designs for research on teaching. In N. L. Gage (Ed.), Handbook of research on teaching (pp. 171–246). Chicago, IL: Rand McNally.

Campbell, D. T., & Stanley, J. C. (1966). Experimental and quasi-experimental designs for research. Boston: Houghton Mifflin Company.

Clark, R. E. (1983). Reconsidering research on learning from media. Review of Educational Research, 53(4), 445-459.

Clark, R.E. (1985). Evidence for confounding in computer-based instruction studies: Analyzing the meta-analyses. Educational Communication and Technology Journal, 33(4), 249-262.
Clark, R. E. (Ed.). (2001). Learning from media: Arguments, analysis, and evidence. Greenwich, CT: Information Age.
Clark, R. E., & Estes, F. (1998). Technology or craft: What are we doing? Educational Technology, 38(5), 5-11.
Clark, R. E., & Estes, F. (1999). The development of authentic educational technologies. Educational Technology, 37(2), 5-16.
Clark, R. E., & Estes, F. (2000). Turning research into results: A guide to selecting the right performance solutions. Atlanta, GA: CEP Press.
Clark, R. E., & Snow, R. E. (1975). Alternative designs for instructional technology research. Audiovisual Communication Review, 23(4), 373-394.
Feak, C. B., & Reinhart, S. M. (2002). An ESP program for students of law. In T. Orr (Ed.), English for Specific Purposes (pp. 7-23). Alexandria, VA: TESOL.
Hanigsberg, J. E. (1994). Swimming lessons: An orientation course for foreign graduate students. Journal of Legal Education 44(4), 588-603.
Kerper, J. (2001). Let's space out: Rethinking the design of law school texts. Journal of Legal Education, 51(2), 267-283.
Khan, B. H. (2001). A framework for web-based learning. In B. H. Khan (Ed.), Web Based Training. Englewood Cliffs, NJ: Educational Technology Publications.
Morton, J. (1991). The bankruptcy of everyday thinking. American Psychologist, 46(1), 32-33.
Pedhazur, E. J., & Schmelkin, L. P. (1991). Measurement, design, and analysis: An integrated approach. Hillsdale, NJ: Lawrence Erlbaum Associates.
Ross. S. M., & Morrison, G. R. (2003). Experimental research methods. In D. J. Jonassen (Ed.), Handbook of research on educational communications, and technology: A project of the Association for Educational Communications and Technology (2nd ed.) (pp. 1021-1045). Mahwah, NJ: Lawrence Erlbaum Associates.
Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). Experimental and quasi-experimental designs for generalized causal inference. Boston: Houghton Mifflin Company.
Stake, R. (1998). Case studies. In N. K. Denzin & Y. S. Lincoln (Eds.), Strategies of Qualitative Inquiry (pp. 86-109). Thousand Oaks, CA: Sage.
Winn, W. D. (2002). Current trends in educational technology research: The study of learning environments. Educational Psychology Review, 14(3), 331–351.