·38·

EXPERIMENTAL RESEARCH METHODS

Steven M. Ross
The University of Memphis

Gary R. Morrison
Wayne State University

38.1 EVOLUTION OF EXPERIMENTAL RESEARCH METHODS

Experimental research has had a long tradition in psychology and education. When psychology emerged as an infant science during the 1900s, it modeled its research methods on the established paradigms of the physical sciences, which for centuries relied on experimentation to derive principals and laws. Subsequent reliance on experimental approaches was strengthened by behavioral approaches to psychology and education that predominated during the first half of this century. Thus, usage of experimentation in educational technology over the past 40 years has been influenced by developments in theory and research practices within its parent disciplines.

In this chapter, we examine practices, issues, and trends related to the application of experimental research methods in educational technology. The purpose is to provide readers with sufficient background to understand and evaluate experimental designs encountered in the literature and to identify designs that will effectively address questions of interest in their own research. In an introductory section, we define experimental research, differentiate it from alternative approaches, and identify important concepts in its use (e.g., internal vs. external validity). We also suggest procedures for conducting experimental studies and publishing them in educational technology research journals. Next, we analyze uses of experimental methods by instructional researchers, extending the analyses of three decades ago by Clark and Snow (1975). In the concluding section, we turn to issues in using experimental research in educational technology, to include balancing internal and external validity, using multiple outcome measures to assess learning processes and products, using item responses vs. aggregate scores as dependent variables, reporting effect size as a complement to statistical significance, and media replications vs. media comparisons.

38.2 WHAT IS EXPERIMENTAL RESEARCH?

The experimental method formally surfaced in educational psychology around the turn of the century, with the classic studies by Thorndike and Woodworth on transfer (Cronbach, 1957). The experimenter's interest in the effect of environmental change, referred to as "treatments," demanded designs using standardized procedures to hold all conditions constant except the independent (experimental) variable. This standardization ensured high internal validity (experimental control) in comparing the experimental group to the control group on the dependent or "outcome" variable. That is, when internal validity was high, differences between groups could be confidently attributed to the treatment, thus ruling out rival hypotheses attributing effects to extraneous factors. Traditionally, experimenters have given less emphasis to external validity, which concerns the generalizability of findings to other settings, particularly realistic ones. One theme of this chapter is that current orientations in instructional theory and research practices necessitate achieving a better balance between internal and external validity levels.

During the past century, the experimental method has remained immune to paradigm shifts in the psychology of learning, including behaviorism to cognitivism, objectivism to

cognitivism, and instructivism to constructivism (see Jonassen, 1991; Jonassen, Campbell, & Davidson, 1994). Clearly, the logical positivism of behavioristic theory created a fertile, inviting framework for attempts to establish causal relationships between variables, using experimental methods. The emergence of cognitive learning theory in the 1970s and 1980s initially did little to change this view, as researchers changed the locus of inquiry from behavior to mental processing but maintained the experimental method as the basic way they searched for scientific truths. Today, the increasing influences of constructivist theories are making the fit between traditional scientific methods and current perspectives on learning more difficult. As Jonassen et al. (1994) state, it is now viewed as much more difficult "... to isolate which components of the learning system, the medium, the attributes, the learner, or the environment affect learning and in what ways" (p. 6). Accordingly, without knowing the ultimate impact or longevity of the constructivist view, we acknowledge its contribution in conveying instruction and learning as less orderly than preceding paradigms had depicted and the learner rather than the "treatment" as deserving more importance in the study of learning processes. Our perspective in this chapter, therefore, is to present experimental methods as continuing to provide valuable "tools" for research but ones whose uses may need to be altered or expanded relative to their traditional functions to accommodate the changing complexion of theory and scientific inquiry in instructional technology.

38.2.1 Types of Experimental Designs

Complete descriptions of alternative experimental designs are provided in Campbell and Stanley (1963) and conventional research textbooks (e.g., Borg, Gall, & Gall, 1993; Creswell, 2002; Gliner & Morgan, 2000). For purposes of providing common background for the present chapter, we have selected four major design approaches to review. These particular designs appeared to be the ones instructional technology researchers would be most likely to use for experimental studies or find in the literature. They are also "core" designs in the sense of including basic components of the more complex or related designs not covered.

38.2.1.1 True Experiments. The ideal design for maximizing internal validity is the true experiment, as diagrammed below. The R means that subjects were randomly assigned, X represents the treatment (in this case, alternative treatments 1 and 2), and O means observation (or outcome), for example, a dependent measure of learning or attitude. What distinguishes the true experiment from less powerful designs is the random assignment of subjects to treatments, thereby eliminating any systematic error that might be associated with using intact groups. The two (or more) groups are then subjected to identical environmental conditions, while being exposed to different treatments. In educational technology research, such

treatments frequently consist of different instructional methods (discussed later).

$$X_1 O$$
$$R \quad X_2 O$$

Example. An example of a true experiment involving an educational technology application is the study by Clariana and Lee (2001) on the use of different types of feedback in computer-delivered instruction. Graduate students were randomly assigned to one of five feedback treatments, approximately 25 subjects per group, comprised of (a) a constructed-response (fill-in-the-blank) study task with feedback and recognition (multiple-choice) tasks with (b) single-try feedback, (c) multiple-response feedback, (d) single-try feedback with overt responding, and (e) multiple-try feedback with overt responding. All subjects were treated identically, with the exception of the manipulation of the assigned feedback treatment. The major outcome variable (observation) was a constructed-response achievement test on the lesson material. Findings favored the recognition-study treatments with feedback followed by overt responding. Given the true experimental design employed, the authors could infer that the learning advantages obtained were due to properties of the overt responding (namely, in their opinion, that it best matched the posttest measure of learning) rather than extraneous factors relating to the lesson, environment, or instructional delivery. In research parlance, "causal" inferences can be made regarding the effects of the independent (manipulated) variable (in this case, type of feedback strategy) on the dependent (outcome) variable (in this case, degree of learning).

38.2.1.2 Repeated Measures. A variation of the above experimental design is the situation where all treatments (X_1, X_2, etc.) are administered to all subjects. Thus, each individual (S1, S2, etc.), in essence, serves as his or her own control and is tested or "observed" (O), as diagrammed below for an experiment using n subjects and k treatments. Note that the diagram shows each subject receiving the same sequence of treatments; a stronger design, where feasible, would involve randomly ordering the treatments to eliminate a sequence effect.

$$S1: \quad X_1 0 - X_2 0 \dots X_k O.$$
$$S2: \quad X_1 0 - X_2 0 \dots X_k O.$$
$$Sn: \quad X_1 0 - X_2 0 \dots X_k O.$$

Suppose that an experimenter is interested in whether learners are more likely to remember words that are italicized or words that are underlined in a computer text presentation. Twenty subjects read a paragraph containing five words in each form. They are then asked to list as many italicized words and as many underlined words as they can remember. (To reduce bias, the forms in which the 10 words are represented are randomly varied for different subjects.) Note that this design has the advantage of using only one group, thereby effectively doubling the number of subjects per treatment relative to a two-group (italics

only vs. underline only) design. It also ensures that the ability level of subjects receiving the two treatments will be the same. But there is a possible disadvantage that may distort results. The observations are not independent. Recalling an italicized word may help or hinder the recall of an underlined word, or vice versa.

Example. An example of a repeated-measures design is the recent study by Gerlic and Jausovec (1999) on the mental effort induced by information present in multimedia and text formats. Three presentation formats (text only, text/sound/video, text/sound/picture) were presented in randomly determined orders to 38 subjects. Brain wave activity while learning the material was recorded as electroencephalographic (EEG) data. Findings supported the assumption that the video and picture presentations induced visualization strategies, whereas the text presentation generated mainly processes related to verbal processing. Again, by using the repeated-measures design, the researchers were able to reduce the number of subjects needed while controlling for individual differences across the alternative presentation modes. That is, every presentation mode was administered to the identical samples. But the disadvantage was the possible "diffusion" of treatment effects caused by earlier experiences with other modes. We will return to diffusion effects, along with other internal validity threats, in a later section.

38.2.1.3 *Quasi-experimental Designs.* Oftentimes in educational studies, it is neither practical nor feasible to assign subjects randomly to treatments. Such is especially likely to occur in school-based research, where classes are formed at the start of the year. These circumstances preclude true-experimental designs, while allowing the quasi-experiment as an option. A common application in educational technology would be to expose two similar classes of students to alternative instructional strategies and compare them on designated dependent measures (e.g., learning, attitude, classroom behavior) during the year.

An important component of the quasi-experimental study is the use of pretesting or analysis of prior achievement to establish group equivalence. Whereas in the true experiment, randomization makes it improbable that one group will be significantly superior in ability to another, in the quasi-experiment, systematic bias can easily (but often unnoticeably) be introduced. For example, although the first- and third-period algebra classes may have the same teacher and identical lessons, it may be the case that honors English is offered third period only, thus restricting those honors students to taking first-period algebra. The quasi-experiment is represented diagrammatically as follows. Note its similarity to the true experiment, with the omission of the randomization component. That is, the Xs and Os show treatments and outcomes, respectively, but there are no Rs to indicate random assignment (by jose at tf)

$$X_1 O$$

$$X_2 O$$

Example. Use of a quasi-experimental design is reflected in a recent study by the present authors on the long-term effects

of computer experiences by elementary students (Ross, Smith, & Morrison, 1991). During their fifth- and sixth-grade years, one class of students at an inner-city school received classroom and home computers as part of a computer-intensive learning program sponsored by Apple Classrooms of Tomorrow (ACOT). A class of similar students, who were exposed to the same curriculum but without computer support, was designated to serve as the control group. To ensure comparability of groups, scores on all subtests of the California Achievement Test (CAT), administered before the ACOT program was initiated, were analyzed as pretests; no class differences were indicated. The Ross et al. (1991) study was designed to find members of the two cohorts and evaluate their adjustment and progress in the seventh-grade year, when, as junior-high students, they were no longer participating in ACOT.

Although many more similarities than differences were found, the ACOT group was significantly superior to the control group on CAT mathematics. Can this advantage be attributed to their ACOT experiences? Perhaps, but in view of the quasi-experimental design employed, this interpretation would need to be made cautiously. Not only was "differential selection" of subjects a validity threat, so was the "history effect" of having each class taught in a separate room by a different teacher during each program year. Quasi-experimental designs have the advantage of convenience and practicality but the disadvantage of reduced internal validity.

38.2.1.4 *Time Series Design.* Another type of quasi-experimental approach is time series designs. This family of designs involves repeated measurement of a group, with the experimental treatment induced between two of the measures. Why is this a quasi-experiment as opposed to a true experiment? The absence of randomly composed, separate experimental and control groups makes it impossible to attribute changes in the dependent measure directly to the effects of the experimental treatment. That is, the individual group participating in the time series design may improve its performances from pretesting to posttesting, but is it the treatment or some other event that produced the change? There is a variety of time series designs, some of which provide a higher internal validity than others.

A single-group time series design can be diagrammed as shown below. As depicted, one group (G) is observed (O) several times prior to receiving the treatment (X) and following the treatment.

$$G \quad O_1 \quad O_2 \quad O_3 \quad X \quad O_4 \quad O_5$$

To illustrate, suppose that we assess on 3 successive days the percentage of students in a class who successfully complete individualized computer-based instructional units. Prior to the fourth day, teams are formed and students are given additional team rewards for completing the units. Performance is then monitored on days 4 and 5. If performance increases relative to the pretreatment phase (days 1 to 3), we may infer that the CBI units contributed to that effect. Lacking a true-experimental design, we make that interpretation with some element of caution.

A variation of the time series design is the single-subject study, in which one individual is examined before and after the introduction of the experimental treatment. The simplest form is the A–B design, where A is the baseline (no treatment) period and B is the treatment. A potentially stronger variation is the A–B–A design, which adds a withdrawal phase following the treatment. Each new phase (A or B) added to the design provides further data to strengthen conclusions about the treatment's impact. On the other hand, each phase may inherit cumulative contaminating effects from prior phases. That is, once B is experienced, subsequent reactions to A and B may be directly altered as a consequence.

Example. An example of a time series design is the study by Alper, Thoresen, and Wright (1972), as described by Clark and Snow (1975). The focus was the effects of a videotape on increasing a teacher's positive attention to appropriate student behavior and decreasing negative responses to inappropriate behavior. Baseline data were collected from a teacher at two times: (a) prior to the presentation of the video and feedback on ignoring inappropriate behavior and (b) prior to the video and feedback on attending to positive behavior. Teacher attention was then assessed at different points following the video modeling and feedback. Interestingly, the analysis revealed that, although the teacher's behavior changed in the predicted directions following the video–feedback interventions, undesirable behavior tended to reappear over time. The time series design, therefore, was especially apt for detecting the unstable behavior pattern. We see relatively few time series designs in the current research literature. Perhaps one reason is that "human subjects" criteria would generally discourage subjecting individuals to prolonged involvement in a study and to repeated assessments.

38.2.1.5 Deceptive Appearances: The Ex Post Facto Design.
Suppose that in reviewing a manuscript for a journal, you come across the following study that the author describes as quasi-experimental (or experimental). The basic design involves giving a class of 100 college educational psychology students the option of using a word processor or paper and pencil to take notes during three full-period lectures on the topic of cognitive theory. Of those who opt for the two media (say, 55 for the word processor and 45 for paper and pencil), 40 from each group are randomly selected for the study. Over the 3 days, their notes are collected, and daily quizzes on the material are evaluated. Results show that the word processor group writes a greater quantity of notes and scores higher on the quizzes.

Despite the appearances of a treatment comparison and random assignment, this research is not an experiment but rather an ex post facto study. No variables are manipulated. Existing groups that are essentially self-selected are being compared: those who chose the word processor vs. those who chose paper and pencil. The random selection merely reduced the number of possible participants to more manageable numbers; it did not assign students to particular treatments. Given these properties, the ex post facto study may look sometimes like an experiment but is closer in design to a correlational study. In our example, the results imply that using a word processor is related to better performance. But a causal interpretation cannot be made, be-

cause other factors could just as easily have accounted for the outcomes (e.g., brighter or more motivated students may have been more likely to select the word-processing option).

38.2.2 Validity Threats

As has been described, internal validity is the degree to which the design of an experiment controls extraneous variables (Borg et al., 1993). For example, suppose that a researcher compares the achievement scores of students who are asked to write elaborations on a computer-based instruction (CBI) lesson vs. those who do not write elaborations on the same lesson. If findings indicate that the elaborations group scored significantly higher on a mastery test than the control group, the implication would be that the elaborations strategy was effective. But what if students in the elaborations group were given more information about how to study the material than were control students? This extraneous variable (i.e., additional information) would weaken the internal validity and the ability to infer causality.

When conducting experiments, instructional technology researchers need to be aware of potential internal validity threats. In 1963, Campbell and Stanley identified different classes of such threats. We briefly describe each below, using an illustration relevant to educational technology interests.

38.2.2.1 History.
This validity threat is present when events, other than the treatments, occurring during the experimental period can influence results.

Example. A researcher investigates the effect of using cooperative learning (treatment) vs. individual learning (control) in CBI. Students from a given class are randomly assigned to different laboratory rooms where they learn either cooperatively or individually. During the period of the study, however, the regular teacher begins to use cooperative learning with all students. Consequently, the control group feels frustrated that, during the CBI activity, they have to work alone. Due to their "history," with cooperative learning, the control group's perceptions were altered.

38.2.2.2 Maturation.
During the experimental period, physical or psychological changes take place within the subjects.

Example. First-grade students receive two types of instruction in learning to use a mouse in operating a computer. One group is given active practice, and the other group observes a skilled model followed by limited practice. At the beginning of the year, neither group performs well. At the end of the year, however, both substantially improve to a comparable level. The researcher (ignoring the fact that students became more dexterous, as well as benefiting from the training) concluded that both treatments were equally effective.

38.2.2.3 Testing.
Exposure to a pretest or intervening assessment influences performance on a posttest.

Example. A researcher who is interested in determining the effects of using animation vs. static graphics in a CBI lesson pretests two randomly composed groups of high-school students on the content of the lesson. Both groups average close

to 55% correct. One of the groups then receives animation, and the other the static graphics on their respective lessons. At the conclusion of the lesson, all students complete a posttest that is nearly identical to the pretest. No treatment differences, however, are found, with both groups averaging close to 90% correct. Students report that the pretest gave them valuable cues about what to study.

38.2.2.4 Instrumentation.
Inconsistent use is made of testing instruments or testing conditions, or the pretest and posttest are uneven in difficulty, suggesting a gain or decline in performance that is not real.

Example. An experiment is designed to test two procedures for teaching students to write nonlinear stories (i.e., stories with branches) using hypermedia. Randomly composed groups of eighth graders learn from a modeling method or a direct instruction method and are then judged by raters on the basis of the complexity and quality of a writing sample they produce. The "modeling" group completes the criterion task in their regular writing laboratory, whereas the "direct instruction" group completes it on similar computers, at the same day and time, but in the journalism room at the local university. Results show significantly superior ratings for the modeling group. In fact, both groups were fairly comparable in skills, but the modeling group had the advantage of performing the criterion task in familiar surroundings.

38.2.2.5 Statistical Regression.
Subjects who score very high or very low on a dependent measure naturally tend to score closer (i.e., regress) to the mean during retesting.

Example. A researcher is interested in the effects of learning programming on the problem-solving skills of high-ability children. A group of 400 sixth graders is pretested on a problem-solving test. The 50 highest scorers are selected and randomly assigned to two groups of 25 each. One group learns programming during the semester, whereas the other learns a spreadsheet application. At the end of the year, the students are posttested on the same problem-solving measure. There are no differences between them; in fact, the means for both groups are actually slightly lower than they were on the pretest. These very high scorers on the pretest had regressed to the mean (due, perhaps, to not having as "good of a day" on the second testing).

38.2.2.6 Selection.
There is a systematic difference in subjects' abilities or characteristics between the treatment groups being compared.

Example. Students in the fourth-period American history class use an electronic encyclopedia during the year as a reference for historical events, whereas those in the sixth-period class use a conventional encyclopedia. The two classes have nearly identical grade point averages and are taught by the same teacher using the exact same materials and curriculum. Comparisons are made between the classes on the frequency with which they use their respective encyclopedias and the quality of the information they select for their reports. The control group is determined to be superior on both of these variables. Further examination of student demographics, however, shows that a much greater percentage of the control students are in advanced placement (AP) courses in English, mathematics, and science. In fact, the reason many were scheduled to take history sixth period was to avoid conflicts with AP offerings. Differential selection therefore resulted in higher-achieving students being members of the control group.

38.2.2.7 Experimental Mortality.
The loss of subjects from one or more treatments during the period of the study may bias the results.

Example. An instructional designer is interested in evaluating a college-level CBI algebra course that uses two learning orientations. One orientation allows the learner to select menu and instructional support options (learner-control treatment); the other prescribes particular options based on what is considered best for "typical" learners (program-control treatment). At the beginning of the semester, 40 students are assigned to each treatment and begin work with the corresponding CBI programs. At the end of the semester, only 50 students remain in the course, 35 in the learner-control group and 15 in the program-control group. Achievement results favor the program-control group. The greater "mortality" in the program-control group probably left a higher proportion of more motivated or more capable learners than in the learner-control group.

38.2.2.8 Diffusion of Treatments.
The implementation of a particular treatment influences subjects in the comparison treatment.

Example. A researcher is interested in examining the influences on attitudes and achievement of fifth graders' writing to pen pals via electronic mail. Half the students are assigned pen pals; the other half complete the identical assignments on the same electronic mail system but send the letters to "fictitious" friends. The students in the latter group, however, become aware that the other group has real pen pals and feel resentful. On the attitude measure, their reactions toward the writing activities are very negative as a consequence. By learning about the experimental group's "treatment," the perceptions and attitudes of the control group were negatively influenced.

38.2.3 Dealing With Validity Threats

In many instances, validity threats cannot be avoided. The presence of a validity threat should not be taken to mean that experimental findings are inaccurate or misleading. By validity "threat," we mean only that a factor has the potential to bias results. Knowing about validity threats gives the experimenter a framework for evaluating the particular situation and making a judgment about its severity. Such knowledge may also permit actions to be taken to limit the influences of the validity threat in question. Examples are as follows:

- Concern that a pretest may bias posttest results leads to the decision not to use a pretest.
- Concern that the two intact groups to be used for treatment comparisons (quasi-experimental design) may not be equal in

ability leads to the decision to pretest subjects on ability and employ a statistical adjustment (analysis of covariance) if the groups significantly differ.

- Concern that subjects may mature or drop out during the period of the experiment leads to the decision to shorten the length of the treatment period, use different types of subjects, and/or introduce noncontaminating conditions (e.g., incentives) to reduce attrition.
- Concern that the posttest may differ in difficulty from the pretest in an experiment design to assess learning gain leads to the decision to use each test form as the pretest for half the students and the posttest for the other half.
- Concern about the artificiality of using abstract symbols such as Xs and Os as the stimulus material for assessing computer screen designs leads to the addition of "realistic" nonsense words and actual words as supplementary treatments.
- Concern that subjects might not be motivated to perform on an experimental task leads to the development of an actual unit of instruction that becomes an alternative form of instruction for the students in a class.

Even after all reasonable actions have been taken to eliminate the operation of one or more validity threats, the experimenter must still make a judgment about the internal validity of the experiment overall. In certain cases, the combined effects of multiple validity threats may be considered inconsequential, whereas in others, the effects of a single threat (e.g., differential sample selection) may be severe enough to preclude meaningful results. When the latter occurs, the experiment needs to be redone. In cases less severe, experimenters have the obligation to note the validity threats and qualify their interpretations of results accordingly.

38.3 THE PRACTICE OF EXPERIMENTATION IN EDUCATIONAL TECHNOLOGY

38.3.1 How to Conduct Experimental Studies: A Brief Course

For the novice researcher, it is often difficult to get started in designing and conducting experimental studies. Seemingly, a common problem is putting the cart before the horse, which in typical cases translates into selecting methodology or a research design before deciding what questions to investigate. Research questions, along with practical constraints (time and resources), should normally dictate what type of study to do, rather than the reverse. To help readers avoid such problems, we have devised the following seven-step model, which presents a sequence of logical steps for planning and conducting research (Ross & Morrison, 1992, 1993, 2001). The model begins at a level where the individual is interested in conducting research (such as for a dissertation or scholarly activity) but has not even identified a topic. More advanced researchers would naturally start at the level appropriate to their needs. To illustrate the various steps, we discuss our recent experiences in designing a

research study on applications of an interactive computer-based chemistry unit.

38.3.1.1 Step 1. Select a Topic. This step is self-explanatory and usually not a problem, except for those who are "required" to do research (e.g., as part of an academic degree program) as opposed to initiating it on their own. The step simply involves identifying a general area that is of personal interest (e.g., learner control, picture perception, mathematics learning) and then narrowing the focus to a researchable problem (step 2).

Chemistry CBI Example. In our situation, Gary Morrison received a grant from FIPSE to develop and evaluate interactive chemistry units. We thus had the interest in as well as the formal responsibility of investigating how the completed units operated.

38.3.1.2 Step 2. Identify the Research Problem. Given the general topic area, what specific problems are of interest? In many cases, the researcher already knows the problems. In others, a trip to the library to read background literature and examine previous studies is probably needed. A key concern is the importance of the problem to the field. Conducting research requires too much time and effort to be examining trivial questions that do not expand existing knowledge. Experienced researchers will usually be attuned to important topics, based on their knowledge of the literature and current research activities. Novices, however, need to be more careful about establishing support for their idea from recent research and issues-oriented publications (see step 3). For experts and novices alike, it is always a good practice to use other researchers as a sounding board for a research focus before getting too far into the study design (steps 4 and 5).

Chemistry CBI Example. The topic and the research problem were presented to us through the objectives of the FIPSE grant and our interest in assessing the "effectiveness" of the completed CBI chemistry units. The research topic was "CBI usage in teaching college chemistry courses"; the research problem was "how effectively interactive CBI units on different chemistry concepts would teach those concepts." Later, this "problem" was narrowed to an examination of the influences on student learning and attitudes of selected features of a specific CBI unit, Gas Laws.

38.3.1.3 Step 3. Conduct a Literature Search. With the research topic and problem identified, it is now time to conduct a more intensive literature search. Of importance is determining what relevant studies have been performed; the designs, instruments, and procedures employed in those studies; and, most critically, the findings. Based on the review, direction will be provided for (a) how to extend or compliment the existing literature base, (b) possible research orientations to use, and (c) specific research questions to address. Helpful information about how to conduct effective literature reviews is provided in other sources (e.g., Borg et al., 1993; Creswell, 2002; Ross & Morrison, 2001).

Chemistry CBI Example. For the chemistry study, the literature proved important in two ways. First, it provided general background information on related studies in the content area

(chemistry) and in CBI applications in general. Second, in considering the many specific features of the chemistry unit that interested us (e.g., usage of color, animation, prediction, elaboration, self-pacing, learner control, active problem solving), the literature review helped to narrow our focus to a restricted, more manageable number of variables and gave us ideas for how the selected set might be simultaneously examined in a study.

38.3.1.4 Step 4. State the Research Questions (or Hypotheses).

This step is probably the most critical part of the planning process. Once stated, the research questions or hypotheses provide the basis for planning all other parts of the study: design, materials, and data analysis. In particular, this step will guide the researcher's decision as to whether an experimental design or some other orientation is the best choice.

For example, in investigating uses of learner control in a math lesson, the researcher must ask what questions he or she really wants to answer. Consider a question such as, How well do learners like using learner control with math lessons? To answer it, an experiment is hardly needed or even appropriate. A much better choice would be a descriptive study in which learners are interviewed, surveyed, and/or observed relative to the activities of concern. In general, if a research question involves determining the "effects" or "influences" of one variable (independent) on another (dependent), use of an experimental design is implied.

Chemistry CB1 Example. The questions of greatest interest to us concerned the effects on learning of (a) animated vs. static graphics, (b) learners predicting outcomes of experiments vs. not making predictions, and (c) learner control vs. program control. The variables concerned were expected to operate in certain ways based on theoretical assumptions and prior empirical support. Accordingly, *hypotheses* such as the following were suggested: "Students who receive animated graphics will perform better on problem-solving tasks than do students who receive static graphics," and "Low achievers will learn less effectively under learner control than program control." Where we felt less confident about predictions or where the interest was descriptive findings, research *questions* were implied: "Would students receiving animated graphics react more positively to the unit than those receiving static graphics?" and "To what extent would learner-control students make use of opportunities for experimenting in the 'lab'?"

38.3.1.5 Step 5. Determine the Research Design.

The next consideration is whether an experimental design is feasible. If not, the researcher will need to consider alternative approaches, recognizing that the original research question may not be answerable as a result. For example, suppose that the research question is to determine the effects of students watching CNN on their knowledge of current events. In planning the experiment, the researcher becomes aware that no control group will be available, as all classrooms to which she has access receive the CNN broadcasts. Whereas an experimental study is implied by the original "cause–effect" question, a descriptive study examining current events scores (perhaps from pretest to posttest) will probably be the most reasonable option. This design may provide some interesting food for thought on the possible effects of CNN on current events learning, but it cannot validly answer the original question.

Chemistry CBI Example. Our hypotheses and research questions implied both experimental and descriptive designs. Specifically, hypotheses concerning the effects of animated vs. static graphics and between prediction vs. no prediction implied controlled experimental comparisons between appropriate treatment conditions. Decisions needed to be made about which treatments to manipulate and how to combine them (e.g., a factorial or balanced design vs. selected treatments). We decided on selected treatments representing targeted conditions of interest. For example, we excluded static graphics with no prediction, as that treatment would have appeared awkward given the way the CBI program was designed, and we had little interest in it for applied evaluation purposes. Because subjects could be randomly assigned to treatments, we decided to use a true-experimental design.

Other research questions, however, implied additional designs. Specifically, comparisons between high and low achievers (in usage of CBI options and relative success in different treatments) required an ex post facto design, because members of these groups would be identified on the basis of existing characteristics. Research questions regarding usage of learner control options would further be examined via a descriptive approach.

38.3.1.6 Step 6. Determine Methods.

Methods of the study include (a) subjects, (b) materials and data collection instruments, and (c) procedures. In determining these components, the researcher must continually use the research questions and/or hypotheses as reference points. A good place to start is with subjects or participants. What kind and how many participants does the research design require? (See, e.g., Glass & Hopkins, 1984, p. 213, for a discussion of sample size and power.) Next consider materials and instrumentation. When the needed resources are not obvious, a good strategy is to construct a listing of data collection instruments needed to answer each question (e.g., attitude survey, achievement test, observation form).

An experiment does not require having access to instruments that are already developed. Particularly in research with new technologies, the creation of novel measures of affect or performance may be implied. From an efficiency standpoint, however, the researcher's first step should be to conduct a thorough search of existing instruments to determine if any can be used in their original form or adapted to present needs. If none is found, it would usually be far more advisable to construct a new instrument rather than "force fit" an existing one. New instruments will need to be pilot tested and validated. Standard test and measurement texts provide useful guidance for this requirement (e.g., Gronlund & Linn, 1990; Popham, 1990). The experimental procedure, then, will be dictated by the research questions and the available resources. Piloting the methodology is essential to ensure that materials and methods work as planned.

Chemistry CB1 Example. Our instructional material consisted of the CBI unit itself. Hypotheses and research questions implied developing alternative forms of instruction (e.g., animation–prediction, animation–no prediction,

static–prediction) to compare, as well as original (new) data collection instruments because the instructional content was unit-specific. These instruments included an achievement test, attitude survey, and on-line assessments for recording of lesson option usage (e.g., number of lab experiments selected), learning time, and predictions.

38.3.1.7 Step 7. Determine Data Analysis Techniques.

Whereas statistical analysis procedures vary widely in complexity, the appropriate options for a particular experiment will be defined by two factors: the research questions and the type of data. For example, a t test for independent samples would be implied for comparing one experimental group (e.g., CBI with animation) to one control group (CBI with static graphics) on an interval-dependent measure (e.g., performance on a problem-solving test). Add a third treatment group (CBI without graphics), and a one-way analysis of variance (ANOVA) would be implied for the same interval data, but now comparing more than two means. If an additional outcome measure were a categorical response on, say, an attitude survey ("liked the lesson" or "didn't like it"), a chi-square analysis would be implied for determining the relationship between treatment and response on the resultant *nominal* data obtained.

Educational technology experimenters do not have to be statisticians. Nor do they have to set analytical procedures in stone prior to completing the research. Clearly formulated research questions and design specifications will provide a solid foundation for working with a statistician (if needed) to select and run appropriate analyses. To provide a convenient guide for considering alternative analysis, Table 38.1 lists common statistical analysis procedures and the main conditions under which they are used. Note that in assessing causal relationships, experiments depend on analysis approaches that compare outcomes associated with treatments (nominal or categorical variables) such as t tests, ANOVA, analysis of covariance, and chi-square, rather than correlational-type approaches.

38.3.2 Reporting and Publishing Experimental Studies

Obviously, for experimental studies to have impact on theory and practice in educational technology, their findings need to be disseminated to the field. Thus, part of the experimenter's role is publishing research in professional journals and presenting it at professional meetings. Discussing these activities in any detail is beyond the scope of the present chapter; also, articles devoted to these subjects can be found elsewhere (e.g., Ross & Morrison, 1991, 1993, 2001; Thyer, 1994). However, given the special features and style conventions of experimental reports compared to other types of educational technology literature, we consider it relevant to review the former, with a specific concentration on journal publications. It is through referred journals—such as *Performance Improvement Quarterly*, and *Educational Technology Research and Development*—that experimental studies are most likely to be disseminated to members of the educational

technology field. The following is a brief description of each major section of the paper.

38.3.2.1 Introduction.

The introduction to reports of experimental studies accomplishes several functions: (a) identifying the general area of the problem (e.g., CBI or cooperative learning), (b) creating a rationale to learn more about the problem (otherwise, why do more research in this area?), (c) reviewing relevant literature, and (d) stating the specific purposes of the study. Hypotheses and/or research questions should directly follow from the preceding discussion and generally be stated explicitly, even though they may be obvious from the literature review. In basic research experiments, usage of hypotheses is usually expected, as a theory or principle is typically being tested. In applied research experiments, hypotheses would be used where there is a logical or empirical basis for expecting a certain result (e.g., "The feedback group will perform better than the no-feedback group"); otherwise, research questions might be preferable (e.g., "Are worked examples more effective than incomplete examples on the CBI math unit developed?").

38.3.2.2 Method.

The *Method* section of an experiment describes the participants or subjects, materials, and procedures. The usual convention is to start with *subjects* (or participants) by clearly describing the population concerned (e.g., age or grade level, background) and the sampling procedure. In reading about an experiment, it is extremely important to know if subjects were randomly assigned to treatments or if intact groups were employed. It is also important to know if participation was voluntary or required and whether the level of performance on the experimental task was consequential to the subjects.

Learner motivation and task investment are critical in educational technology research, because such variables are likely to impact directly on subjects' usage of media attributes and instructional strategies (see Morrison, Ross, Gopalakrishnan, & Casey, 1995; Song & Keller, 2001). For example, when learning from a CBI lesson is perceived as part of an experiment rather than actual course, a volunteer subject may be concerned primarily with completing the material as quickly as possible and, therefore, not select any optional instructional support features. In contrast, subjects who were completing the lesson for a grade would probably be motivated to take advantage of those options. A given treatment variable (e.g., learner control or elaborated feedback) could therefore take very different forms and have different effects in the two experiments.

Once subjects are described, the type of design employed (e.g., quasi-experiment, true experiment) should be indicated. Both the independent and the dependent variables also need to be identified.

Materials and *instrumentation* are covered next. A frequent limitation of descriptions of educational technology experiments is lack of information on the learning task and the context in which it was delivered. Since media attributes can impact learning and performance in unique ways (see Clark, 1983, 2001; 1994; Kozma, 1991, 1994; Ullmer, 1994), their full

TABLE 38.1. Common Statistical Analysis Procedures Used in Educational Technology Research

Analysis	Types of Data	Features	Example	Test of Causal Effects?
t test Independent samples	Independent variable = nominal; dependent = one interval-ratio measure	Tests the differences between 2 treatment groups	Does the problem-based treatment group surpass the traditional instruction treatment group?	Yes
t test Dependent samples	Independent variable = nominal (repeated measure); dependent = one interval-ratio measure	Tests the difference between 2 treatment means for a *given group*	Will participants change their attitudes toward drugs, from pretest to posttest, following a videotape on drug effects?	Yes
Analysis of variance (ANOVA)	Independent variable = nominal; dependent = one interval-ratio measure	Tests the differences between 3 or more treatment means. If ANOVA is significant, follow-up comparisons of means are performed.	Will there be differences in learning among three groups that paraphrase, summarize, or neither?	Yes
Multivariate analysis of variance (MANOVA)	Independent variable = nominal; dependent = two or more interval-ratio measures	Tests the difference between 2 or more treatment group means on 2 or more learning measures. Controls Type I error rate across the measure. If MANOVA is significant, an ANOVA on each individual measure is performed.	Will there be differences among 3 feedback strategies on problem solving and knowledge learning?	Yes
Analysis of covariance (ANCOVA) or multivariate analysis of covariance (MANCOVA)	Independent variable = nominal; dependent = one or more interval-ratio measures; covariate = one or more measures	Replicates ANOVA or MANOVA but employs an additional variable to control for treatment group differences in aptitude and/or to reduce error variance in the dependent variable(s)	Will there be differences in concept learning among learner-control, program-control, and advisement strategies, with differences in prior knowledge controlled?	Yes
Pearson *r*	Two ordinal or interval-ratio measures	Tests relationship between two variables	Is anxiety related to test performance?	No
Multiple linear regression	Independent variable = two or more ordinal or interval-ratio measures; dependent = one ordinal or interval-ratio measure	Tests relationship between set of predictor (independent) variables and outcome variable. Shows the relative contribution of each predictor in accounting for variability in the outcome variable.	How well do experience, age, gender, and grade point average predict time spent on completing a task?	No
Discriminant analysis	Nominal variable (groups) and 2 or more ordinal or interval-ratio variables	Tests relationship between a set of predictor variables and subjects' membership in particular groups	Do students who favor learning from print materials vs. computers vs. television differ with regard to ability, age, and motivation?	No
Chi-square test of independence	Two nominal variables	Tests relationship between two nominal variables	Is there a relationship between gender (male vs. females) and attitudes toward the instruction (liked, no opinion, disliked)?	

description is particularly important to the educational technologist. Knowing only that a "CBI" presentation was compared to a "textbook" presentation suggests the type of senseless media comparison experiment criticized by Clark (1983, 2001) and others (Hagler & Knowlton, 1987; Knowlton, 1964; Morrison, 2001; Ross & Morrison, 1989). In contrast, knowing the specific attributes of the CBI (e.g., animation, immediate feedback, prompting) and textbook presentations permits more meaningful interpretation of results relative to the influences of these attributes on the learning process.

Aside from describing the instructional task, the overall method section should also detail the instruments used for data collection. For illustrative purposes, consider the following excerpts from a highly thorough description of the instructional materials used by Schnackenberg and Sullivan (2000).

The program was developed in four versions that represented the four different treatment conditions. Each of the 13 objectives was taught through a number of screens that present the instruction, practice and feedback, summaries, and reviews. Of the objectives, 9 required selected responses in a multiple-choice format and 4 required constructed responses. The program tracked each participant's response choice on a screen-by-screen basis. (p. 22)

The next main methodology section is the *procedure.* It provides a reasonably detailed description of the steps employed in carrying out the study (e.g., implementing different treatments, distributing materials, observing behaviors, testing). Here, the rule of thumb is to provide sufficient information on what was done to perform the experiment so that another researcher could replicate the study. This section should also provide a time line that describes sequence of the treatments and data collection. For example, the reader should understand that the attitude survey was administered *after* the subjects completed the treatment and *before* they completed the posttest.

38.3.2.3 Results.
This major section describes the analyses and the findings. Typically, it should be organized such that the most important dependent measures are reported first. Tables and/or figures should be used judiciously to supplement (not repeat) the text.

Statistical significance vs. practical importance. Traditionally, researchers followed the convention of determining the "importance" of findings based on statistical significance. Simply put, if the experimental group's mean of 85% on the posttest was found to be significantly higher (say, at $p < .01$) than the control group's mean of 80%, then the "effect" was regarded as having theoretical or practical value. If the result was not significant (i.e., the null hypothesis could not be rejected), the effect was dismissed as not reliable or important.

In recent years, however, considerable attention has been given to the benefits of distinguishing between "statistical significance" and "practical importance" (Thompson, 1998). Statistical significance indicates whether an effect can be considered attributable to factors other than chance. But a significant effect does not necessary mean a "large" effect. Consider this example:

Suppose that 342 students who were randomly selected to participate in a Web-based writing skills course averaged 3.3 (out of 5.0) on the state assessment of writing skills. The 355 students in the control group, however, averaged 3.1, which, due to the large sample sizes, was significantly lower than the experimental group mean, at $p = .032$. Would you advocate the Web-based course as a means of increasing writing skill? Perhaps, but the findings basically indicate a "reliable but small" effect. If improving writing skill is a priority goal, the Web-based course might not be the most effective and useful intervention.

To supplement statistical significance, the reporting of effect sizes is recommended. In fact, in the most recent (fifth) edition of the *APA Publication Manual (2001)*, effect sizes are recommended as "almost always necessary" to include in the results section (pp. 25–26). Effect size indicates the number of standard deviations by which the experimental treatment mean differs from the control treatment mean. Thus an effect size of $+1.00$ indicates a full standard deviation advantage, a large and educationally important effect (Cohen, 1988). Effect sizes of $+0.20$ and $+0.50$ would indicate small and medium effects, respectively. Calculation of effect sizes is relatively straightforward. Helpful guidance and formulas are provided in the recent article by Bruce Thompson (2002), who has served over the past decade as one of the strongest advocates of reporting effect sizes in research papers. Many journals, including *Educational Technology Research and Development (ETR&D)*, presently require effect sizes to be reported.

38.3.2.4 Discussion.
To conclude the report, the *discussion* section explains and interprets the findings relative to the hypotheses or research questions, previous studies, and relevant theory and practice. Where appropriate, weaknesses in procedures that may have impacted results should be identified. Other conventional features of a discussion may include suggestions for further research and conclusions regarding the research hypotheses/questions. For educational technology experiments, drawing implications for practice in the area concerned is highly desirable.

38.3.3 Why Experimental Studies Are Rejected for Publication

After considering the above discussion, readers may question what makes an experimental study "publishable or perishable" in professional research journals. Given that we have not done a formal investigation of this topic, we make only a brief subjective analysis based on our experiences with *ETR&D*. We strongly believe, however, that all of the following factors would apply to every educational technology research journal, although the relative importance they are assigned may vary. Our "top 10" listing is as follows.

Low internal validity of conditions: Treatment and comparison groups are not uniformly implemented. One or more groups have an advantage on a particular condition (time, materials, encouragement) other than the independent (treatment) variable. Example: The treatment group that receives illustrations and text takes 1 hr to study the electricity unit, whereas the text-only group takes only 0.5 hr.

Low internal validity of subject selection/assignment: Groups assigned to treatment and comparison conditions are not comparable (e.g., a more experienced group receives the treatment strategy). Example: In comparing learner control vs. program control, the researcher allows students to select the orientation they want. The higher-aptitude students tend to select program control, which, not surprisingly, yields the better results!

Invalid testing: Outcomes are not measured in a controlled and scientific way (e.g., observations are done by the author without validation of the system or reliability checks of the data). Example: In a qualitative study of teachers' adaptations to technology, only one researcher (the author) observes each of the 10 teachers in a school in which she works part-time as an aide.

Low external validity: Application or importance of topic or findings is weak. Example: The findings show that nonsense syllables take more time to be identified if embedded in a border than if they are isolated. We should note, however, that there are journals that do publish basic research that has a low external validity but a high internal validity.

Poor writing: Writing style is unclear, weak in quality (syntax, construction), and/or does not use appropriate (APA) style. Example: The *method* section contains no subheadings and intermixes descriptions of participants and materials, then discusses the procedures, and ends with introducing the design of the study. Note that the design would be much more useful as an organizer if presented first.

Trivial/inappropriate outcome measures: Outcomes are assessed using irrelevant, trivial, or insubstantial measures. Example: A 10–item multiple-choice test is the only achievement outcome in a study of cooperative learning effects.

Inadequate description of methodology: Instruments, materials, or procedures are not described sufficiently to evaluate the quality of the study. Example: The author describes dependent measures using only the following: "A 10–item posttest was used to assess learning of the unit. It was followed by a 20–item attitude scale regarding reactions to the unit. Other materials used...."

Inappropriate analyses: Quantitative or qualitative analyses needed to address research objectives are not properly used or sufficiently described. Example: In a qualitative study, the author presents the "analysis" of 30 classroom observations exclusively as "holistic impressions," without reference to any application of systematic methods of documenting, transcribing, synthesizing, and verifying what was observed. *Inappropriate discussion of results*: Results are not interpreted accurately or meaningfully to convey appropriate implications of the study. Example: After finding that motivation and performance correlated significantly but very weakly at $r = +.15$, the author discusses for several paragraphs the importance of motivation to learning, "as supported by this study." (Note that although "reliable" in this study, the .15 correlation indicates that motivation accounted for only about 2.25% of the variable in performance: $.15 \times .15 = .0225$).

Insufficient theoretical base or rationale: The basis for the study is conveyed as manipulating some combination of variables essentially "to see what happens." Example: After

reviewing the literature on the use of highlighting text, the author establishes the rationale for his study by stating, "No one, however, has examined these effects using color vs. no-color with males vs. females." The author subsequently fails to provide any theoretical rationales or hypotheses relating to the color or gender variable. A similar fault is providing an adequate literature review, but the hypotheses and/or problem statement are not related or supported by the review.

38.4 THE STATUS OF EXPERIMENTATION IN EDUCATIONAL TECHNOLOGY

38.4.1 Uses and Abuses of Experiments

The behavioral roots of educational technology and its parent disciplines have fostered usage of experimentation as the predominant mode of research. As we show in a later section, experiments comprise the overwhelming proportion of studies published in the research section of *Educational Technology Research and Development (ETR&D)*. The representation of alternative paradigms, however, is gradually increasing.

38.4.1.1 The Historical Predominance of Experimentation. Why is the experimental paradigm so dominant? According to Hannafin (1986), aside from the impetus provided from behavioral psychology, there are three reasons. First, experimentation has been traditionally viewed as the definition of "acceptable" research in the field. Researchers have developed the mentality that a study is of higher quality if it is experimental in design. Positivistic views have reinforced beliefs about the importance of scientific rigor, control, statistical verification, and hypothesis testing as the "correct" approaches to research in the field. Qualitative researchers have challenged this way of thinking, but until recently, acceptance of alternative paradigms has been reluctant and of minimal consequence (Creswell, 2002, pp. 47–48).

Second, Hannafin (1986) proposes that promotion and tenure criteria at colleges and universities have been strongly biased toward experimental studies. If this bias occurs, it is probably attributable mainly to the more respected journals having been more likely to publish experimental designs (see next paragraph). In any case, such practices are perpetuated by creating standards that are naturally favored by faculty and passed down to their graduate students.

Third, the research journals have published proportionately more experimental studies than alternative types. This factor also creates a self-perpetuating situation in which increased exposure to experimental studies increases the likelihood that beginning researchers will also favor the experimental method in their research.

As discussed in later sections, in the 17 years since Hannafin presented these arguments, practices have changed considerably in the direction of greater acceptance of alternative methodologies, such as qualitative methods. The pendulum may have even swung far enough to make the highly controlled experiment with a low external validity less valued than eclectic

orientations that use a variety of strategies to balance internal and external validity (Kozma, 2000).

38.4.1.2 When to Experiment.

The purpose of this chapter is neither to promote nor to criticize the experimental method but, rather, to provide direction for its effective usage in educational technology research. On the one hand, it is fair to say that, probably for the reasons just described, experimentation has been overused by educational technology researchers. The result has frequently been "force-fitting" the experiment in situations where research questions could have been much more meaningfully answered using an alternative design or a combination of several designs.

For example, we recall a study on learner control that was submitted to *ETR&D* for review several years ago. The major research question concerned the benefits of allowing learners to select practice items and review questions as they proceeded through a self-paced lesson. The results showed no effects for the learner-control strategy compared to conventional instruction on either an achievement test or an attitude survey. Despite the study's being well designed, competently conducted, and well described, the decision was *not* to accept the manuscript for publication. In the manner of pure scientists, the authors had carefully measured outcomes but totally omitted any observation or recording of how the subjects used learner-control. Nor did they bother to question the learners on their usage and reactions toward the learner-control options. The experiment thus showed that learner control did not "work" but failed to provide any insights into why.

On the other hand, we disagree with the sentiments expressed by some writers that experimental research conflicts with the goal of improving instruction (Guba, 1969; Heinich, 1984). The fact that carpentry tools, if used improperly, can potentially damage a bookcase does not detract from the value of such tools to skilled carpenters who know how to use them appropriately to build bookcases. Unfortunately, the experimental method has frequently been applied in a very strict, formal way that has blinded the experimenter from looking past the testing of the null hypothesis to inquire why a particular outcome occurs. In this chapter, we take the view that the experiment is simply another valuable way, no more or less sacrosanct than any other, of increasing understanding about methods and applications of educational technology. We also emphasize sensitivity to the much greater concern today than there was 20 or 30 years ago with applying experimental methods to "ecologically valid" (realistic) settings. This orientation implies assigning relatively greater focus on external validity and increased tolerance for minor violations (due to uncontrollable real-world factors) of internal validity. A concomitant need is for contextually sensitive interpretations of findings coupled with replicability studies in similar and diverse contexts.

38.4.1.3 Experiments in Evaluation Research.

In applied instructional design contexts, experiments could potentially offer practitioners much useful information about their products but will typically be impractical to perform. Consider, for example, an instructional designer who develops an innovative way of using an interactive medium to teach principles of chemistry. Systematic evaluation of this instructional method (and of the unit in particular) would comprise an important component of the design process (Dick, Carey, & Carey, 2001; Morrison, Ross, & Kemp, 2001). Of major interest in the evaluation would certainly be how effectively the new method supports instructional objectives compared to conventional teaching procedures. Under normal conditions, it would be difficult logistically to address this question via a true experiment. But if conditions permitted random assignment of students to "treatments?" without compromising the integrity (external validity) of the instruction, a true experimental design would likely provide the most meaningful test. If random assignment were not viable, but two comparable groups of learners were available to experience the instructional alternatives, a quasi-experimental design might well be the next-best choice. The results of either category of experiment would provide useful information for the evaluator, particularly when combined with outcomes from other measures, for either judging the method's effectiveness (summative evaluation) or making recommendations to improve it (formative evaluation). Only a very narrow, shortsighted approach would use the experimental results as isolated evidence for "proving" or "disapproving" program effects.

In the concluding sections of this chapter, we further examine applications and potentialities of "applied research" experiments as sources of information for understanding and improving instruction. First, to provide a better sense of historical practices in the field, we will turn to an analysis of how often and in what ways experiments have been employed in educational technology research.

38.4.2 Experimental Methods in Educational Technology Research

To determine practices and trends in experimental research on educational technology, we decided to examine comprehensively the studies published in a single journal. The journal, *Educational Technology Research and Development* (*ETR&D*), is published quarterly by the Association for Educational Communications and Technology (AECT). *ETR&D* is AECT's only research journal, is distributed internationally, and is generally considered a leading research publication in educational technology. The journal started in 1953 as *AV Communication Review* (*AVCR*) and was renamed *Educational Communication and Technology Journal* (*ECTJ*) in 1978. *ETR&D* was established in 1989 to combine *ECTJ* (AECT's research journal) with the *Journal of Instructional Development* (AECT's design/development journal) by including a research section and a development section. The research section, which is of present interest, solicits manuscripts dealing with "research in educational technology and related topics." Nearly all published articles are blind refereed, with the exception of infrequent solicited manuscripts as part of special issues.

38.4.2.1 Analysis Procedure.

The present analysis began with the Volume I issue of *AVCR* (1953) and ended with Volume

49 (2001) of *ETR&D*. All research studies in these issues were examined and classified in terms of the following categories.

Experimental Studies. This category included (a) true experimental, (b) quasi-experimental, (c) single-subject time series, and (d) repeated-measures time series studies.

Nonexperimental (Descriptive) Studies. The nonexperimental or descriptive studies included correlational, ex post facto, survey, and observational/ethnographic approaches, but these were not used as separate categories—only the experimental studies were classified. A total of 424 articles was classified into one of the four experimental categories or into the overall nonexperimental category. Experimental studies were then classified according to the two additional criteria described below.

Stimulus Materials: *Actual content.* Stimulus materials classified in this category were based on actual content taught in a course from which the subjects were drawn. For example, Tennyson, Welsh, Christensen, and Hajovy (1985) worked with a high-school English teacher to develop stimulus materials that were based on content covered in the English class.

Realistic content. Studies classified in this category used stimulus materials that were factually correct and potentially usable in an actual teaching situation. For example, in examining Taiwanese students' leaning of mathematics, Ku and Sullivan (2000) developed word problems that were taken directly from the fifth-grade textbook used by the students.

Contrived content. This stimulus material category included both nonsense words (Morrison, 1986) and fictional material. For example, Feliciano, Powers, and Kearl (1963) constructed fictitious agricultural data to test different formats for presenting statistical data. Studies in this category generally used stimulus materials with little if any relevance to subjects' knowledge base or interests.

Experimental setting: *Actual setting.* Studies in this category were conducted in either the regular classroom, the computer lab, or an other room used by the subjects for real-life instruction. For example, Nath and Ross (2001) examined student activities in cooperative learning groups on real lessons in their actual classrooms.

Realistic setting. This category consisted of new environments designed to simulate a realistic situation. For example, in the study by Koolstra and Beentjes (1999), elementary students participated in different television-based treatments in vacant school rooms similar to their actual classrooms.

Contrived setting. Studies requiring special equipment or environments were classified in this study. For example, Niekamp's (1981) eye movement study required special equipment that was in-lab designed especially for the data collection.

The final analysis yielded 311 articles classified as experimental (81%) and 71 classified as descriptive (19%). In instances where more than one approach was used, a decision was made by the authors as to which individual approach was predominant. The study was then classified into the latter category. The authors were able to classify all studies into individual design categories. Articles that appeared as literature reviews or studies that clearly lacked the rigor of other articles in the volume were not included in the list of 388 studies. The results of the analysis are described below.

38.4.2.2 Utilization of Varied Experimental Designs. Of the 311 articles classified as experimental, 223 (72%) were classified as true experiments using random assignment of subjects, 77 (25%) of the studies were classified as using quasi-experimental designs, and 11 (35%) were classified as employing time series designs. Thus, following the traditions of the physical sciences and behavioral psychology, use of true-experimental designs has predominated in educational technology research.

An analysis of the publications by decade (e.g., 1953–1962, 1963–1972, 1973–1982) revealed the increased use of true-experimental designs and decreased use of quasi-experimental designs since 1953 (see Fig. 38.1). In the first 10 years of the journal (1953–1962), there was a total of only six experimental studies and three descriptive studies. The experimental studies included two true-experimental and four quasi-experimental designs. During the next 30 years, there was an increase in the number of true-experimental articles. However, in the most recent (abbreviated) decade, from 1993–2001, the percentage of true experiments decreased from the prior decade from 77% to 53% of the total studies, whereas descriptive studies increased from 13% to 45%. This pattern reflects the growing influence of
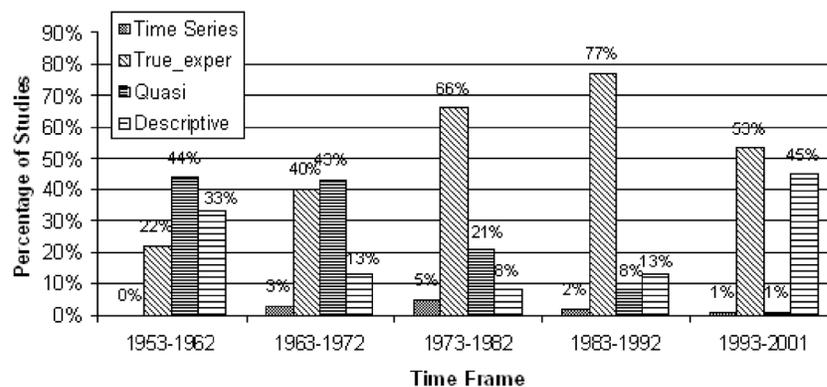

FIGURE 38.1. Experimental design trends.

TABLE 38.2. Designs × Time Frame

Design	1953–1962	1963–1972	1973–1982	1983–1992	1993–2001
Time series	0	3	5	2	1
True experimental	2	40	70	70	41
Quasi-experimental	4	43	22	7	1
Descriptive	3	13	9	12	34

qualitative designs such as case studies. Table 38.2 presents the number of articles published with each design in each of the five decades. It is interesting to note that quasi-experimental designs reached a peak during the 1963–1972 period, with 43 articles, and then decreased to only 1 article in the 1993–2001 time period.

38.4.2.3 Utilization of Stimulus Materials. An additional focus of our analysis was the types of stimulus materials used in the studies. For example, did researchers use actual materials that were either part of the curriculum or derived from the curriculum? Such materials would have a high external validity and provide additional incentive for the subjects to engage in learning process. Figure 38.2 illustrates the three classifications of materials used by the various studies published during the past 40 years. In the period 1963 to 1972, actual materials were clearly used more often than realistic or contrived. Then, starting in 1972, the use of actual materials began a rapid decline, whereas the use of realistic materials tended to increase. There are two possible explanations for this shift from actual to realistic materials. First is the increasing availability of technology and improved media production techniques. During the 1963–1972 time frame, the primary subject of study was film instruction (actual materials). The increased utilization of realistic materials during the 1973–1982 period may have been the result of the availability of other media, increased media production capabilities, and a growing interest in instructional design as opposed to message design. Similarly, in the 1983–1992 time frame, the high utilization of realistic materials may have been due to the increase in experimenter-designed CBI materials using topics appropriate for the subjects but not necessarily based on curriculum objectives. Interestingly, in 1993–2001, relative to the prior decade, the percentage of studies using realistic content almost doubled, from 18% to 31%. This trend seems

attributable to increased interest in external validity in contemporary education technology research.

38.4.2.4 Utilization of Settings. The third question concerns the settings used to conduct the studies. As shown in Fig. 38.3, actual classrooms have remained the most preferred locations for researchers, with a strong resurgence in the past 9 years. Again, it appears that increased concern about the applicability (external validity) of findings has an created impetus for moving from the controlled laboratory setting into real-world contexts, such as classrooms and training centers.

38.4.2.5 Interaction Between Usage Variables. Extending the preceding analyses is the question of which types of stimulus materials are more or less likely to be used in different designs. As shown in Fig. 38.4, realistic materials were more likely to be used in true-experimental designs (48%), whereas actual materials were used most frequently in quasi-experimental designs. Further, as shown in Fig. 38.5, classroom settings were more likely to be chosen for studies using quasi-experimental (82%) than for those using true-experimental (44%) designs. These relationships are predictable, since naturalistic contexts would generally favor quasi-experimental designs over true-experimental designs given the difficulty of making the random assignments needed for the latter. The nature of educational technology research seems to create preferences for realistic as opposed to contrived applications. Yet the trend over time has been to emphasize true-experimental designs and a growing number of classroom applications. Better balances between internal and external validity are therefore being achieved than in the past. Changes in publishing conventions and standards in favor of high experimental control have certainly been influential. Affecting present patterns is the substantive and still growing usage and acceptance of qualitative methods in educational
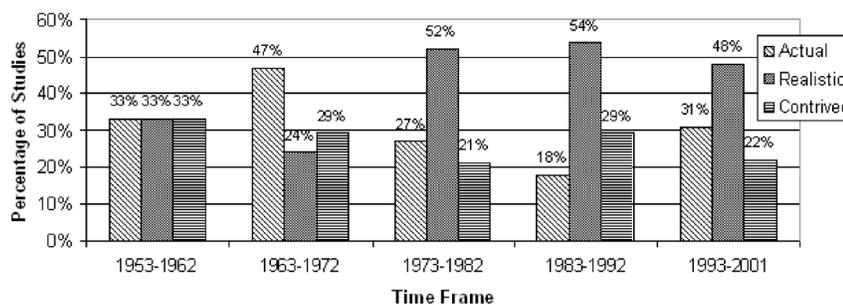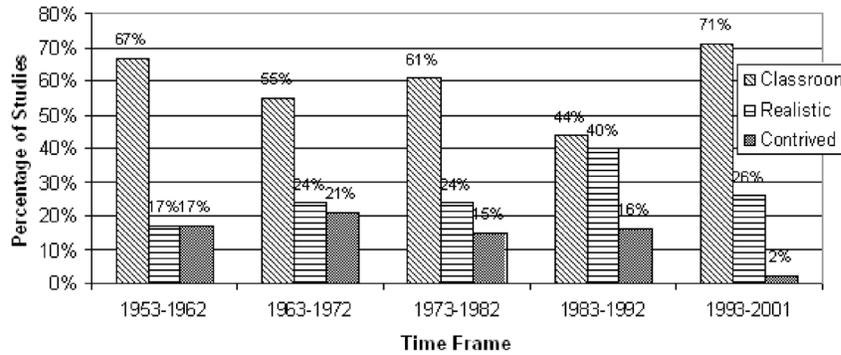

FIGURE 38.2. Trends in stimulus material.

FIGURE 38.3. Trends in settings.

technology research. In our trends analysis, that pattern first became noticeable in the reviewed studies published in 1992 or later.

38.5 CONTEMPORARY ISSUES IN EDUCATIONAL TECHNOLOGY EXPERIMENTATION

38.5.1 Balancing Internal and External Validity

Frequently in this chapter, we have discussed the traditional importance to experimenters of establishing a high internal validity by eliminating sources of extraneous variance in testing treatment effects. Consequently, any differences favoring one treatment over another can be attributed confidently to the intrinsic properties of those treatments rather than to confounding variables, such as one group having a better teacher or more comfortable conditions for learning (see, e.g., reviews by Ross & Morrison, 1989; Slavin, 1993).

The quest for high internal validity orients researchers to design experiments in which treatment manipulations can be tightly controlled. In the process, using naturalistic conditions (e.g., real classrooms) is discouraged, given the many extraneous sources of variance that are likely to operate in those contexts. For example, the extensive research conducted on "verbal

learning" in the 1960s and 1970s largely involved associative learning tasks using simple words and nonsense syllables (e.g., see Underwood, 1966). With simplicity and artificiality comes greater opportunity for control.

This orientation directly supports the objectives of the basic learning or educational psychology researcher whose interests lie in testing the generalized theory associated with treatment strategies, independent of the specific methods used in their administration. Educational technology researchers, however, are directly interested in the interaction of medium and method (Kozma, 1991, 1994; Ullmer, 1994). To learn about this interaction, realistic media applications rather than artificial ones need to be established. In other words, external validity becomes as important a concern as internal validity.

Discussing these issues brings to mind a manuscript that one of us was asked to review a number of years ago for publication in an educational research journal. The author's intent was to compare, using an experimental design, the effects on learning of programmed instruction and CBI. To avoid Clark's (1983) criticism of performing a media comparison, i.e., confounding media with instructional strategies, the author decided to make the two "treatments" as similar as possible in all characteristics except delivery mode. This essentially involved replicating the exact programmed instruction design in the CBI condition. Not surprisingly, the findings showed no difference between treatments, a direct justification of Clark's (1983) position. But, unfortunately, this result (or one showing an actual treatment
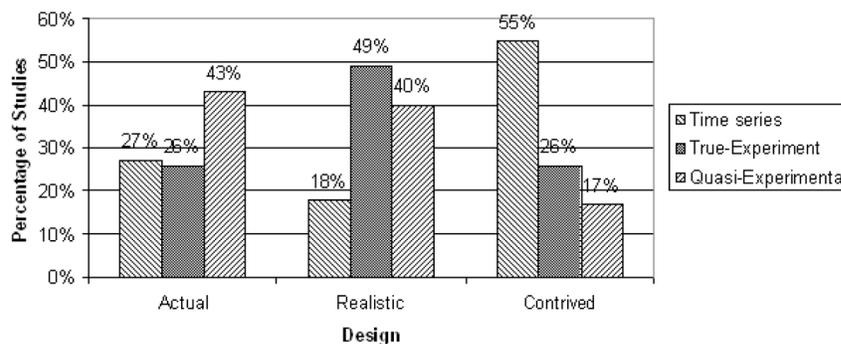

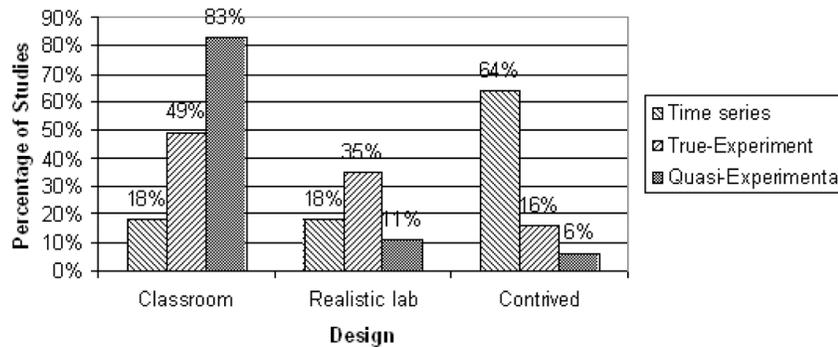FIGURE 38.4. Experimental designs × materials.

FIGURE 38.5. Experimental design × setting.

effect as well) would be meaningless for advancing theory or practice in educational technology. By stripping away the special attributes of a normal CBI lesson (e.g., interaction, sound, adaptive feedback, animation), all that remained were alternative forms of programmed instruction and the unexciting finding, to use Clark's (1983) metaphor, that groceries delivered in different, but fundamentally similar, ways still have the same nutritional value. Needless to say, this study, with its high internal validity but very low external validity, was evaluated as unsuitable for publication. Two more appropriate orientations for educational technology experiments are proposed in the following sections.

38.5.1.1 Randomized Field Experiments.
Given the importance of balancing external validity (application) and internal validity (control) in educational technology research, an especially appropriate design is the randomized field experiment (Slavin, 1997), in which instructional programs are evaluated over relatively long periods of time under realistic conditions. In contrast to descriptive or quasi-experimental designs, the randomized field experiment requires random assignment of subjects to treatment groups, thus eliminating differential selection as a validity threat.

For example, Nath and Ross (2001) randomly assigned elementary students working in cooperative learning dyads to two groups. The treatment group received training in cooperative learning over seven sessions during the school year, while the control group participated in unrelated ("placebo") group activities. At eight different times, the cooperative dyads were observed using a standardized instrument to determine the level and types of cooperative activities demonstrated. Results indicated that in general the treatment group surpassed the control group in both communication and cooperative skills. Students in grades 2–3 showed substantially more improvement than students in grades 4–6. The obvious advantage of the randomized field experiment is the high external validity. Had Nath and Ross (2001) tried to establish cooperative groupings outside the regular classroom, using volunteer students, the actual conditions of peer interactions would have been substantially altered and likely to have yielded different results. On the other hand, the randomized field experiment concomitantly sacrifices internal validity, because its length

and complexity permit interactions to occur with confounding variables. Nath and Ross' (2001) results, for example, might have been influenced by students' discussing the study and its different conditions with one another after class (e.g., diffusion of treatments). It was definitely influenced, as the authors describe in detail, by the teachers' level of expertise in cooperative learning pedagogy, the cooperative learning tasks assigned, and the ways in which learning conditions were established in the particular class. The experimental results from such studies, therefore, reflect "what really happens" from combined effects of treatment and environmental variables rather than the pure effects of an isolated instructional strategy.

38.5.1.2 Basic–Applied Design Replications.
Basic research designs demand a high degree of control to provide valid tests of principles of instruction and learning. Once a principle has been thoroughly tested with consistent results, the natural progression is to evaluate its use in a real-world application. For educational technologists interested in how learners are affected by new technologies, the question of which route to take, basic vs. applied, may pose a real dilemma. Typically, existing theory and prior research on related interventions will be sufficient to raise the possibility that further basic research may not be necessary. Making the leap to a real-life application, however, runs the risk of clouding the underlying causes of obtained treatment effects due to their confounding with extraneous variables.

To avoid the limitations of addressing one perspective only, a potentially advantageous approach is to look at both using a replication design. "Experiment 1," the basic research part, would examine the variables of interest by establishing a relatively high degree of control and high internal validity. "Experiment 2," the applied component, would then reexamine the same learning variables by establishing more realistic conditions and a high external validity. Consistency of findings across experiments would provide strong convergent evidence supporting the obtained effects and underlying theoretical principles. Inconsistency of findings, however, would suggest influences of intervening variables that alter the effects of the variables of interest when converted from their "pure" form to realistic applications. Such contamination may often represent "media effects," as might occur, for example, when feedback strategies

used with print material are naturally made more adaptive (i.e., powerful and effectual) via interactive CBI (see Kozma, 1991). (For example, a learner who confuses discovery learning with inquiry learning in response to an inserted lesson question may be branched immediately to a remedial CBI frame that differentiates between the two approaches, whereas his or her counterpart in a parallel print lesson might experience the same type of feedback by having to reference the response selected on an answer page and manually locate the appropriate response-sensitive feedback in another section of the lesson.) The next implied step of a replication design would be further experimentation on the nature and locus of the altered effects in the applied situation. Several examples from the literature of the basic–applied replication orientation follow.

Example 1. In a repeated-measures experiment that we conducted several years ago, we asked adult subjects to indicate their preferences for screen designs representing differing degrees of text density (Morrison, Ross, Schultz, & O'Dell, 1989). In one experiment, high internal validity was established by having learners judge only the initial screen of a given text presentation, thus keeping the number of displays across higher- and lower-density variations constant. In realistic lessons, however, using lower-density displays requires the use of additional screens (or more scrolling) to view the content fully. Accordingly, a parallel experiment, having a higher external validity but a lower internal validity, was conducted in which the number of screens was allowed to vary naturally in accord with the selected density level.

Both experiments produced similar results, supporting higher- over lower-density displays, regardless of the quantity of screens that conveyed a particular density condition. Consequently, we were able to make a stronger case both for the theoretical assumption that higher density would provide greater contextual support for comprehending expository text and for the practical recommendation that such density levels be considered for the design of actual CBI lessons.

Example 2. In a design used by Winn and Solomon (1993), nonsense syllables served as verbal stimuli in experiment 1. Findings indicated that the interpretation of diagrams containing verbal labels (e.g., "Yutcur" in box A and "Nipden" in box B) was determined mainly by syntactic rules of English. For example, if box B were embedded in box A, subjects were more likely to select, as an interpretation, "Yutcur are Nipden" than the converse description. However, when English words were substituted for the nonsense syllables (e.g., "sugar" in box A and "spice" in box B) in experiment 2, this effect was overridden by common semantic meanings. For example, "Sugar is spice" would be a more probable response than the converse, regardless of the diagram arrangement. Taken together, the two experiments supported theoretical assumptions about the influences of diagram arrangement on the interpreted meaning of concepts, while suggesting for designers that appropriate diagram arrangements become increasing critical as the meaningfulness of the material decreases.

Example 3. Although using a descriptive rather than experimental design, Grabinger (1993) asked subjects to judge the readability of "model" screens that presented symbolic notation as opposed to real content in different formats (e.g., using

or not using illustrations, status bars, headings). Using multidimensional scaling analysis, he found that evaluations were made along two dimensions: organization and structure. In a second study, he replicated the procedure using real content screens. Results yielded only one evaluative dimension that emphasized organization and visual interest. In this case, somewhat conflicting results from the basic and applied designs required the researcher to evaluate the implications of each relative to the research objectives. The basic conclusion reached was that although the results of study I were free from content bias, the results of study 2 more meaningfully reflected the types of decisions that learners make in viewing CBI information screens.

Example 4. Morrison et al. (1995) examined uses of different feedback strategies in learning from CBI. Built into the experimental design was a factor representing the conditions under which college student subjects participated in the experiment: simulated or realistic. Specifically, in the simulated condition, the students from selected education courses completed the CBI lesson to earn extra credit toward their course grade. The advantage of using this sample was increased internal validity, given that students were not expected to be familiar with the lesson content (writing instructional objectives) or to be studying it during the period of their participation. In the realistic condition, subjects were students in an instructional technology course for which performance on the CBI unit (posttest score) would be computed in their final average.

Interestingly, the results showed similar relative effects of the different feedback conditions; for example, knowledge of correct response (KCR) and delayed feedback tended to surpass no-feedback and answer-until-correct (AUC) feedback. Examination of learning process variables, however, further revealed that students in the realistic conditions performed better, while making greater and more appropriate use of instructional support options provided in association with the feedback. Whereas the simulated condition was valuable as a more basic and purer test of theoretical assumptions, the realistic condition provided more valid insights into how the different forms of feedback would likely be used in combination with other learning resources on an actual learning task.

38.5.2 Assessing Multiple Outcomes in Educational Technology Experiments

The classic conception of an experiment might be to imagine two groups of white rats, one trained in a Skinner Box under a continuous schedule of reinforcement and the other under an intermittent schedule. After a designated period of training, reinforcement (food) is discontinued, and the two groups of rats are compared on the number of trials to extinction. That is, how long will they continue to press the bar even though food is withheld?

In this type of experiment, it is probable that the single dependent measure of "trials" would be sufficient to answer the research question of interest. In educational technology research, however, research questions are not likely to be resolved in so straightforward a manner. Merely knowing that

one instructional strategy produced better achievement than another provides little insight into how those effects occurred or about other possible effects of the strategies. Earlier educational technology experiments, influenced by behavioristic approaches to learning, were often subject to this limitation.

For example, Shettel, Faison, Roshal, and Lumsdaine (1956) compared live lectures and identical film lectures on subjects (Air Force technicians) learning fuel and rudder systems. The dependent measure was immediate and delayed multiple-choice tests on three content areas. Two outcomes were significant, both favoring the live-lecture condition on the immediate test. Although the authors concluded that the films taught the material less well than the "live" lectures, they were unable to provide any interpretation as to why. Observation of students might have revealed greater attentiveness to the live lecture, student interviews might have indicated that the film audio was hard to hear, or a problem-solving test might have shown that application skills were low (or high) under both presentations.

Released from the rigidity of behavioristic approaches, contemporary educational technology experimenters are likely to employ more and richer outcome measures than did their predecessors. Two factors have been influential in promoting this development. One is the predominance of cognitive learning perspectives in the past two decades (Bransford, Brown, & Cocking, 1999; Snow & Lohman, 1989; Tennyson, 1992); the other has been the growing influence of qualitative research methods.

38.5.2.1 Cognitive Applications.
In their comprehensive review paper, Snow and Lohman (1989) discuss influences of cognitive theory on contemporary educational measurement practices. One key contribution has been the expansion of conventional assessment instruments so as to describe more fully the "cognitive character" of the target. Among the newer, cognitively derived measurement applications that are receiving greater usage in research are tests of declarative and procedural knowledge, componential analysis, computer simulations, faceted tests, and coaching methods, to name only a few.

Whereas behavioral theory stressed learning products, such as accuracy and rate, cognitive approaches also emphasize learning processes (Brownell, 1992). The underlying assumption is that learners may appear to reach similar destinations in terms of observable outcomes but take qualitatively different routes to arrive at those points. Importantly, the routes or "processes" used determine the durability and transferability of what is learned (Mayer, 1989). Process measures may include such variables as the problem-solving approach employed, level of task interest, resources selected, learning strategies used, and responses made on the task. At the same time, the cognitive approach expands the measurement of products to include varied, multiple learning outcomes such as declarative knowledge, procedural knowledge, long-term retention, and transfer (Tennyson & Rasch, 1988).

This expanded approach to assessment is exemplified in a recent experiment by Cavalier and Klein (1998). The focus of the study was comparing the effects of implementing cooperative versus individual learning and orienting activities during CBI. Students working in cooperative dyads or individually completed a CBI earth science program that contained advance organizers, instructional objectives, or no orienting activities. Results indicated that students who received the instructional objectives performed highest on the posttest. This information alone, however, would have provided little insight into how learning objectives might be used by students and, in the case of dyads, how they might influence the dynamics of learner interactions. Accordingly, Cavalier and Klein also examined interaction behaviors while students were learning under the different orienting activities. Findings revealed, for example, that cooperative dyads receiving objectives exhibited more helping behaviors and on-task behaviors than those not receiving orienting activities. Qualitative data from attitude surveys provided further insight into how students approached the instructional task and learning structure. Using these multiple outcome measures, the researchers acquired a clearer perspective on how processes induced by the different strategies culminated in the learning products obtained.

Use of special assessments that directly relate to the treatment is illustrated in a study by Shin, Schallert, and Savenye (1994). Both quantitative and qualitative data were collected to determine the effectiveness of leaner control with elementary students who varied in prior knowledge. An advisement condition that provided the subject with specific directions as to what action to take next was also employed. Quantitative data collected consisted of both immediate and delayed posttest scores, preferences for the method, self-ratings of difficulty, and lesson completion time. The qualitative data included an analysis of the path each learner took through the materials. This analysis revealed that nonadvisement students became lost in the hypertext "maze" and often went back and forth between two sections of the lessons as though searching for a way to complete the lesson. In contrast, students who received advisement used the information to make the proper decisions regarding navigation more than 70% of the time. Based on the qualitative analysis, they concluded that advisement (e.g., orientation information, what to do next) was necessary when learners could freely access (e.g., learner control) different parts of the instruction at will. They also concluded that advisement was not necessary when the program controlled access to the instruction.

Another example of multiple and treatment-oriented assessments is found in Neuman's (1994) study on the applicability of databases for instruction. Neuman used observations of the students using the database, informal interviews, and document analysis (e.g., review of assignment, search plans, and search results). This triangulation of data provided information on the design and interface of the database. If the data collection were limited to the number of citations found or used in the students' assignment, the results might have shown that the database was quite effective. Using a variety of sources allowed the researcher to make specific recommendations for improving the database rather than simply concluding that it was beneficial or was not.

38.5.2.2 Qualitative Research.
In recent years, educational researchers have shown increasing interest in qualitative research approaches. Such research involves naturalistic inquiries using techniques such as in-depth interviews, direct

observation, and document analysis (Patton, 1990). Our position, in congruence with the philosophy expressed throughout this chapter is that quantitative and qualitative research are more useful when used together than when either is used alone (see, e.g., Gliner & Morgan, 2000, pp. 16–28). Both provide unique perspectives, which, when combined, are likely to yield a richer and more valid understanding.

Presently, in educational technology research, experimentalists have been slow to incorporate qualitative measures as part of their overall research methodology. To illustrate how such an integration could be useful, we recall conducting an editorial review of a manuscript submitted by Klein and Pridemore (1992) for publication in *ETR&D*. The focus of their study was the effects of cooperative learning and need for affiliation on performance and satisfaction in learning from instructional television. Findings showed benefits for cooperative learning over individual learning, particularly when students were high in affiliation needs. Although we and the reviewers evaluated the manuscript positively, a shared criticism was the lack of data reflecting the nature of the cooperative interactions. It was felt that such qualitative information would have increased understanding of why the treatment effects obtained occurred. Seemingly, the same recommendation could be made for nearly any applied experiment on educational technology uses. The following excerpt from the *published* version of the Klien and Pridemore paper illustrates the potential value of this approach:

. . . Observations of subjects who worked cooperatively suggested that they did, in fact, implement these directions [to work together, discuss feedback, etc.]. After each segment of the tape was stopped, one member of the dyad usually read the practice question aloud. If the question was unclear to either member, the other would spend time explaining it . . . [in contrast to individuals who worked alone] read each question quietly and would either immediately write their answer in the workbook or would check the feedback for the correct answer. These informal observations tend to suggest that subjects who worked cooperatively were more engaged than those who worked alone. (p. 45)

Qualitative and quantitative measures can thus be used collectively in experiments to provide complementary perspectives on research outcomes.

38.5.3 Item Responses vs. Aggregate Scores as Dependent Variables

Consistent with the "expanded assessment" trend, educational technology experiments are likely to include dependent variables consisting of one or more achievement (learning) measures, attitude measures, or a combination of both types. In the typical case, the achievement or attitude measure will be a test comprised of multiple items. By summing item scores across items, a total or "aggregate" score is derived. To support the validity of this score, the experimenter may report the test's internal-consistency reliability (computed using Cronbach's alpha or the KR-20 formula) or some other reliability index. Internal consistency represents "equivalence reliability"—the extent to which parts of a test are equivalent (Wiersma & Jurs, 1985). Depending on the situation, these procedures could

prove limiting or even misleading with regard to answering the experimental research questions.

A fundamental question to consider is whether the test is designed to measure a unitary construct (e.g., ability to reduce fractions or level of test anxiety) or multiple constructs (e.g., how much students liked the lesson and how much they liked using a computer). In the latter cases, internal consistency reliability might well be low, because students vary in how they perform or how they feel across the separate measures. Specifically, there may be no logical reason why good performances on, say, the "math facts" portion of the test should be highly correlated with those on the problem-solving portion (or why reactions to the lesson should strongly correlate with reactions to the computer). It may even be the case that the treatments being investigated are geared to affect one type of performance or attitude more than another. Accordingly, one caution is that, where multiple constructs are being assessed by *design,* internal-consistency reliability may be a poor indicator of construct validity. More appropriate indexes would assess the degree to which (a) items within the separate subscales intercorrelate (subscale internal consistency), (b) the makeup of the instruments conforms with measurement objectives (content validity), (c) students answer particular questions in the same way on repeated administrations (test–retest reliability), and (d) subscale scores correlate with measures of similar constructs or identified criteria (construct or predictive validity).

Separate from the test validation issue is the concern that aggregate scores may mask revealing patterns that occur across different subscales and items. We explore this issue further by examining some negative and positive examples from actual studies.

38.5.3.1 Aggregating Achievement Results. We recall evaluating a manuscript for publication that described an experimental study on graphic aids. The main hypothesis was that such aids would primarily promote better understanding of the science concepts being taught. The dependent measure was an achievement test consisting of factual (fill-in-the-blank), application (multiple-choice and short answer), and problem-solving questions. The analysis, however, examined total score only in comparing treatments. Because the authors had not recorded subtest scores and were unable to rerun the analysis to provide such breakdowns (and, thereby, directly address the main research question), the manuscript was rejected.

38.5.3.2 Aggregating Attitude Results. More commonly, educational technology experimenters commit comparable oversights in analyzing attitude data. When attitude questions concern different properties of the learning experience or instructional context, it may make little sense to compute a total score, unless there is an interest in an overall attitude score. For example, in a study using elaborative feedback as a treatment strategy, students may respond that they liked the learning material but did not use the feedback. The overall attitude score would mask the latter, important finding.

For a brief illustration, we recall a manuscript submitted to *ETR&D* in which the author reported only aggregate results on a postlesson attitude survey. When the need for individual item

information was requested, the author replied, "The KR-20 reliability of the scale was .84; therefore, all items are measuring the same thing." Although a high internal consistency reliability implies that the items are "pulling in the same direction," it does not also mean necessarily that all yielded equally positive responses. For example, as a group, learners might have rated the lesson material very high, but the instructional delivery very low. Such specific information might have been useful in furthering understanding of why certain achievement results occurred.

Effective reporting of item results was done by Ku and Sullivan (2000) in a study assessing the effects of personalizing mathematics word problems on Taiwanese students' learning. One of the dependent measures was a six-item attitude measure used to determine student reactions to different aspects of the learning experience. Rather than combining the items to form a global attitude measure, the authors performed a MANOVA comparing the personalized and control treatments on the various items. The MANOVA was significant, thereby justifying follow-up univariate treatment comparisons on each item. Findings revealed that although the personalized group tended to have more favorable reactions toward the lesson, the differences were concentrated (and statistically significant) on only three of the items—ones concerning the students' interest, their familiarity with the referents (people and events) in the problems, and their motivation to do more of that type of math problem. More insight into learner experiences was thus obtained relative to examining the aggregate score only. It is important to keep in mind, however, that the multiple statistical tests resulting from individual item analyses can drastically inflate the chances of making a Type I error (falsely concluding that treatment effects exists). As exemplified in the Ku and Sullivan (2000) study, use of appropriate statistical controls, such as MANOVA (see Table 38.1) or a reduced alpha (significance) level, is required.

38.5.4 Media Studies vs. Media Comparisons

As confirmed by our analysis of trends in educational technology experimentation, a popular focus of the past was comparing different types of media-based instruction to one another or to teacher-based instruction to determine which approach was "best." The fallacy or, at least, unreasonableness of this orientation, now known as "media comparison studies," was forcibly explicated by Clark (1983) in his now classic article (see also Hagler & Knowlton, 1987; Petkovich & Tennyson, 1984; Ross & Morrison, 1989; Salomon & Clark, 1977). As previously discussed, in that paper, Clark argued that media were analogous to grocery trucks that carry food but do not in themselves provide nourishment (i.e., instruction). It, therefore, makes little sense to compare delivery methods when instructional strategies are the variables that impact learning.

For present purposes, these considerations present a strong case against experimentation that simply compares media. Specifically, two types of experimental designs seem particularly unproductive in this regard. One of these represents treatments as amorphous or "generic" media applications, such as

CBI, interactive video, and Web-based instruction. The focus of the experiment then becomes which medium "produces" the highest achievement. The obvious problem with such research is the confounding of results with numerous media attributes. For example, because CBI may offer immediate feedback, animation, and sound, whereas a print lesson does not, differences in outcomes from the two types of presentations would be expected to the extent that differentiating attributes impact criterion performance. More recently, this type of study has been used to "prove" the effectiveness of distance education courses. A better approach is an evaluation study that determines if the students were able to achieve the objectives for the course (Morrison, 2001). Little can be gained by comparing two delivery systems in comparison to determining if a course and the strategies are effective in helping the students achieve the stated objectives. A second type of inappropriate media comparison experiment is to create artificially comparable alternative media presentations, such that both variations contain identical attributes but use different modes of delivery. In an earlier section, we described a study in which CBI and a print manual were used to deliver the identical programmed instruction lesson. The results, which predictably showed no treatment differences, revealed little about CBI's capabilities as a medium compared to those of print lessons. Similarly, to learn about television's "effects" as a medium, it seems to make more sense to use an actual television program, as in Koolstra and Beentjes' (1999) study of subtitle effects, than a simulation done with a home videocamera. So where does this leave us with regard to experimentation on media differences? We propose that researchers consider two related orientations for "media studies." Both orientations involve conveying media applications realistically, whether "conventional" or "ideal" (cutting edge) in form. Both also directly compare educational outcomes from the alternative media presentations. However, as explained below, one orientation is deductive in nature and the other is inductive.

38.5.4.1 Deductive Approach: Testing Hypotheses About Media Differences. In this first approach, the purpose of the experiment is to test a priori hypotheses of differences between the two media presentations based directly on analyses of their different attributes (see Kozma, 1991, 1994). For example, it might be hypothesized that for teaching an instructional unit on a cardiac surgery procedure, a conventional lecture presentation would be superior to an interactive video presentation for facilitating retention of factual information, whereas the converse would be true for facilitating meaningful understanding of the procedure. The rationale for these hypotheses would be based directly on analyses of the special capabilities (embedded attributes or instructional strategies) of each medium in relation to the type of material taught. Findings would be used to support or refute these assumptions.

An example of this a priori search for media differences is the study by Aust, Kelley, and Roby (1993) on "hypereference" (on line) and conventional paper dictionary use in foreign-language learning. Because hypereferences offer immediate access to supportive information, it was hypothesized and

confirmed that learners would consult such dictionaries more frequently and with greater efficiency than they would conventional dictionaries.

38.5.4.2 Inductive Approach: Replicating Findings Across Media.

The second type of study, which we have called *media replications* (Ross & Morrison, 1989), examines the consistency of effects of given instructional strategies delivered by alternative media. Consistent findings, if obtained, are treated as corroborative evidence to strengthen the theoretical understanding of the instructional variables in question as well as claims concerning the associated strategy's effectiveness for learning. If inconsistent outcomes are obtained, methods and theoretical assumptions are reexamined and the target strategy subjected to further empirical tests using diverse learners and conditions. Key interests are why results were better or worse with a particular medium and how the strategy might be more powerfully represented by the alternative media. Subsequent developmental research might then explore ways of incorporating the suggested refinements in actual systems and evaluating those applications. In this manner, media replication experiments use an inductive, post hoc procedure to identify media attributes that differentially impact learning. At the same time, they provide valuable generalizability tests of the effects of particular instructional strategies.

The classic debate on media effects (Clark, 1983, 1994, 2001; Kozma, 1994) is important for sharpening conceptualization of the role of media in enhancing instruction. However, Clark's focal argument that media do not affect learning should not be used as a basis for discouraging experimentation that compares educational outcomes using different media. In the first orientation reviewed above, the focus of the experiment is hypothesized effects on learning of instructional strategies embedded in media. In the second orientation, the focus is the identified effects of media in altering how those strategies are conveyed. In neither case is the medium itself conceptualized as the direct cause of learning. In both cases, the common goal is increasing theoretical and practical understanding of how to use media more effectively to deliver instruction.

38.6 SUMMARY

In this chapter, we have examined the historical roots and current practices of experimentation in educational technology. Initial usage of experimental methods received impetus from behavioral psychology and the physical sciences. The basic interest was to employ standardized procedures to investigate the effects of treatments. Such standardization ensured a high internal validity or the ability to attribute findings to treatment variations as opposed to extraneous factors.

Common forms of experimentation consist of true experiments, repeated-measures designs, quasi-experiments, and time series designs. Internal validity is generally highest with true experiments due to the random assignment of subjects to different treatments. Typical threats to internal validity consist of history, maturation, testing, instrumentation, statistical regression, selection, experimental mortality, and diffusion of treatments.

Conducting experiments is facilitated by following a systematic planning and application process. A seven-step model suggested consists of (1) selecting a topic, (2) identifying the research problem, (3) conducting a literature search, (4) stating research questions or hypotheses, (5) identifying the research design, (6) determining methods, and (7) identifying data analysis approaches.

For experimental studies to have an impact on theory and practice in educational technology, their findings need to be disseminated to other researchers and practitioners. Getting a research article published in a good journal requires careful attention to writing quality and style conventions. Typical write-ups of experiments include as major sections an introduction (problem area, literature review, rationale, and hypotheses), method (subjects, design, materials, instruments, and procedure), results (analyses and findings), and discussion. Today, there is increasing emphasis by the research community and professional journals on reporting effects sizes (showing the magnitude or "importance" of experimental effects) in addition to statistical significance.

Given their long tradition and prevalence in educational research, experiments are sometimes criticized as being overemphasized and conflicting with the improvement of instruction. However, experiments are not intrinsically problematic as a research approach but have sometimes been used in very strict, formal ways that have blinded educational researchers from looking past results to gain understanding about learning processes. To increase their utility to the field, experiments should be used in conjunction with other research approaches and with nontraditional, supplementary ways of collecting and analyzing results.

Analysis of trends in using experiments in educational technology, as reflected by publications in *ETR&D* (and its predecessors) over the last five decades, show consistent trends as well as some changing ones. True experiments have been much more frequently conducted over the years relative to quasi-experiments, time series designs, and descriptive studies. However, greater balancing of internal and external validity has been evidenced over time by increasing usage in experiments of realistic but simulated materials and contexts as opposed to either contrived or completely naturalistic materials and contexts.

Several issues seem important to current uses of experimentation as a research methodology in educational technology. One is balancing internal validity and external validity, so that experiments are adequately controlled while yielding meaningful and applicable findings. Two orientations suggested for achieving such balance are the randomized field experiment and the "basic–applied" design replication. Influenced and aided by advancements in cognitive learning approaches and qualitative research methodologies, today's experimenters are also more likely than their predecessors to use multiple data sources to obtain corroborative and supplementary evidence regarding the learning processes and products associated with the strategies evaluated. Looking at individual item results as opposed to only aggregate scores from cognitive and attitude measures is consistent with the orientation.

Finally, the continuing debate regarding "media effects" notwithstanding, media comparison experiments remain

interesting and viable in our field. The goal is not to compare media generically to determine which are "best" but, rather, to further understanding of (a) how media differ in their capabilities for conveying instructional strategies and (b) how the influences of instructional strategies are maintained or altered via different media presentations.

References

Alper, T., Thoresen, C. E., & Wright, J. (1972). *The use of film mediated modeling and feedback to change a classroom teacher's classroom responses.* Palo Alto, CA: Stanford University, School of Education, R&D Memorandum 91.

Aust, R., Kelley, M. J. & Roby, W. (1993). The use of hypereference and conventional dictionaries. *Educational Technology Research & Development, 41*(4), 63-71.

Borg, W. R., Gall, J. P., & Gall, M. D. (1993). *Applying educational research* (3rd ed.). New York: Longman.

Bransford, J. D., Brown, A. L., & Cocking, R. R. (1999). *How people learn: Brain, mind, experience, and school*. Washington, DC: National Academy Press.

Brownell, W. A. (1992). Reprint of criteria of learning in educational research. *Journal of Educational Psychology, 84,* 400-404.

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.

Cavalier, J. C., & Klein, J. D. (1998). Effects of cooperative versus individual learning and orienting activities during computer-based instruction. *Educational Technology Research and Development, 46*(1), 5-18.

Clariana. R. B., & Lee, D. (2001). The effects of recognition and recall study tasks with feedback in a computer-based vocabulary lesson. *Educational Technology Research and Development, 49*(3), 23-36.

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

Clark, R. E. (1994). Media will never influence learning. *Educational Technology Research and Development, 42*(2), 21-29.

Clark, R. E. (Ed.). (2001). *Learning from media: Arguments, analysis, and evidence.* Greenwich, CT: Information Age.

Clark, R. E., & Snow, R. E. (1975). Alternative designs for instructional technology research. *AV Communication Review, 23,* 373-394.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Creswell, J. W. (2002). *Educational research.* Upper Saddle River, NJ: Pearson Education.

Cronbach, L. J. (1957). The two disciplines of scientific psychology. *American Psychologist, 12,* 671-684.

Dick, W., Carey, L., & Carey, J. (2001). *The systematic design of instruction* 4th ed.). New York: HarperCollins College.

Feliciano, G. D., Powers, R. D., & Kearl, B. E. (1963). The presentation of statistical information. *AV Communication Review, 11,* 32-39.

Gerlic, I., & Jausovec, N. (1999). Multimedia differences in cognitive processes observed with EEG. *Educational Technology Research and Development, 47*(3), 5-14.

Glass, G. V., & Hopkins, K. D. (1984). *Statistical methods in education and psychology.* Englewood Cliffs, NJ: Prentice Hall.

Gliner, J. A., & Morgan, G. A. (2000). *Research methods in applied settings: An integrated approach to design and analysis.* Mahwah, NJ: Lawrence Erlbaum Associates.

Grabinger, R. S. (1993). Computer screen designs: Viewer judgments. *Educational Technology Research & Development, 41*(2), 35-73.

Gronlund, N. E., & Linn, R. L. (1990). *Measurement and evaluation in teaching* (6th ed.). New York: Macmillan.

Guba, E. G. (1969). The failure of educational evaluation. *Educational Technology, 9*(5), 29-38.

Hagler, P., & Knowlton, J. (1987). Invalid implicit assumption in CBI comparison research. *Journal of Computer-Based Instruction, 14,* 84-88.

Hannafin, M. J. (1986). Ile status and future of research in instructional design and technology. *Journal of Instructional Development, 8,* 24-30.

Heinich, R. (1984). The proper study of educational technology. *Educational Communication and Technology Journal, 32,* 67-87.

Jonassen, D. H. (1991). Chaos in instructional design. *Educational Technology, 30,* 32-34.

Jonassen, D. H., Campbell, J. P., & Davidson, M. E. (1994). Learning with media: Restructuring the debate. *Educational Technology Research & Development, 42*(2), 31-39.

Klein, J. D., & Pridemore, D. R. (1992). Effects of cooperative learning and the need for affiliation on performance, time on task, and satisfaction. *Educational Technology Research & Development, 40*(4), 39-48.

Knowlton, J. Q. (1964). A conceptual scheme for the audiovisual field. *Bulletin of the School of Education: Indiana University, 40*(3), 1-44.

Koolstra, C. M., & Beentjes, J. W. J. (1999). Children's vocabulary acquisition in a foreign language through watching subtitled television programs at home. *Educational Technology Research & Development, 47*(1), 51-50.

Kozma, R. B. (1991). Learning with media. *Review of Educational Research, 61,* 179-212.

Kozma, R. B. (1994). Will media influence learning? Refraining the debate. *Educational Technology Research and Development, 42*(2) 7-19.

Kozma, R. B. (2000). Reflections on the state of educational technology research and development: A reply to Richey. *Educational Technology Research and Development, 48*(1), 7-19.

Ku, H-Y., & Sullivan, H. (2000). Learner control over full and lean computer-based instruction under personalization of mathematics word problems in Taiwan. *Educational Technology Research and Development, 48*(3), 49-60.

Mayer, R. E. (1989). Models for understanding. *Review of Educational Research, 59,* 43-64.

Morrison, G. R. (1986). Communicability of the emotional connotation of type. *Educational Communication and Technology Journal, 43*(1), 235-244.

Morrison, G. R. (2001). New directions: Equivalent evaluation of instructional media: The next round of media comparison studies. In R. E. Clark (Ed.), *Learning from media: Arguments, analysis, and evidence* (pp. 319-326). Greenwich, CT: Information Age.

Morrison, G. R., Ross, S. M., Schultz, C. X., & O'Dell, J. K. (1989). Learner preferences for varying screen densities using realistic stimulus materials with single and multiple designs. *Educational Technology Research & Development, 37*(3), 53-62.

Morrison, G. R., Ross, S. M., Gopalakrishnan, M., & Casey, J. (1995). The effects of incentives and feedback on achievement in computer-based instruction. *Contemporary Educational Psychology, 20,* 32–50.

Morrison, G. R., Ross, S. M., & Kemp, J. E. (2001). *Designing effective instruction: Applications of instructional design.* New York: John Wiley & Sons.

Nath, L. R., & Ross, S. M. (2001). The influences of a peer-tutoring training model for implementing cooperative groupings with elementary students. *Educational Technology Research & Development, 49*(2), 41–56.

Neuman, D. (1994). Designing databases as tools for higherlevel learning: Insights from instructional systems design. *Educational Technology Research & Development, 41*(4), 25–46.

Niekamp, W. (1981). An exploratory investigation into factors affecting visual balance. *Educational Communication and Technology Journal, 29,* 37–48.

Patton, M. G. (1990). *Qualitative evaluation and research methods,* (2nd ed.). Newbury Park, CA: Sage.

Petkovich, M. D., & Tennyson, R. D. (1984). Clark's "Learning from media": A critique. *Educational Communication and Technology Journal, 32*(4), 233–241.

Popham, J. X. (1990). *Modern educational measurement: A practitioner's perspective* (2nd ed.). Englewood Cliffs, NJ: Prentice Hall.

Ross, S. M., & Morrison, G. R. (1989). In search of a happy medium in instructional technology research: Issues concerning external validity, media replications, and learner control. *Educational Technology Research and Development, 37*(1), 19–34.

Ross, S. M., & Morrison, G. R. (1991). Delivering your convention presentations at AECT. *Tech Trends, 36,* 66–68.

Ross, S. M., & Morrison, G. R. (1992). Getting started as a researcher: Designing and conducting research studies in instructional technology. *Tech Trends, 37,* 19–22.

Ross, S. M., & Morrison, G. R. (1993). How to get research articles published in professional journals. *Tech Trends, 38,* 29–33.

Ross, S. M., & Morrison, G. R. (2001). *Getting started in instructional technology research* (3rd ed.). Bloomington, IN: Association for Educational Communication and Technology. https://www.aect.org/intranet/Publications/Research/index.html.

Ross, S. M., Smith, L. S., & Morrison, G. R. (1991). The longitudinal influences of computer-intensive learning experiences on at risk elementary students. *Educational Technology Research & Development, 39*(4), 33–46.

Salomon, G., & Clark, R. W. (1977). Reexamining the methodology of research on media and technology in education. *Review of Educational Research, 47,* 99–120.

Schnackenberg, H. L., & Sullivan, H. J. (2000). Learner control over full and lean computer-based instruction. *Educational Technology Research and Development, 48*(2), 19–36.

Shettel, H. H., Faison, E. J., Roshal, S. M., & Lumsdaine, A. A. (1956). An experimental comparison of "live" and filmed lectures employing mobile training devices. *AV Communication Review, 4,* 216–222.

Shin, E. C., Schallert, D. L., & Savenye, W. (1994). Effects of learner control, advisement, and prior knowledge on students' learning in a hypertext environment. *Educational Technology Research and Development, 42*(1), 33–46.

Slavin, R. E. (1993). *Educational psychology.* Englewood Cliffs, NJ: Prentice Hall.

Slavin, R. E. (1997). *Educational psychology* (5th ed.). Needham Heights, MA: Allyn & Bacon.

Snow, R. E., & Lohman, D. F. (1989). Implications of cognitive psychology for educational measurement. *In* R. L. Linn (ed.). *Educational measurement* (3rd ed., pp. 263–331). New York: Macmillan.

Song, H. S., & Keller, J. M. (2001). Effectiveness of motivationally-adaptive computer-assisted instruction on the dynamic aspects of motivation. *Educational Technology Research and Development, 49*(2), 5–22.

Tennyson, R. D., (1992). An educational learning theory for instructional design. *Educational Technology, 32,* 36–41.

Tennyson, R. D., & Rasch, M. (1988). Linking cognitive learning theory to instructional prescriptions. *Instructional Science, 17,* 369–385.

Tennyson, R. D., Welsh, J. C., Christensen, D. L., & Hajovy, H. (1985). Interactive effect of information structure sequence of information and process learning time on rule learning using computer-based instruction. *Educational Communication and Technology Journal, 33,* 212–223.

Thompson, B. (1998). Review of *What if there were no significance tests? Educational and Psychological Measurement, 58,* 332–344.

Thompson, B. (2002). "Statistical," "practical," and "clinical": How many kinds of significance do counselors need to consider? *Journal of Counseling & Development, 80,* 64–70.

Thyer, B. A. (1994). *Successful publishing in scholarly journals.* Thousand Oaks, CA: Sage.

Ullmer, E. J. (1994). Media and learning: Are there two kinds of truth? *Educational Technology Research & Development, 42*(1), 21–32.

Underwood, B. J. (1966). *Experimental psychology.* New York: Appleton–Ceatury–Crofts.

Wiersma, W., & Jurs, S. G. (1985). *Educational measurement and testing.* Newton, MA: Allyn & Bacon.

Winn, W., & Solomon, C. (1993). The effect of spatial arrangement of simple diagrams on the interpretation of English and nonsense sentences. *Educational Technology Research & Development, 41,* 29–41.