Comparative Experiments in Educational Research

Andrew C. Porter
University of Wisconsin-Madison

Overview

Of all empirical work, comparative experiments provide the strongest evidence about the effects of education interventions. Unlike other empirical methods, these comparative experiments can supply answers to questions like the following: Does site-based decision making result in increased student achievement? Does the Cognitively Guided Instruction project increase primary school students' ability to use their mathematical knowledge to solve novel problems?

Comparative experiments also provide the best answers to theory-driven questions about cause and effect. What are comparative experiments and why are they so useful in determining whether or not causal relationships exist?

This introduction to comparative experiments in educational research is organized into four main sections. First, experiments are presented as one formal method for investigating whether or not a specific set of actions causes predictable changes in behavior. Key terms necessary for understanding experiments are defined. In the second section, requirements for arguing cause are considered, and experiments are defined in light of these requirements. Illustrations are provided of both experimental and nonexperimental research, making clear that experiments represent only one form of research.

In the third section, criteria are presented for judging the quality of an experiment. The first criterion concerns the ambiguity in deciding which of several possibilities were the causal factors in an experiment. The criterion is called internal validity. The second criterion, precision, reflects a concern for the accuracy of experimental results. The third criterion is called external validity and reflects the fact that a good experiment provides results that can be generalized for use elsewhere.

Having defined what an experiment is and how to judge its quality, the role of experiments in educational research is considered in the fourth and final section of the paper. Here, the strengths and limitations of experiments are outlined.

Starting With a Question

Much of what educators do with the belief that through their actions students will profit in some way. Not surprisingly, therefore, a great deal of educational research is conducted with an eye toward seeing whether beliefs about alternative educational practices are correct. For example, is cooperative
learning to be preferred over whole-group instruction? Should basic skills be mastered before students learn to apply knowledge, or is instruction more effective when basic skills and applications are integrated? These questions have generated a common concern for choosing among alternative educational practices. If so, will the result be better if I do something else? Experiments represent a formal method for investigating the relative merits of educational practices.

Most experiments begin with just such a broad and general question. Which is better, cooperative learning or whole-group instruction? Certainly, it would be nice to know the answer to this general question, but unfortunately, the question has many interpretations. Each interpretation might result in different answers. For example, what exactly is meant by cooperative learning? Does it mean students work in groups where each group member plays a unique role (Aronson, Blaney, Stephan, Sikes, & Snapp, 1978)? Are students graded according to the quality and quantity of group products, or are grades based only on individual work (Slavin, 1983)? And what is meant by whole-group instruction? Are you imagining teacher lectures, or are you imagining group discussions? Any of these alternative interpretations would be consistent with the initial question, and the goals of instruction have still to be considered. Are cognitive skills of interest and, if so, in what subject? Some people would be equally interested (or even more interested) in student affect, such as self-concept and how the student feels toward school and learning. Finally, under what conditions are cooperative learning and whole-group instruction to be compared? Is the interest in school learning something else, such as military training? If the initial general question brought school learning to your mind, what types of teachers and students did you imagine?

From a general question about which alternative educational practice is preferred follow a very large number of interpretations, each of which might lead to a different specific question and so a different answer. If there are different answers depending upon the specific interpretation, then the general question has no answer.

Hypotheses

Thus, a very important part of experimental research is restating the general question into one or more of its specific interpretations. Each specific interpretation is then translated into a statement of belief about alternative practices. These statements of belief are hypothesis. To illustrate, a researcher might start with the general question, "Which is better, cooperative learning or whole-group instruction?" A specific question might be, "Which is better, education in which students work in groups and in which each student plays a unique role or education in which there is group discussion?" The resulting hypothesis might be "Education in which students work in groups and each student plays a unique role is better than education in which there is group discussion."

In hypotheses, the alternative practices would represent what is called the independent variable. A variable is simply something that can differ, for example, from student to student, teacher to teacher, or school to school. Because students might be taught by either cooperative learning or by whole-group instruction, those two methods of instruction represent a variable. The goals of the methods of instruction represent what are called dependent variables. In the example about cooperative learning and whole-group instruction, the goal might be to increase reading comprehension. Reading comprehension would then be the dependent variable in the hypothesis. Because reading comprehension differs from student to student, reading comprehension is a variable.

Independent and Dependent Variables

To reiterate, experiments begin with a general question that is then made into a specific statement of belief called a hypothesis. Two important parts of a hypothesis are the independent variable and the dependent variable. Both the independent and the dependent variables represent characteristics that can be described to describe differences of interest, for example, differences among students. The independent variable is something that is believed to predict or bring about other differences (e.g., method of instruction). The resulting differences represent the dependent variable (e.g., reading comprehension).

Experiments are conducted to check the validity of a hypothesis. For example, it might be hypothesized that: Cooperative learning results in better reading comprehension than does whole-group instruction. Still, the independent variable, type of instruction, and the dependent variable, reading comprehension, need more explicit definitions prior to conducting an experiment. The objective is to state a hypothesis that is specific enough that it can be shown to be either generally true or generally false.

The idea of a statement being either generally true or generally false is very important to experiments. It must be possible to imagine a study that could lead you either to believe more strongly in your original hypothesis or to reject that hypothesis and so come to change your original belief. In short, a hypothesis must be testable. The original question about cooperative learning and whole-group instruction was not directly testable because it was too general. To answer the original question would have required considering all the different methods of cooperative learning and all the different methods of whole-group instruction. Similarly, all of the different possible conceptions of reading comprehension would need to be considered.

Population

Even given explicit definitions for the types of instruction and reading comprehension, the hypothesis remains ambiguous about the conditions for which it is to be true. The validity of the hypothesis may depend on answers to such questions: What types of students? What types of teachers? In what types of physical arrangements will students be learning? The answers to these and similar questions must be provided before the hypothesis is complete. The answers define what is called the population for the hypothesis. A population is the collection of instances for which the statement about independent and dependent variables is believed to hold. When an experiment is done to test a
found that students who have relatively positive self-concepts of academic ability have also achieved relatively well on cognitive skills in school. But these non-experimental studies leave ambiguous which came first, positive self-concept or high achievement. Even when it seems clear that the independent variable preceded differences in the dependent variable, there are an infinity of "third" variables that singly or in combination might have been responsible for a correlation between the independent and dependent variables. To further complicate matters, a lack of correlation can also result from the effects of "third" variables. For example, many people have argued that evaluations conducted to investigate the effects of Title I programs were biased in that only students most in need were eligible to participate in Title I. Because students in Title I programs started out behind, they would have done well just to catch up with students not in Title I. Campbell and Erlebach (1970) claimed that evaluations reporting no difference between students in Title I and not in Title I were really evidence in favor of program success.

Because a great deal of educational research (not just experiments) is used to support statements of cause, the three conditions for cause just given (i.e., temporal antecedence, correlation of experimental conditions and outcomes, and the lack of plausible alternative explanations) are useful in planning and understanding educational research, beyond just experiments. Even when the word cause is not used explicitly, researchers and individual readers of a research report may still infer causality in the report. For example, all educational evaluations conducted to assess an innovation or to aid decisions about a preferred educational practice are attempts to establish a causal link between practice and outcomes. Some researchers go out of their way to avoid use of the word cause because the criteria for arguing a cause are so demanding. But many words carry the connotation of cause: produce, create, induce, evoke, elicit, affect, institute, bring about. If cause is of interest, then arguing cause should be one of the explicit goals of the research. The reporting of results and the critiquing of their validity would then be done with both eyes open for plausible alternative explanations of results.

So experiments are done to test hypotheses about causal relationships between independent and dependent variables. But what are experiments and what are they not?

Experiments: A More Formal Definition

The word experiment is used a great deal, and in a number of different contexts. For purposes of describing educational research methods, placing some limits on what is meant by an experiment is useful. First, an experiment is comparative. It includes at least two different conditions represented by the independent variable, as in the example, cooperative learning versus whole-group instruction.

Second, the independent variable is under the direct control of the researcher. For purposes here, an experiment is not the study of naturally occurring changes in the independent variable; rather, it is the study of planned changes. Simply locating 20 teachers who said they were using cooperative
learning and comparing the achievement of their students to that of students of 20 teachers who said they used whole-group instruction would not be an experiment. In the words of the famous statistician George Box (1966), “To find out what happens to a system when you interfere with it you have to interfere with it (not just passively observe it).”

Third, the decision about which groups will receive which experimental conditions is made randomly. Random assignment is a process that gives each subject an equally likely chance of experiencing any one of the experimental conditions under investigation. These three limits are motivated by the three conditions described earlier for establishing cause.

Experiments: An Illustration

An illustration may help. A study was conducted to investigate the accuracy of diagnoses of school psychologists (France, Clarizio, & Porter, 1984). Briefly, it was hypothesized that the diagnosis of a student as learning disabled depends on the student’s race (black/white) and socioeconomic status (high/low) and the general achievement level of the school the student attends (high/low). The school psychologists were all from the state of Michigan and were experienced professionals. The experimental task did not involve diagnoses of actual students. Rather, a single student was represented by a file containing the full range of information that school psychologists attempt to collect prior to making a diagnosis. Thus, the main difference between actual practice and the experimental conditions was that in the experiment, the school psychologists requested information from the experimenters about a simulated child instead of working with a child.

As you have probably concluded, the independent variables of the study were the simulated case’s race and socioeconomic status and the student’s achievement level. Sometimes when the independent variables of interest are difficult to bring under experimental control, a simulation can be used to address questions of cause and effect. As is usually true, there were several dependent variables including the nature of the diagnosis and the extent to which diagnoses among school psychologists were in agreement.

By considering all possible combinations of race, socioeconomic status, and school achievement, eight experimental conditions were created. For example, one such condition was a white student from a family with high socioeconomic status who attends a school where achievement is high. Because a simulated student was used, it was possible to keep all information about the student the same for each of the eight experimental conditions except, of course, the three independent variables. School psychologists were then randomly assigned to diagnose one of the eight simulated students, and their diagnoses were recorded.

Consider again the three requirements placed on experiments. The study was comparative for each of the three independent variables. Each of the independent variables was under the direct control of the researcher. Finally, the decision about which school psychologists would diagnose which experimental conditions was made randomly. The study was an experiment. Incidentally, diagnoses were found to differ according to experimental conditions. For example, school psychologists were less likely to prescribe special class instruction for the black simulated student than for the white simulated student. Of course, the questions remain about whether or not the results from school psychologists diagnosing a simulated child represent what school psychologists do in the field.

To further understand what experiments are, consider some types of research that are not experiments. First, only research studies designed to test the validity of statements about independent variables causing changes in dependent variables are candidates for consideration as experiments. This immediately rules out large classes of research studies. For example, studies done to investigate the utility of the Graduate Record Examination for predicting success in graduate school are not experiments. In general, studies that investigate the utility and fairness of criteria used for making decisions about educational opportunities are not experimental studies. Studies with the purpose of describing education as it is are not experiments either because they lack an interest in causal relationships. A well-known example is the Gallup poll of public opinion about education.

Even research to investigate causal relationships between independent and dependent variables is not always experimental. For example, most research on the effectiveness of education programs has not been experimental, though there are exceptions. The typical study of the effects of programs involves comparing schools or classrooms using a specific program to schools or classrooms not using the program. Those using the program typically made that decision themselves, just as those in the comparison group decided not to use the program. In this research the independent variable is not under the control of the researcher, and there is no random assignment of students, classrooms, or schools. The research is not experimental. Still the interest is in determining whether or not the program produced the benefits anticipated, a causal question. Nonexperimental studies of causal relations must find ways to rule out “plausible third variable” explanations of results. For example, if the program schools look better, is it because of the program or is it because the program schools were volunteers anxious to improve in whatever ways they could?

Another common method of investigating program effects is to test student performance against the goals of an innovation, give the students the innovation, and then test the students a second time. For example, the innovation might be some form of computer-aided instruction. Although such a study gives the researcher direct control over the independent variable (before innovation versus after innovation), it does not involve random assignment of students to those two conditions. Because most students benefit at least to some extent from any reasonable type of instruction, the question of cause must be, did students gain more in achievement from the computer-aided instruction than they would have gained otherwise? Because all students are first without the innovation and later all students receive the innovation, the design fails to provide an estimate of what students would have gained from “regular” instruction. An experiment would randomly assign some students to receive the innovation and other students to receive regular instruction.
COMPLEMENTARY METHODS

Three Goals of Experimental Design

Simply being able to distinguish between what is and what is not an experiment is not terribly useful. What is more important is knowing how to assess the strengths and weaknesses of an experiment.

To understand the advantages and difficulties of experiments for testing hypotheses about cause-and-effect relationships, it is useful to understand three goals of experimental design. The first goal is directly related to the interest in cause and is sometimes stated as, an experiment should have internal validity. Consider again the hypothesis that cooperative learning results in better reading comprehension than does whole-group instruction. To test this hypothesis, a researcher might have some students taught by a particular method of cooperative learning and some other students taught by a particular method of whole-group instruction. Students in both groups would then be given a test of reading comprehension. If method of instruction is the only plausible interpretation for any differences in reading comprehension between the two groups of students in the study, then method of instruction must be the cause of those differences. If the independent variable is the only reasonable explanation for differences in the dependent variable, the study is said to have internal validity.

The second goal of experimental design reflects the possibility that differences among individual students on reading comprehension may be considerable, regardless of the method of instruction used to teach these skills. For example, even if cooperative learning was somewhat better than whole-group instruction, it would be quite likely that several of the best students who received whole-group instruction would have better reading comprehension skills than several of the poorest students who received cooperative-learning instruction. The test results for these students would make more difficult the objective of determining whether whole-group instruction or cooperative learning was better. The second goal of experimental design, then, is to conduct a study in which even small differences caused by the independent variable are measured with sufficient accuracy that they will not be overlooked, even though they are embedded in relatively large individual differences among subjects. This goal is called precision.

The third goal of experimental design is that valid generalizations can be made from the study. For example, if cooperative-learning instruction were superior to whole-group instruction in a particular study, the next question would be, Does the finding generalize to other students, teachers, and conditions? To the extent that the study had external validity, at least some of these desired generalizations would be appropriate. The issue of external validity was raised with the results from the simulation experiment of diagnosing learning disabilities given earlier.

When interpreting the results of an experiment or when designing an experiment, the experiment should be considered from the perspectives of internal validity, precision, and external validity. Each of these three goals is discussed separately in greater detail.

COMPARATIVE EXPERIMENTAL METHODS

Internal Validity

Judging the extent to which you believe an experiment has internal validity is equivalent to deciding how much of the correlation between the independent and dependent variables you believe was caused by the independent variable. How effective was the design of the experiment in allowing you to rule out explanations, other than the independent variable, for differences between experimental conditions on the dependent variable? Clearly, if the brightest students received cooperative-learning instruction and the poorest students received whole-group instruction, a finding in favor of cooperative learning would be suspect.

Confounding Variables

The most important concept for judging the internal validity of an experiment is the concept of confounding variables. A variable is said to be confounded with the independent variable of a study if the two variables are inseparable. In the example above, the brightness of students was confounded with the method of instruction. Students' brightness is then an explanation for differences between the two groups of students taught by different methods of instruction. Because method of instruction is also an explanation, the two explanations are confounded. An experiment has internal validity, therefore, to the extent that no variables are confounded with the independent variable.

Random Assignment

As stated previously, a crucial part of the definition of a comparative experiment is that the experimenter uses random assignment in determining which subjects will experience which experimental condition. (Subjects need not be individual students and can be classrooms or even whole schools.) The essential idea behind the process of random assignment is that, for an initial pool of subjects, each has an equally likely chance of being assigned to any one of the experimental conditions to be compared. Random assignment is, therefore, a way of guarding against confounding variables.

Before considering the strengths and weaknesses of random assignment of subjects to experimental conditions, a brief description of the process is appropriate. Imagine that you have 40 subjects that you wish to randomly assign to one or the other of two experimental conditions. You need a process that gives each of the subjects an equal chance of being assigned a particular experimental condition. One method for random assignment is to use a table of random numbers. The method begins by putting the subjects' names in a list in alphabetical order. A subject can then be represented by a number indicating that subject's position on the list. The table of random numbers has been prepared such that each number has an equally likely chance of having been placed anywhere in the table. The researcher begins reading the table in any row and any column. Reading from the table of random numbers, the first 20 numbers between 1 and 40 indicate the positions of subjects on the list of subjects to be assigned one of
the two experimental conditions. The remaining subjects are assigned to the
other experimental condition. Random assignment is a process that ensures each
subject's having an equally likely chance of being assigned to each experimental
condition. Random assignment is not accidental, fortuitous, or casual.

In thinking about the value of random assignment in eliminating confound-
ing variables, a few points are worth emphasizing. First, random assignment
takes place at the beginning of an experiment and is followed by the experi-
mental conditions. The dependent variable is not observed until the end of the
experiment. Thus, for experiments, the independent variable precedes the
dependent variable. The temporal antecedence requirement for arguing caus-
ality unambiguously established. In this sense, all experiments include at least the
sequential activities: (a) random assignment, followed by (b) application of
experimental conditions, and finally, (c) observation of the dependent variable.
Second, randomization is a process of assignment and not a result of the
experiment. It is virtually impossible to look at the composition of experimental
groups created by someone else and from that to judge accurately whether or
not they were created randomly. Third, randomization creates comparison
groups at the outset of the experiment that differ only by chance on all possi-
ble confounding variables. Randomization not only eliminates the confounding
variables that the researcher may have considered, but it also eliminates the con-
founding variables the researcher may have overlooked. This is the real power
of random assignment.

The utility of random assignment for controlling confounding variables must
of course, be tempered by the realization that the process is based on chance
and by chance alone experimental groups will differ at least to some extent. For
example, an experiment used to compare cooperative learning and whole-group
instruction would begin by identifying a large enough pool of classrooms to
institute both methods of instruction. If the initial pool for assignment consisted
of only two classrooms, however, the two groups would differ at the outset of
the experiment to the same degree that the two classrooms differed (and that
might be considerable). In such a case, many characteristics of the two class-
rooms would be totally confounded with the independent variable and would
constitute a major threat to the goal of providing a strong statement about
cause. If there were four classrooms in the initial pool, the chances would be a
little less that the best students (or the best teachers) would end up in a single
group. The more classrooms in the initial pool, the smaller the chance that there
will be worrisome differences between experimental groups at the outset of the
experiment. Still, chance plays a part in the final results of any experiment.

Assessing the Utility of Random Assignment

Imagine 20 students randomly assigned to receive a version of cooperative
learning and 20 students to receive a version of whole-group instruction. One
method for delivering the two modes of instruction would be to form two
experimental classes of 20 students each and then have one teacher teach the
cooperative-learning instruction class while another teacher taught the whole-
group instruction class. Even though students were randomly assigned to the
two classes and even if the two teachers were randomly assigned to classes, the
characteristics of teachers would be confounded with methods of instruction.
For most people, this would represent a serious threat to internal validity. One
teacher might be more effective, regardless of method of instruction.

To unconfound teachers from experimental conditions would require identi-
fying several existing classrooms of students each with a different teacher and
then randomly assigning classrooms to methods of instruction. Each method
would then be represented by several teachers. Of course, the need for several
teachers increases the cost of the experiment. You may wonder why one teacher
couldn't teach both classrooms. Although using one instructor would prevent
teachers from becoming confounded with the method of instruction, the strategy
introduces another confounding variable. If one teacher taught both class-
rooms, then one classroom would have to be taught first and the other second.
Among other difficulties, this confounds both order of presentation and time
of day with the independent variable, methods of instruction. The possibilities
for confounding variables to enter a design are virtually unlimited.

The internal validity that is gained through random assignment of subjects to
experimental conditions is the real strength of experiments, compared to other
research methods. Consider a typical study of teaching practices. Several class-
rooms and their teachers are recruited, and students are tested on school
achievement variables in both fall and spring of a school year. The researcher
observes teaching practices during the school year and attempts to relate them
to student gains in achievement. Useful as these studies may be, they always suf-
fer from the presence of confounding variables. There is always ambiguity in
deciding whether teaching practices caused the relationships between teaching
practices and student achievement. What confounding variables are you imag-
ining? Probably the first one that came to your mind was student aptitude. The
most able students will probably have the biggest achievement gains regardless
do the quality of the instruction they receive. There will be a tendency for the
behavioral outcomes of teachers who have the most able students to appear as
though they produced the greatest gains in student achievement. Teachers of
the most able students may have actually employed the most effective teaching
methods, or whatever teaching practices were used with the most able students
appeared to be most effective. It might even be that able students cause teach-
ers to follow certain instructional practices (e.g., teaching at a faster pace or
giving more frequent positive feedback to students). The implications for practice
are quite different if one believes that faster paced instruction produces more
student learning than if one believes that teachers who have more able students
cover more content. There are, of course, many other potentially confounding
variables for this nonexperimental design.

The ever-popular pretest/posttest design is yet another context in which to
understand the implication of confounding variables. One group of students
takes a test on the dependent variable, it then receives an educational innova-
tion, and finally the students are tested once more on the dependent variable.
The researcher is interested in student gains on the dependent variable and would like to attribute those gains to the innovation. But what are the possibilities for confounding to have occurred between the two times of testing? Campbell and Stanley (1963), in their classic review on experimental and quasi-experimental designs, provided a taxonomy of the types of possible confounding. Students may do better on the second test because they learned from taking the first test; the second test may be different from the first in some way that makes it easier; the students themselves are older and more mature; or during the course of the study the students may have received some experiences (other than the experimental condition) that changed them and improved their performance on the dependent variable.

Experiments are the single most effective method for ruling out confounding variables, but experiments are not infallible indicators of the presence or absence of causal relationships. Randomization only starts experimental groups out in the right way. Absence of confounding variables at the beginning of an experiment does not guarantee absence of confounding variables at the end of an experiment. There are plenty of opportunities for important confounding variables to creep into the design while the study is being conducted. Subjects may drop out of the experimental groups at different rates or for different reasons. Observations on the dependent variables may inadvertently result in observers being confounded with experimental conditions or even reflect experimenter bias. Data can be incorrectly coded in ways that create confounding. In fact, anything that happens systematically during the course of the study and results in a different treatment of subjects in one experimental condition from subjects in another must either be included in the definition of a treatment or become a confounding variable.

There is no foolproof method for ensuring internal validity. Experiments are a big help, but even experiments differ greatly in their ability to convince. One hopes that, as confounding variables are identified (usually in retrospect), their probable effects can be judged either by common sense or from the research literature and in that way be taken into account.

**Precision**

For an experiment to have internal validity is important, but that alone is not enough. An experiment must also be designed so that, if the independent variable causes differences in the dependent variable, those differences will be detected. In short, the experiment must have precision.

Because of random assignment, chance plays a part in the results of any experiment. Through random assignment there is always the chance that by bad luck all students of a particular type will be assigned to the same experimental condition. One way to think about precision is that the more precise an experiment, the less likely it is that the experiment will yield large chance differences between experimental groups.

To better understand the concept of precision, it may be useful to digress a bit and consider the typical ways that results of experiments are reported. At its simplest, an experiment can be thought of as describing the average difference between two experimental conditions on a dependent variable (e.g., student achievement). Thus an experiment might report that, on the average, students who received cooperative-learning instruction scored four points higher on a particular test of reading comprehension than did students who received whole-group instruction. Because classrooms were randomly assigned to experimental conditions, an important question remains. Is it reasonable to believe that cooperative-learning instruction was more effective than whole-group instruction, or should the experimenter conclude that the four-point difference was likely to happen by the chance results of random assignment? Deciding between these two alternative interpretations of the four-point difference is an important aspect of experimental research.

**Statistical Significance**

The results of an experiment are often reported as statistically significant or not. If they are statistically significant, the researcher decided that the difference between experimental conditions did not happen by chance. If they are not statistically significant, the study failed to provide evidence of the superiority of one experimental condition over another. The rules for deciding whether or not the results of an experiment were statistically significant are called the procedures of inferential statistics.

Briefly, experimenters use the rules of statistical inference to decide for their particular studies how large the observed difference between experimental conditions must be to be judged statistically significant. Of course, the more certain the experimenter wishes to be that a difference called statistically significant was not due to chance, the larger the required difference must be. Without going into details, the procedures of statistical significance begin with an experimenter stating how willing he or she is to make a mistake when concluding that a difference does exist. A common criterion is to take a 5% chance of making such a mistake.

Returning to the concept of precision, sometimes experiments result in quite large differences between experimental conditions, and still the rules of statistical inference lead to a decision of no significance. This is a particularly frustrating situation. Even if the difference appears large, if it is no larger than might be expected by chance, the researcher is forced to conclude that the independent variable did not affect the dependent variable. The only information gained is that the experiment lacked precision. The experiment should not have been conducted because the unacceptable precision could have been predicted.

In short, the second goal of experimental design is to conduct a study such that if a "large difference" is found between experimental conditions, that difference will also be judged by the rules of statistical inference to be significant.

**Judging the Importance of a Difference**

Before considering ways in which the precision of an experiment can be enhanced, a comment on the cavalier use of the term large difference is in order. What may be a large difference in the eyes of one person may not be a large difference in the eyes of another. How would you go about deciding what consti-
COMPARATIVE EXPERIMENTAL METHODS

were willing to take a 5% chance of mistakenly thinking a difference was real, this
same researcher might decide instead to build a 95% confidence interval.

If a confidence interval is used to report the results of an experiment, precision

can be thought of in terms of the width of the interval. The narrower the
interval, the more precise the experiment is. An experiment with perfect pre-
cision would yield a confidence interval with no width at all. The interval would
be a single point.

Improving Precision

So precision is an important attribute for any experiment to have. Without
precision, even an internally valid experiment may fail to detect important dif-
erences between experimental conditions and leave the mistaken impression
that the various conditions lead to equivalent results. But how can the precision
of an experiment be made acceptable?

The most direct method for increasing the precision of an experiment is to
increase the number of subjects assigned to each experimental condition. As
indicated earlier, the larger the number of subjects to be randomly assigned, the
less likely it is that random assignment will result in unusual groups of subjects.
Increasing the number of subjects is often expensive, however, and there are
other ways to improve precision.

A straightforward way to make sure that not all of the most able students (or
classrooms) are assigned to a single experimental condition is first to group stu-
dents according to aptitude. Then, for each group of students at a specified
aptitude, randomly assign equal numbers to each experimental condition. This
insures that each experimental condition has subjects with similar aptitude lev-
els. The procedure is called blocking and is one way to improve precision.

There are several other ways to improve precision, some of which are quite
complicated and technical. For illustration, consider just one more. Clearly, if
all subjects for an experiment were identical, random assignment could not
result in an unusual composition of experimental groups. Precision would be
assured. You are probably saying that identical subjects are an impossibility. In
a sense you are right. Still, some researchers have used identical twins as sub-
jects in an attempt to improve precision. My earlier example of an experiment
to study the accuracy of school psychologists provides another illustration. In
that study school psychologists diagnosed a single simulated student. Using a
simulated student increased precision over what it would have been had each
psychologist diagnosed a different actual student.

External Validity

The first two goals of experimental design, internal validity and precision, are
primarily concerned with interpreting the results of a particular study as it was
conducted. But an experiment would have little utility if the results only
revealed what was true for that particular study. The third goal of experimental
design, then, is to have an experiment from which it is possible to make valid
generalizations. The goal is called external validity. The word external indicates

COMPLEMENTARY METHODS

nutes a large difference in average reading comprehension between cooperative
learning and whole-group instruction? Would you consider the four-point dif-
ference I mentioned earlier to be important if you knew it had been caused by dif-
ferences in the method of instruction? Of course you can't say. You are probably
wondering about the types of items on the test of reading comprehension. You
may also be wondering what types of decisions are to be made about school prac-
tices based on the four-point difference. You may even be wondering what type
of students were involved in the study. Deciding what constitutes large or impor-
tant differences is a highly personal activity. There is no right or wrong answer.

Saying that a difference between experimental conditions was statistically sig-
nificant is not synonymous with saying that a difference was large or important.
Using statistical significance as an indication of importance is a common mis-
take that should be avoided. Statistical significance only indicates that a differ-
ence was not due to the chances of random assignment. A statement of statistical sig-
nificance leaves completely unanswered the question, How important was the
difference that can be attributed to experimental conditions? It is possible to
design a study that is so precise that a difference as small as a fraction of a test
score point would be judged significant by methods of statistical inference.

Confidence Intervals

Although it has been common practice to report the results of an experiment
as simply statistically significant or not, many argue that this practice does not
go far enough (e.g., Carver, 1978). First, that kind of reporting encourages the
belief that statistical significance implies importance, which isn't true. Second,
a researcher is not likely to be satisfied by knowing simply whether or not a dif-
erence happened by chance. A researcher will probably want to have some idea
about how big the difference was. For these reasons, many researchers prefer to
report the size of difference found between experimental conditions and then
to indicate that because of random assignment the difference might have been
a little higher or a little lower. The procedure of establishing upper and lower
limits on an observed difference between experimental conditions is called
building a confidence interval. Confidence intervals represent an alternative to
reporting results as simply statistically significant or not.

Basically, the same procedures are used to build confidence intervals as are
used to decide whether or not a difference was statistically significant. Returning
to the example of a four-point difference between cooperative learning and
whole-group instruction, a confidence interval could have indicated that the true
difference between instructional methods might be as low as three points or as
high as five points. This information indicates that there was a statistically signi-
ficant difference between methods of instruction and also provides a range of val-
ues, three to five, for thinking about how big the difference actually was.
Alternatively, had the interval included zero (i.e., no difference between meth-
ods), the conclusion would have been that the difference was not statistically sig-
ificant. Just as deciding on statistical significance requires setting a criterion
level for chance errors, so does building a confidence interval. If a researcher
COMPLEMENTARY METHODS

an interest in concluding that the findings of a particular study have validity beyond or external to that study.

One way to think about external validity is to imagine a particular experiment and then attempt to identify situations in which you believe the results might apply. To illustrate the point, consider again the experiment to investigate the accuracy of school psychologists' diagnoses. The experiment used a simulated learning-disabled student to investigate the effects of student race, socioeconomic level, and school general achievement level on school psychologists' diagnoses. The school psychologists in the study were experienced professionals recruited from school districts in Michigan. School psychologists were randomly assigned to diagnose one of the eight versions of the simulated student. To what extent is the result generalizable to the black simulated student? To what extent would you be unwill- ing to generalize the result, the experiment lacked external validity for you.

In the words of Bricht and Glass (1968):

"Threats to external validity appear to fall into two broad classes: (1) those dealing with generalizations to populations or conditions (What population of subjects can be expected to behave in the same way as did the... experimental subject?) and (2) those dealing with the "environment" of the experiment (Under what conditions... can the same results be expected?). (p. 438)"

The results of the school psychologist experiment may be valid for school psychologists outside of Michigan and for inexperienced school psychologists, but there is no evidence from the experiment that this is so. If generalizations are to be made to other states or to inexperienced school psychologists, the grounds for such generalization would need to be based on information not provided by the experiment.

Random Selection

Depending on how school psychologists were recruited, it might not even be appropriate to generalize the results of the experiment to all experienced school psychologists in Michigan. In an ideal experiment, all experienced school psychologists in Michigan would have been identified and put on a list. The list would represent the population of school psychologists referred to in the hypotheses for the experiment. Participants in the experiment would then be selected from the list, using a process that gave each person on the list an equally likely chance of being selected and ultimately included in the study. The process is called random selection. Random selection from a well-defined population is ideal, because the process insures, within the limits of chance, that the participants in the study are representative of the population from which they were selected. Had random selection been used to select school psychologists, the results of the study could be straightforwardly generalized to the population of experienced school psychologists in Michigan.

Do not confuse random selection with the earlier concept of random assignment. Random selection and random assignment serve two distinctly different purposes. Random assignment is motivated by internal validity, and random selection is motivated by external validity. Random assignment is an essential ingredient of an experiment, but because of costs and subject availability, random selection is almost never part of an experiment's design. Rather, subjects are recruited as best they can be. In the case for the experienced school psychologists, volunteers were recruited from the pool of graduates from several different degree programs.

Generalizing Across Subjects

How then should the external validity of the experiment with school psychologists be thought about? Is generalization limited to the subjects immediately involved? More generally, because random selection is rarely used in experiments, are the results of experiments typically limited to the subjects used? Most people answer these questions with a decided, "No." People who are willing to generalize from an experiment, however, require a careful description of the subjects. The description defines a hypothetical population to be used for purposes of interpreting results.

We already know that the subjects were experienced school psychologists working in Michigan. But clearly we need to know more about these psychologists if we are to generalize the results of the experiment. What was meant by the word experienced? What were the theoretical persuasions of the school psychologists? In general, the more complete the description, the easier it becomes to think about the range of external validity of the experiment. Given a thorough description of the subjects in the experiment on school psychologists, people will differ in their willingness to generalize the results. Some people will conclude that the results are even valid for states other than Michigan.

So the results of an experiment may be valid for types of subjects that were not included in the experiment. There is another way, however, that external validity may be threatened even for the types of subjects used. Imagine that the experienced school psychologists could be categorized into two groups. For example, half the school psychologists may have received their training from one institution and half from another institution. School psychologists from both institutions are represented in the study, and so, on the surface, it would appear that the results can be generalized to both types of school psychologists. But the results of experiments are reported as averages across all subjects. An experiment may indicate no average difference between experimental conditions when in fact there was a difference between experimental conditions for each of the two types of subjects. This apparent paradox will occur when the difference between the averages computed. When the size and/or direction of an experimental effect depends on some characteristics of the subjects (e.g., the institution from which they received their training), there is said to be a treatment by subject characteristic interaction. More generally, interactions limit (or at least describe) the external validity of an experiment. When interactions are hypothesized, the
COMPLEMENTARY METHODS

experiment should be designed so that the interaction hypothesis can be tested. The straightforward way to test for interaction is to first block on the subject characteristic and then, within each block of subjects, to randomly assign subjects to experimental conditions. In the school psychologist example, subjects could have been blocked by the institution from which they received their training and then for each block separately randomly assigned to the eight experimental conditions.

Generalizing Across Experimental Conditions

Threats to external validity are not limited, however, to concerns about types of subjects. One must also consider the conditions under which the experiment was conducted.

The school psychologist experiment provides fertile ground for illustrating several ways in which experimental conditions can limit external validity. Remember that school psychologists diagnosed a simulated student who was the same in all eight experimental conditions. Clearly, only one example of learning disability was considered. If the results of the study are to be generalized, one would need to know the nature of that specific example of learning disability. The experiment provides no direct evidence about whether the same results would have been obtained for a different type of learning disability.

The school psychologist experiment is also an example of what might be called a laboratory study. The subjects were not working in their natural environment with real students. There is no question but that they realized they were in an experiment, although they were not told the purposes of the experiment. Still, one must at least raise the question of whether their behavior in the simulation was representative of how they would behave as psychologists in their school districts. Did the subjects merely respond as they felt they should have responded and not as they would ordinarily? Because they were confronted with only a single student, did they perform their work more carefully than when they are under the pressure of a heavy work schedule? Did the experimenter inadvertently reveal a desired set of responses? Unfortunately, the experiment did not provide answers to these questions. Each of us must decide the answers on our own. The point is that these and similar questions should be raised when judging the external validity of any experiment.

More generally there can sometimes be tension between designing an experiment that has good internal validity and precision and designing an experiment that has good external validity. In the school psychologist experiment, a simulated case increased internal validity through increased experimenter control and increased precision by using a single simulated case to control variance. At the same time, the simulation aspect of the experiment raised a number of serious questions about whether the results could be generalized to practice. Every empirical investigation suffers from some weaknesses that call into question the results, no matter how carefully designed and conducted. Experiments are no exception, even though experiments have unique strengths for investigating causal relationships. The best way to deal with the tensions among internal validity, precision, and external validity is to conduct multiple experiments, all of which investigate the same general question but each of which is designed to have unique strengths and few if any overlapping weaknesses.

Generalizing Across Outcomes

External validity can also be limited to the extent that results apply for only certain narrowly defined dependent variables. To illustrate this point, consider again the example of cooperative learning and whole-group instruction. The dependent variable was reading comprehension, but how reading comprehension was to be measured was not made explicit. Suppose the task for each subject was to read some short paragraphs and then to answer a series of multiple-choice questions. If so, how difficult were the passages to be read? What vocabulary was included in the tasks? What was the nature of the multiple-choice questions? Was the subject also asked to demonstrate reading comprehension of longer passages, perhaps a whole book? What type of book? Was comprehension assessed through an oral examination? Was the subject asked to demonstrate that the information acquired from a passage could also be put to use in solving problems? Reading comprehension can mean many different things. A good experiment would include several definitions of reading comprehension as multiple dependent variables so that external validity could be assessed directly.

Experimental conditions can also lead to unintended results. The method of instruction that promotes greatest gains in reading comprehension might at the same time leave subjects with a distaste for reading. Thus, an experiment should reflect a concern for unintended results by including appropriate dependent variables.

Finally, differences on a dependent variable (e.g., student achievement) observed immediately following the experiment may not be present one year later. Education is typically interested in practices that promote sustained effects. It may not be valid to generalize from measures taken immediately at the end of an experiment to what will be true weeks, months, or even years later.

Of course, it is impossible to measure subjective things on everything that might be of interest. Nevertheless, one should be cautious about generalizing the results of experiments using one dependent variable to what is true for another.

Limitations to Experiments

Now you know what experiments are and how to think about their quality against the three criteria: internal validity, precision, and external validity. What remains to be done is to provide a sense of the role of experiments in educational research.

Among all the methods of educational research, experiments provide the most unimpeachable evidence of whether or not an independent variable causes differences in a dependent variable. The key to the internal validity of experiments lies with random assignment. Without random assignment there will always be variables that are confounded with the independent variable and,
therefore, offer alternative explanations of results. Explanations of cause have greater ambiguity when random assignment is not used.

But comparative experiments are not the sole methodology of educational research. One obvious reason is that not all educational research is interested in questions about what causes what. Even if cause is of interest, however, research methods other than comparative experiments are frequently used. The reasons for this lie primarily with the nature of the research questions asked.

By definition, experiments involve random assignment to experimental conditions. Thus to conduct any experiment, the researcher must first create experimental conditions. Experiments are anticipatory, not retrospective. They are not appropriate for investigating the antecedents of a historical event. For example, at this time, experiments cannot be used to investigate the effects of the "new mathematics" curriculum reform.

Even when an educational event has been anticipated, experiments are not always possible. The researcher must have the authority to randomly assign subjects. Imagine that a teacher strike is anticipated in a school district, and you are interested in determining the effects of the strike on student achievement, teacher/principal relations, and the like. An experiment would require a comparison of striking teachers and nonstriking teachers, and the comparison would need to be built through randomly assigning teachers (or school districts) to the two conditions. Can you imagine an experimenter having the authority to tell some teachers to strike and others not to strike? Here, the requirements of an experiment are clearly impossible.

Finally, experiments necessarily focus on a few selected variables. Because creating experimental conditions and random assignment is difficult to accomplish, the typical experiment investigates the causal effects of only a single independent variable. (There may, of course, be other independent variables in the design of the study for purposes of testing interactions that describe the range of external validity.) Even the selection of dependent variables is restricted by the requirement that they be formally observed in a way that is common to all subjects. Experiments attempt to facilitate understanding of a few variables well. Sometimes the goal of educational research is to understand a community or a school so completely that, regardless of the situation, it would be possible to accurately predict the actions of the individuals involved. Anthropologically oriented research of this type requires the consideration of many variables simultaneously. Potentially, many of the variables have causal influences on the behaviors of individuals involved. Experiments do not lend themselves well to answering questions about complicated causal models that involve many independent variables, which are all believed to influence a dependent variable, some directly and some indirectly through influencing other independent variables.

Using Experiments

If one defines educational research as that research of potential relevance to improving understanding of the processes of education, then presumably a considerable portion of research conducted in the following disciplines is of interest: psychology, sociology, anthropology, economics, history, philosophy, and political science. Of these disciplines, psychology stands out as the predominant user of comparative experiments.

Moving beyond the disciplines, a great deal of important educational research is conducted within an educational context by educators. Much of this research has been conducted with an eye toward developing new programs, curricula, and teacher training procedures. Historically, these efforts have been heavily influenced by psychology, and not surprisingly they evidence considerable use of comparative experiments. However, there has been a strong tendency to drop the requirement of random assignment, leading to a distinction between true experiments, the topic here, and quasi-experiments, which are comparative studies without random assignment. Educational researchers have been too quick to conclude that random assignment was not possible. The overreliance on nonrandomized quasi-experiments has led to a great many results that leave ambiguous the causal antecedents of behavior.

It is impossible to define the range of educational research questions that might profitably be addressed through experiments. Experimental studies have been conducted to investigate the effects on achievement of peer tutoring, teacher questioning techniques, sequencing of instructional material, types of teacher lecture notes, different types of feedback on past achievement, and self-paced versus instructor pacing. Experimental studies have investigated the effects on reading comprehension of positioning of questions in text; organization of prose using semantic, temporal, and random strategies; and different types of pictures to accompany the text. There have been experimental studies of the effects of bimodal learning teams on race relations as well as experimental studies on the effects of achievement and aptitude testing on a variety of dependent variables, including teachers' expectations for student achievement.

Experiments have been used to determine the effects of education programs. (An example is provided in the reading that follows this chapter.) Unfortunately, experiments to determine the effects of education programs are not as prevalent as needed. The result has been the widespread promotion of education programs without solid evidence that they have the touted effects. Today school restructuring provides a disappointingly large number of examples. Hopefully, many of these "name brand" school improvement efforts are effective. If they are not, most will ultimately be tossed on the garbage pile of education facts, but not without having first wasted the time, energy, and money of a good many well-intentioned educators. A strong commitment to requiring that education innovations first prove themselves through experimental validation might slow the reform process a bit, but at the same time make the reform process much more effective.

Summary

The term experiment means many different things to many different people. In this chapter the term has been limited to a rather narrow, but useful, definition—an investigation, involving random assignment, of the causal relationship between one or more independent variables and one or more dependent variables. The methodological quality of an experiment should be judged
Suggestions for Further Reading

A great deal has been written that is relevant to acquiring a better understanding of experiments in educational research. Virtually all textbooks designed to provide an introduction to research methods in education contain a chapter on experiments or experimental methodology. The following list contains some of the better known introductions to educational research. I suggest that you read the chapters on experimental research in two or three of them.


Many textbooks on statistics in psychological or educational research contain chapters or sections that are relevant to the design of educational experiments. Some of these textbooks even contain the words experimental design in their title although, in my opinion, they are more textbooks on statistics than they are textbooks on experimental design. Although knowledge of statistics is extremely useful for designing or interpreting educational experiments, a separate body of knowledge is appropriately labeled experimental design. Following is a list of some statistically oriented textbooks for educational researchers as well as some of the classics on experimental design. These references will be of interest if you have some background in inferential statistics and if you plan to conduct experiments in the future.


References


A most important part of becoming familiar with experiments in educational research is to read reports of studies that have used comparative experiments in at least one aspect of their design. The following reading provides one opportunity. The best approach, of course, is to look for examples of experiments that have been conducted in your substantive areas of interest. A good exercise would be to select a few issues that seem to be of interest and see if you can identify examples of comparative experiments in the journals that dominate your field of work. For the experiments identified, see if you can judge the quality against the three main goals of experimental design.

**Study Questions**

1. What is a hypothesis and how is a hypothesis used in experimental research?
2. Define the terms independent variable, dependent variable, and population.
3. What is meant by a causal relationship? What are the requirements for arguing cause?
4. What is an experiment? Give an example of an experiment and an example of a study that is not an experiment.
5. State the three criteria for judging experimental research and define what is meant by each.
6. Why are confounding variables of concern to experimenters?
7. What is random assignment, and why is it useful?
8. What is meant by the description of a difference as statistically significant? Why is statistical significance not synonymous with importance?
9. State at least two ways to improve the precision of an experiment.
10. What is random selection, how does it differ from random assignment, and why is random selection important?
11. Describe at least three threats to the external validity of an experiment. How can an interaction between the treatment independent variable and another independent variable help to define the range of external validity?
12. Give two illustrations of when experiments would be an appropriate research method.
13. Give two illustrations of when experiments would not be an appropriate research method.
Reading

Introduction
A Comparative Experiment in Educational Research

Andrew C. Porter
University of Wisconsin-Madison

In the following reprint of an article from the American Educational Research Journal, Carpenter, Fennema, Peterson, Chiang, and Loef report findings from a well-designed and well-executed experiment to test the effects of Cognitively Guided Instruction (CGI), a teacher education program designed to improve student achievement in the areas of addition and subtraction. The article is an excellent illustration of how research builds upon results from previous research and also the progression from basic research to increasingly more applied research and, ultimately, into classroom practice. Incidentally, the article also illustrates that education experiments are sometimes planned, carried out, and reported by a team of researchers and that sometimes education experiments are supported by federal funds. Most important to purposes here, however, the article illustrates many of the concepts and terms introduced in the preceding chapter on comparative experimental methods in education research.

Perhaps unfortunately for purposes here, the article illustrates that data from experiments are typically analyzed with fairly sophisticated statistical procedures. Clearly, now is not the time or place for a minicourse on inferential statistics. Still, it may be helpful to know the following:

- Reliability (Cronbach's Alpha) indicates the extent to which a variable has been measured with precision (in the sense that if the same subject were measured in the same way again, the same results would be found). Reliability can be any value from 0.0 to 1.0, with higher values meaning better reliability.
- Standard deviation is a descriptive statistic indicating the amount of variability or dispersion in scores on a variable. For example, in Table 4, CGI classrooms spent, on average, 60 percent of time in whole-class instruction. Around this average, however, classrooms differed considerably. The reported standard deviation was approximately 23. For most distributions, a range around the mean of a 2 standard deviations includes 90% or more of the observations.

- Several different tests of significance were used, depending on the dependent variables that were analyzed. Tests were used for observation data, analysis of variance was used for teacher belief data, and analysis of covariance was used for student achievement data.

Each technique is used to decide whether mean differences between CGI and control classrooms were sufficiently large that they could not be dismissed as simply having occurred by chance. Results were reported at both 0.05 and 0.01 levels of significance, allowing readers to decide which of the two chances they want to take in concluding that the treatment had an effect, when in fact it did not.

The study of Cognitively Guided Instruction is of interest in its own right, but the purpose here is to illustrate the methodology of comparative experiments. With that in mind, the following may be worth thinking about while reading the article:

1. Like most experiments, the researchers were originally motivated by a general question, "Whether providing teachers access to explicit knowledge derived from research on children's thinking in a specific content domain would influence the teachers' instruction and their students' achievement." The researchers move from there to testable hypotheses. What are the hypotheses tested? What are the independent variables, the dependent variables, the population?
2. Are the hypotheses (or the initial motivating question, for that matter) about causal relationships of independent variables on dependent variables? In short, are the purposes of the study consistent with the purposes of comparative experiments?
3. Does the study satisfy the requirement of a comparative experiment, as presented in the preceding chapter? Where did random assignment enter the design?
4. Internal validity is one of three main criteria for judging the quality of an experiment. Beyond random assignment, were any measures taken to see that confounding variables did not occur? For example, is it possible that the observations of classrooms and/or the interviews of students and teachers, might reflect an experimenter bias in favor of CGI? The possibilities of experimenter bias entering a study during measures of dependent variables are often controlled through using observers that are "blind" as to whether or not they are observing a person in the treatment group or the control group.
5. Most of the conclusions reached in the study are based on comparisons between CGI and a control group. In the discussion section, however, the following is stated: "Both CGI and control teachers increased in their agreement with the perspective that children construct mathematical knowledge. These results suggest that both the CGI workshop and the problem-solving workshop for the control teachers had an effect on teachers' beliefs about children's problem solving." Is this conclusion based on the experimental portion of the study, or is it represented on a weaker pretest/posttest comparison? What, if any, confounding variables might offer an alternative interpretation to the data?
6. Another criterion for judging the quality of an experiment is precision. Clearly, the study had sufficient precision for achieving statistically significant results for some but not all of the dependent variables. Several techniques were used to give the study good precision: relatively large sample sizes (e.g., 40 classrooms with 12 target students from each), highly reliable measures of dependent variables (most of the reliability coefficients were above 0.80), blocking on schools when assigning teachers to experimental conditions, and using prior achievement as control variables in analysis of covariance when analyzing achievement test results.

7. The authors reported CGI and control means on each dependent variable and tests of statistical significance. They did not report confidence intervals. Do you believe the authors provided sufficient help to the readers to judge for themselves whether the statistically significant differences among treatment conditions were also important in some educational sense (as opposed to real but trivial differences)?

8. Did the researchers randomly select subjects from well-defined populations? Regardless of your answer, how would you judge the quality of this study against the criterion of external validity? One of the strengths of the study was the inclusion of several types of dependent variables (across types of outcomes). A similar strength was that the authors checked for an interaction between treatment conditions and prior levels of student achievement. Good experiments should contain design features that help test the limits of external validity in just this way. Reports should also describe the research procedures in sufficient detail that readers can make educated guesses about external validity on other dimensions not explored systematically in the design. Did you feel that the descriptions of students, teachers, and experimental procedures were adequate? In what ways, if any, did the experimental conditions differ from those of regular school settings?

9. Sometimes education interventions have unanticipated effects, and occasionally these are negative. The experiment checked to see whether the problem-solving orientation of CGI might have led inadvertently to weaker student achievement on computational skills. Perhaps surprisingly the results were just the opposite. CGI students surpassed control students in performance on computational skills of addition and subtraction. Might CGI have caused teachers to spend more time on mathematics instruction of addition and subtraction than they had in the past and less time on other subjects and other mathematics? Presumably, the observation data would have allowed the experimenters to check whether CGI teachers spent more time on addition and subtraction than did control teachers, although these results were not reported. The experimenters investigated whether, within the amount of time allocated to instruction of addition and subtraction, CGI teachers spent more on problem solving than did control teachers, and that was found to be the case.

Don’t be discouraged when reading the following article. Few people understand exactly everything they read in a report of an experiment, not even seasoned veterans of education research. Don’t dwell on what seems beyond your present grasp. Rather, concentrate on any new understandings of comparative experiment methodology that may occur. Don’t be embarrassed if some parts of the article need more than one reading; some parts did for me. Finally, don’t be surprised if you spot what appear to be flaws in the study (beyond those few hinted at already). There is no such thing as a perfect experiment. A flaw in one experiment is simply a motivation for another slightly different experiment. Empirical studies are not meant to stand alone but rather to serve as pieces of the puzzle.
Using Knowledge of Children's Mathematics Thinking in Classroom Teaching: An Experimental Study

Thomas P. Carpenter and Elizabeth Fennema
University of Wisconsin—Madison

Penelope L. Peterson
Michigan State University

and

Chi-Pang Chiang and Megan Loef
University of Wisconsin-Madison

This study investigated teachers' use of knowledge from research on children's mathematical thinking and how their students' achievement is influenced as a result. Twenty first grade teachers, assigned randomly to an experimental treatment, participated in a month-long workshop in which they studied a research-based analysis of children's development of problem-solving skills in addition and subtraction. Other first grade teachers (n = 20) were assigned randomly to a control group. Although instructional practices were not prescribed, experimental teachers taught problem solving significantly more and number facts significantly less than did control teachers. Experimental teachers encouraged students to use a variety of problem-solving strategies, and they listened to processes their students used significantly more than did control teachers. Experimental teachers knew more about individual students' problem-solving processes.

Assisting in all phases of the research were Deborah Carey, Janice Gratch, and Cheryl Lubinski. Both the experimental treatment and the data collection were facilitated by Eileen Johnson and Peter Christiansen III of the Madison, Wisconsin, Metropolitan School District and Carolyn Stoner of the Watertown, Wisconsin, Unified School District. The research reported in this paper was supported in part by a grant from the National Science Foundation (MDR-8550236). The opinions expressed in this paper do not necessarily reflect the position, policy, or endorsement of the National Science Foundation. A previous version of this paper was presented at the annual meeting of the American Educational Research Association, New Orleans, April 1988. More detailed information about the instructional treatment can be obtained from the authors.
and they believed that instruction should build on students' existing knowledge more than did control teachers. Students in experimental classes exceeded students in control classes in number fact knowledge, problem solving, reported understanding, and reported confidence in their problem-solving abilities.

One of the critical problems facing educators and researchers is how to apply the rapidly expanding body of knowledge on children's learning and problem solving to classroom instruction. Implications for instruction do not follow immediately from research on thinking and problem solving, and explicit programs of research are needed to establish how the findings of descriptive research on children's thinking can be applied by teachers in actual classrooms with all the complexity that is involved (Romberg & Carpenter, 1986).

The purpose of this study was to investigate whether providing teachers access to explicit knowledge derived from research on children's thinking in a specific content domain would influence the teachers' instruction and their students' achievement. We hypothesized that knowledge about differences among problems, children's strategies for solving different problems, and how children's knowledge and skills evolve would affect directly how and what teachers did in classrooms. We also hypothesized that such knowledge would affect teachers' ability to assess their own students, which would be reflected in teachers' knowledge about their students. Knowledge about their students would allow teachers to better match instruction to students' knowledge and problem-solving abilities. As a consequence, students' meaningful learning and problem solving in mathematics would be facilitated.

Background

Research on children's addition and subtraction concepts. Many of the recent studies on children's thinking have focused on a specified content area, and the analysis of the task domain represents an important component of the research. This study draws on research on the development of addition and subtraction concepts and skills in young children. (For reviews of this research see Carpenter, 1985; Carpenter & Moser, 1983; or Riley, Greeno, & Heller, 1983). Within this domain researchers have provided a highly structured analysis of the development of addition and subtraction concepts and skills as reflected in children's solutions of different types of word problems. In spite of differences in details and emphasis, researchers in this area have reported remarkably consistent findings across a number of studies and have drawn similar conclusions about how children solve different problems. This research provides a basis for studying how children's thinking might be applied to instruction.

The research is based on a detailed analysis of the domain. Addition and subtraction word problems are partitioned into several basic classes, which distinguish between different types of action or relationships that represent different interpretations of addition and subtraction. Within each class three distinct problem types can be generated by systematically varying the unknowns. This scheme provides a highly principled analysis of problem type such that knowledge of a few general rules is sufficient to generate the complete range of problems.

The analysis is consistent with the way children think about and solve problems. Empirical research on children's solutions of addition and subtraction problems has shown that children initially solve word problems by directly representing the action or relationships in the problems. Research also has identified the major levels that children pass through in acquiring more advanced procedures for solving addition and subtraction problems. These more advanced procedures include counting strategies, use of derived facts, and recall of number facts. Thus, the taxonomy of problem types provides a framework for identifying the processes children are likely to use to solve different problems and to distinguish between problems in terms of their relative difficulty.

When children enter first grade, most of them are able to solve a variety of problems, and the processes they initially use to solve problems do not appear to have been learned through formal instruction. However, this informal knowledge may provide a basis to give meaning to the formal operations and symbol systems that are taught in school, provided that children are able to make the connections between their informal knowledge and the formal mathematics of school.

A cognitive view of the teacher. Consistent with the emerging cognitive view of the learner reflected in the research on addition and subtraction, those conducting research on teaching have begun to have a cognitive perspective on the teacher (Clark & Peterson, 1986; Peterson, 1988). Like their students, teachers are thinking individuals who approach the complex task of teaching in much the same way that problem solvers deal with other complex tasks. However, previous research on teachers' decision making suggests that teachers do not tend to base instructional decisions on their assessment of children's knowledge or misconceptions (Clark & Peterson, 1986). Putnam (1987) and Putnam and Leinhardt (1986) proposed that assessment of students' knowledge is not a primary goal of most teachers. They argued that keeping track of the knowledge of 25 students would create an overwhelming demand on the cognitive resources of the teacher. Putnam and Leinhardt hypothesized that teachers follow curriculum scripts in which they make only minor adjustments based on student feedback. The evidence is far from conclusive, however, that teachers do not or cannot monitor students' knowledge and use that information in instruction. Furthermore, Lampert (1986) has argued that a concern for
thinking. In this study we addressed the following questions about teachers and their students.

1. Did teachers who had participated in a program designed to help them understand children's thinking (a) employ different instructional processes in their classrooms than did teachers who had not participated in the program? (b) have different beliefs about teaching mathematics, about how students learn, and about the role of the teacher in facilitating that learning than did teachers who did not participate in the program? and (c) know more about their students' abilities than did teachers who did not participate in the program?

2. Did the students of teachers who participated in a program designed to help them understand children's thinking (a) have higher levels of achievement than did the students of teachers who did not participate in the program? (b) have higher levels of confidence in their ability in mathematics than did the students of teachers who did not participate in the program? and (c) have different beliefs about themselves and mathematics than did students of teachers who did not participate in the program?

**Method**

**Overview**

Forty first grade teachers participated in the study. Half (n = 20) were assigned randomly by school to the treatment group. These teachers participated in a 4-week summer workshop designed to familiarize them with the findings of research on the learning and development of addition and subtraction concepts in young children and to provide them with an opportunity to think about and plan instruction based on this knowledge. The other teachers (n = 20) served as a control group who participated in two 2-hour workshops focused on nontime problem solving. Throughout the following school year, all 40 teachers and their students were observed during mathematics instruction by trained observers using two coding systems developed especially for this study. Near the end of the instructional year, teachers' knowledge of their students was measured by asking each teacher to predict how individual students in his or her class would solve specific problems and if correct answers would be obtained. Teachers' predictions were then matched with students' actual responses to obtain a measure of teachers' knowledge of their students' thinking and performance. Teachers' beliefs were measured using a 48-item questionnaire designed to assess their assumptions about the learning and teaching of addition and subtraction. Students in the 40 teachers' classes completed a standardized mathematics achievement pretest in September and a series of posttests in April and May. The posttests included standardized tests of computation and problem solving as well as experimenter-constructed