Becoming an Educated Consumer of Research:

A Quick Look at the Basics of Research Methodologies and Design

Taylor Dimsdale
Mark Kutner

Meeting of the Minds
Practitioner-Researcher Symposium

December 2004

American Institutes for Research
www.air.org

Introduction

In recent years, the debate about the utilization of research studies that have been used to inform educational practice has been steadily growing. This debate is primarily the result of frustration; despite the investment of billions of dollars, learner outcomes have not significantly improved. One possible explanation is that most education programs and specific interventions are not informed by the same type of scientifically based research (SBR) used in medical and other scientific fields. There is a growing call for concentrating future research investments in the field of education on experimental SBR because other methods, including quasi-experimental and non-experimental studies, cannot identify effective practices and programs. Proponents of scientifically based research in education argue that education is a science, like chemistry, medicine, or psychology, and therefore is subject to the same rules of evidence as any other applied field.

Federally supported adult education and literacy research activities are increasingly emphasizing experimental studies. The landmark No Child Left Behind statute highlights the importance of experimental research that can identify educational programs that have demonstrated their effectiveness through scientifically based research.

Research in the adult education and literacy field is no exception. The Department of Education's Institute for Education Sciences (IES) has just commissioned an impact evaluation of direct literacy instruction for adult ESL students. The National Institute for Child Health and Human Development (NICHD) has commissioned a series of projects that use an experimental design to assess the impact of decoding and phonic instructional approaches for adults. The National Center for the Study of Adult Learning and Literacy (NCSALL) has been operating two "Lab Schools" in Portland, Oregon, and New Brunswick, New Jersey, to provide

> stable environments for conducting high quality research, facilitate close collaborations between researchers and practitioners, allow for systematic innovation, experimentation and evaluation of promising new instructional

> methods, materials and technologies, and create knowledge that improves our understanding of adult learning and literacy and improves practice (NCSALL, 2004).

Undoubtedly in the coming years, the amount of scientific research in education will continue to expand. The drive to produce, digest, and incorporate scientific research has already been taken up by endeavors such as the Department of Education's What Works Clearinghouse, the National Center for Education Evaluation, and the Coalition for Evidence-based Policy. Yet not all policymakers, researchers, and practitioners have bought into what is viewed as a stampede in support of experimental research projects to the exclusion of other quasi-experimental and non-experimental methods.

Those who are more cautious about, or downright hostile to, experimental research in education contend that too many contextual variables are involved in education issues for experiments to be useful; that, unlike molecules or chemicals, for example, human beings do not always act or respond in predictable patterns of behavior; and that humans come from such a wide variety of backgrounds that no two schools or programs would respond to the same treatment in the same way. Those who also value quasi-experimental and non-experimental methods posit that teachers learn best practices from years of trial and error in the classroom; they claim that data are an unacceptable and cold way to describe students and accuse researchers of trying to explain and dictate education policy far from the classroom by using numbers rather than personal experiences.

In an effort to address some of these issues, the National Research Council (NRC) commissioned a study to investigate the nature of scientifically based research in education. Its 2002 report, *Scientific Research in Education*, reached a wide range of conclusions, perhaps none more integral than this: "At its core, scientific research is the same in all fields" (Shavelson & Towne, 2002). Although the NRC does not discount the importance of quasi-experimental or non-experimental methods, including "professional" or "folk" wisdom in education, the report concludes that scientific research should be a vital part of the study of education in the future, in the same way it is considered vital to research in related fields. Although researchers in education most certainly face unique challenges that will necessitate new and uncommon approaches, the report contends that SBR can and should be part of the solution instead of being branded part of the problem.

Challenges Practitioners Face as Consumers of Research

So what is a practitioner to do? Practitioners face many challenges in interpreting research studies and their conclusions. How can you figure out whether and for what purposes you should be able to use research studies to inform practice? The proliferation of research is both encouraging and problematic because it cannot be assumed that a study's conclusions are valid simply because it is "published." The sheer volume of research produced is daunting for any consumer, as is the variety of sources in which it can be found. Whereas the majority of research used to be published almost exclusively in scholarly journals, research is now published by think tanks and partisan organizations, on websites, and by advocacy organizations. Each source is subject to different sets of standards and varying levels of peer review. Unfortunately, bad research does get published and is not always given a "red flag" by the research community (McEwan & McEwan, 2003).

In considering research, practitioners must be careful to understand how and for what purposes they can use the findings to inform practice. All research studies using an experimental and scientifically based research approach are not necessarily well designed and well implemented. Scientific jargon and explanations of advanced statistical procedures that accompany many research studies often add yet another obstacle to the process of understanding because practitioners do not always have training in advanced research methods to determine research quality. Also, practitioners cannot always easily determine the appropriateness of the specific data collection and analysis methods, a necessary step before interpreting the quality of research findings.

Becoming an Educated Consumer of Research

Despite these challenges, with the appropriate knowledge practitioners can rely on research studies as effective and important sources of information to inform practice. Well-designed studies using an experimental and scientifically based approach are extremely useful tools for teachers and administrators alike in improving educational practices. Other types of research projects that use quasi-experimental and non-experimental designs also can provide useful information and should not be dismissed simply because of the lack of a control group.

This paper is a primer from which practitioners can draw when they are faced with new and unfamiliar research. It is an attempt to flesh out the most important aspects of quality research and to explain how those not trained in advanced research methods can be effective consumers of education research. The paper describes the following:

- The continuum of research methods so that you as a practitioner will have some rudimentary understanding of the features of each of these methods. The amount of certainty that research findings are both valid and generalizable to other settings depends on the type of research method used. Because of the emphasis increasingly being placed on experimental and quasi-experimental designs, they receive the most attention in this paper. We do, however, also describe non-experimental designs at some length and do not discount them as useful tools in education research.

- Two aspects of research designs—the research question and research findings—that are important to consider when assessing the quality of research studies.

Research methods are reviewed first because the methodology that is used is a good indicator of how much trust consumers of research can have about the generalizability of findings, although a relationship clearly exists among the research questions being investigated, the research method used, and the type of findings that result. Appended to the paper is a glossary of research terms with which consumers of research need to be familiar.

Understanding the Continuum of Research Methods

Consumers of research need to do more than read the research question and then skip directly to the study findings and conclusions or implications. Although these elements of a research study can immediately inform the reader about what was investigated and whether there were significant findings, it is also important for consumers of research to be certain that the studies are methodologically appropriate for both the research questions guiding the study and the findings resulting from the

study. Interpreting research would be a much easier process if it could always be safely assumed that studies were methodologically sound.

The methods section of a research study should be used to judge whether the study findings can be supported by the study methodology. This section also gives the reader important insights into how to use the results and in what contexts the results may be applicable. A practitioner interested in using a piece of research should first understand the methodology used and its associated strengths and weaknesses. In this paper, we examine three main types of research designs:

- Experimental designs;
- Quasi-experimental designs;
- Non-experimental or descriptive designs.

In examining these methods, with special attention paid to experimental and quasi-experimental designs, we attempt to touch on four key issues:

- The type of questions that can be addressed;
- The features of the design;
- The types of data collection and analysis used;
- The implications for practice and policy.

Using these issues, we summarize in the following exhibit the three research methodologies, with more detail provided in the remainder of this paper.

Comparing Research Methodologies

Design	Questions Addressed	Principle Features	Data Collection and Analyses	Implications for Practice and Policy
Experimental	Causal questions Does a treatment work?	• Control groups • Random assignment • Sample selection • Non-contamination of participants • Heavily quantitative	• Standardized assessments • T-test	Definitive evidence for effectiveness or ineffectiveness of various educational interventions
Quasi-experimental	Does a treatment work? What is the relationship between program practices and features? Program outcomes	• Comparison of groups thought to be similar before intervention • Matching • Nonequivalent groups • Interrupted time series • Quantitative and qualitative	• Standardized assessments • Surveys • Interviews • Observations • Correlation • Regression • Multiple regression • Factor analysis • Chi-square	More robust understanding of program practices and features Can have greater external validity than true experiments in some cases
Nonexperimental	How does a treatment work? Descriptive rather than causal	• Naturalistic • Descriptive • Focus on meaning and explanation • Qualitative and quantitative	• Surveys • Case studies • Ethnographies • Correlation studies	New advanced statistical procedures have increased the quality of nonexperimental results. Will always be important in education because of the large part contextual variables play

Experimental Research

Over the past 5 years, the application of scientifically based research to the field of education has received considerable attention. In addition to the research requirements cited in the No Child Left Behind statute, the U.S. Department of Education's IES has made it a primary goal of the agency to "operate consistently with the standards of a science-based research agency…so that the research, evaluation, and statistical activities we fund lead to solving problems and answering questions of high relevance to education policy" (Whitehurst, 2002).

Experimental studies are highly quantitative studies that attempt to empirically test a hypothesis by using “hard” data and statistical techniques. This methodology is especially attractive in examining social issues, including education, because it seeks to produce definitive conclusions and measurable results. Results from experimental studies are primarily concerned with whether or not a treatment works. Experimental studies do not emphasize understanding how contextual factors and conditions influence outcomes.

Data and findings produced by experimental studies provide the most definitive conclusions possible and thus are very appealing to policymakers, researchers, and stakeholders. Experimental or “true” designs are the gold standard for research because these studies attempt to establish causal relationships between two or more factors. A well-designed experimental study allows the researchers to answer a research question with a high degree of certainty because their conclusions are backed up by concrete data.

Questions Addressed

The primary question an experimental study allows a researcher to answer is: Did the treatment or intervention have a significant effect on the treatment group? Or, put another way, does the treatment work? In the adult education and literacy field, for example, an experimental study can answer the question: Is phonics-based instruction an effective instructional approach for adult education and literacy learners?

Features of Experimental Research

Experimental studies can establish or disprove a causal relationship between two variables because random assignment has ensured that the groups being compared have the same socio-economic and education characteristics (and are thus equivalent) prior to the study. These are, however, complex and complicated studies that must be carefully designed and executed. Essential features of an experimental research design that practitioners should focus on when evaluating the quality of a study follow:

- Control group – All experimental studies are performed by comparing two groups: a treatment group that is exposed to some kind of intervention and a control group that does not receive the intervention. The control group provides

the groundwork for a comparison to be made and is perhaps the most essential feature of experimental research. It gives researchers an insight into how the treatment group would have theoretically behaved if group members had never received the treatment, and therefore it allows researchers to judge the effectiveness of the treatment.

- Sample selection – Identifying participants for the experimental and control groups is part of the sample selection process. In selecting samples it is essential for researchers to have a clear understanding of the characteristics of experimental and control group participants, such as age, race, economic status, and gender, among other things.

- Random Assignment – The use of comparison groups alone does not qualify a study as a true experiment. The participants in the study must also be randomly assigned to either the experimental group or the control group to ensure that there are no pre-existing differences between the groups. Randomly assigning participants to one of the two groups allows researchers to say with confidence that any differences shown between the groups after the intervention was a result of the intervention alone. If instead of being randomly assigned to the treatment and control groups, participants volunteer or are selected through a "convenience" sample, the experimental nature of the study is compromised.

- Noncontamination of participants – One of the greatest threats to the internal validity of an experimental study is contamination of participants. This can occur when the participants of the experimental and control groups communicate or when the instruction in the control group adopts key aspects of the treatment instruction.

- Specifying the treatment – Although every experimental research study includes a section on the treatment used, the treatment is not always explained in an unambiguous way. If a study is conducted to examine the relationship between test scores and an ESL instructional curriculum for adults, the report should clearly specify the contents of the curriculum, as well as the duration of each unit and method of instruction. Words such as "more" or "less" and "high" or "low" are not sufficiently descriptive. After reading about the treatment, the consumer of

research should understand exactly how the treatment group was treated differently from the control group.

- Reliability of the instrument used to assess learner outcomes – Because educational impact studies generally relate the effect of an instructional intervention on learner performance, standardized assessments are common tools. A successful experimental study should use an assessment instrument that directly measures the skills being taught by the intervention. Reliability refers to the consistency of the results of an assessment and is a necessary but not sufficient condition for validity. If an assessment used to measure reading comprehension is given to the same student three times without any instruction to the student, you would expect the student's test score to be about the same each time. If it was not, the instrument would not be providing reliable measures.

- Study outcomes – Beyond understanding the research question, consumers should also make sure they are aware of what the researchers expect to find following the intervention, as well as the instruments they plan to use. The study outcomes should be directly related to the research questions that guide the study. Experimental studies in education generally focus on learner outcomes as measured by standardized assessments.

Data Collection and Analysis

Experimental studies in education are used to determine the impact of an educational intervention or program. These impacts are typically learner outcomes: how learners' knowledge, skills, or performance have changed or improved as a result of the intervention. Standardized assessments are therefore the most commonly used data collection instrument in experimental studies (Marzano, Pickering, & Pollock, 2001). At a minimum, assessments must be administered to both experimental and control group participants prior to the beginning of the intervention (the pretest) and at the conclusion of the intervention (the posttest). Assessments may also be administered to both the experimental and control groups periodically during the intervention period, as well as at a predetermined period of time after the intervention has been completed if the study's objective is to measure long-term program impacts.

During experimental research studies it is also important to measure and document that the intervention is being delivered as designed. This is especially important when an experimental study involves multiple sites. Site visits by researchers to observe instruction and instructional logs completed by the instructors of both experimental and control groups are useful data collection instruments to measure the fidelity of the intervention and to ensure that instruction received by the control group participants is different from the instruction received by experimental group participants throughout the period of study.

Data analyses in experimental research design are usually very straightforward. Because the control and treatment groups are randomly assigned, researchers do not need to use advanced statistical techniques to account for pre-existing differences that may have compromised the effect size. The most common method of data analysis used in an experimental study is a test of significance, or a t-test. A t-test is a technique to compare the mean scores of the two groups and determine whether the difference between them is statistically significant.

Implications for Practice and Policy

Historically, the amount of experimental research done in the field of education has not been large because of the methodological difficulties and costs associated with designing and conducting such studies. However, as mentioned earlier, this method of research has been increasing and should continue to increase as a result of recent mandates from the Department of Education demanding increases in scientific or evidence-based research. The position of IES is that “randomized trials are the only sure method for determining the effectiveness of education programs and practices” (Whitehurst, 2003).

An increase in experimental research studies should produce an increase in definitive evidence for the effectiveness or ineffectiveness of various education treatments. A just commissioned national experimental impact study, sponsored by the IES, is currently under way to test the effectiveness of an explicit literacy curriculum in improving the reading, writing, and speaking skills of low-literate adult ESL learners. It is the first rigorous evaluation of an explicit literacy curriculum for ESL learners and will include a random assignment procedure at the beginning of the class instructional period.

Quasi-Experimental Research

Experimental studies are expensive and difficult to successfully conduct. They also focus on impact, without providing an understanding about other program implementation and contextual factors and conditions that affect learning. When it is not possible to use an experimental design, a quasi-experimental research design is the next best thing. A quasi-experiment attempts to replicate the conditions of a true experiment to the greatest extent possible and allows researchers to retain some control over the variables involved.

Researchers would employ a quasi-experimental design in two circumstances:

- Naturally existing groups are being studied – If researchers wanted to make a comparison between different age groups or genders, for example, they would need to use a quasi-experimental design. They would not randomly assign male participants to the female group in an experiment.

- Random assignment is restricted by external factors – This situation occurs frequently in education research. If a study is conducted to compare a treatment given to all students in each of two separate classes, then each classroom is a pre-existing group and therefore has not been randomly assigned.

The examples above demonstrate the shortcomings of quasi-experimental studies relative to true experiments. Because the comparison groups are formed before the study and without random assignment, pre-existing differences between the groups could contribute to differences between the groups in their response to the treatment.

Questions Addressed

Like studies using experimental research design, quasi-experimental studies also attempt to address a causal question or, in other words, to determine whether or not a treatment works. In some types of quasi-experimental studies, the difference with experimental studies is not in the questions being asked but rather in the certainty with which they can be answered. Other types of quasi-experimental studies can assess

whether the skills of learners receiving a specific educational treatment (e.g., phonics-based reading instruction) have improved, but they are unable to determine whether the improvement is based on the treatment or on other program practices or features. Quasi-experimental studies are also an appropriate methodology to use when the purpose of the study is to produce quantifiable data that help provide an understanding about how program features influence learner outcomes, but without comparing the effects to learner outcomes in programs without those features.

Features of Quasi-Experimental Research

Quasi-experimental research projects seek to compare two groups of individuals who are thought to have similar characteristics. There are many different types of quasi-experiments, and listing them all is beyond the scope of this paper. We do, however, touch on a few of the more common approaches, especially matching designs, nonequivalent groups, and interrupted time series.

Matching

A matched experiment is one of the most commonly found quasi-experimental designs in education. In a matched study, researchers attempt to replicate the advantages of randomization by comparing groups that can be shown at the outset to be similar in demographic and other characteristics, such as prior academic achievement. Statistical techniques are available for controlling for small differences, although these methods have their own limits. The main problem for consumers to consider when evaluating a matched study is that it is never possible to measure each and every characteristic between the two groups to ensure equivalence.

Consumers of research should also make sure that matched experiments include an adequate number of participants or schools and avoid comparing volunteers with nonvolunteers (Slavin, 2003). McEwan and McEwan (2003) describe five characteristics of a good matching experiment:

- It should describe the variables used to conduct the match;
- It should describe the procedures used to conduct the match;
- It should provide some minimal data to support the assertion that treatment and comparison groups are similar;

- It should provide some discussion of potential pitfalls in the matching process;
- It should not oversell or misrepresent its methods.

Nonequivalent Groups

This design is used to compare a treatment between groups that are likely to be different before the beginning of the study. Random assignment is not used, and participants are not matched to groups. Participants can often choose whether or not they receive the treatment. In some cases, the two groups are given a pretest at the beginning of the study and a posttest after the treatment; in other cases, they are tested only after the treatment. Giving both a pretest and a posttest allows researchers to identify any differences in performance between the two groups at the onset of the study.

Interrupted Time Series

Interrupted time series designs are used when nonequivalent treatment and control groups are given assessments multiple times over the course of a research study. Using this design can give more insight into the ongoing effect of a treatment or treatments of participants over time and can shed light on any trends that may be occurring. As with any nonequivalent groups design, however, these studies are subject to the same limitations that occur when a study lacks random assignment.

Data Collection and Analysis

Quasi-experimental studies typically use the same data collection methods as do experimental studies; therefore, the information about methods described in the previous section, including caveats about methods used, are also appropriate. Quasi-experimental studies are also likely to measure implementation issues and program operation to provide a contextual basis for analyses.

Analyses of quasi-experimental studies are not as straightforward as analyses in experimental projects. Advanced statistical procedures need to be used to account for the lack of randomization in the comparison groups. Research has shown that an understanding of eight statistical procedures can lead to an understanding of 90% of

quantitative research studies encountered (Valentin, 1997). Below are five such procedures associated with quasi-experimental research, along with basic definitions:

- Correlation – Used to find a relationship between two variables without making any claims about causation. If a study has used this procedure, the report will give a p-value, which represents the strength of the relationship between the two variables. The range of the p-value is from –1.0 to 1.0, with –1.0 representing a perfect negative correlation (both variables decrease) and 1.0 representing a perfect positive correlation (both variables increase).

- Regression – Whereas a correlation analysis tells you that two variables are related, a regression analysis is used when researchers assert that the relationship occurs one way or another and want to show how changes in the independent variable (e.g., program features) influence changes in the dependent variable (e.g., learner outcomes)—in other words, how much the one independent variable predicts the behavior of the other dependent variable.

- Multiple Regression – This procedure is an extension of single regression analysis. It is used when researchers want to test the cumulative effect of two or more variables (e.g., a program feature and staff development) on another variable (e.g., learner outcomes).

- Factor Analysis – If a study uses a large number of variables, researchers can employ a factor analysis to reduce the number of variables and to detect any possible existing relationships between or among them. Factor analysis is often used in survey research when the surveys ask a large number of questions.

- Chi Square Analysis – This method is used in studies examining the relationship between two categorical variables, which are any variables that are not quantitative (e.g., eye color or gender). It allows researchers to test whether the differences in behavior between two samples are large enough in the research study being performed that the differences can be generalized to the population at large.

Implications for Practice and Policy

Although not a true experiment, a quasi-experimental research design allows researchers to have some control over the treatment conditions. In fact, in some cases quasi-experiments may even demonstrate greater external validity than do true experiments. External validity refers to the extent that a study's findings can be generalized to other participants, settings, or times. When participants within a single school are randomly assigned to treatment or control groups, for example, there is a possibility that they may talk with each other about their experiences, thus jeopardizing the validity of the experiment. In a quasi-experiment, there is no random assignment and therefore no danger of this type of contamination.

Nonexperimental Research

Nonexperimental research designs have historically been the designs most often used in education. A defining feature of a nonexperiment is the absence of researcher control over the variables being studied (e.g., program factors), including the composition of the group of participants or the way the treatment is applied. Although researchers using an experimental design can randomly assign participants and have control over the independent variable, and quasi-experimental designs allow control over the independent variable but lack random assignment, a nonexperiment allows neither. A nonexperimental design lacks the validity of experimental or quasi-experimental designs but is often a great deal more practical to conduct. They have often been used to inform educational practice.

Questions Addressed

Whereas experimental and quasi-experimental studies are concerned with *if* something works, nonexperimental studies are conducted to address the question of *why* something occurs. The strength of nonexperimental research lies in researchers' ability to include large amounts of description.

Features of Nonexperimental Research

Nonexperimental studies give researchers more insight into *how* a treatment or intervention works, with researchers typically going into more detail and providing richer

descriptions about program practices and procedures. Nonexperimental research studies are more likely to include qualitative data that are more subjective, as well as descriptive (rather than causal) quantitative data about programs and procedures. The descriptive data available through nonexperimental research studies often have a greater appeal to consumers of research than do the results of experimental or quasi-experimental studies because anecdotes have a more emotional impact and are more easily relatable to personal experiences (Kutner et al., 1997).

In their book *Making Sense of Research*, McEwan and McEwan (2003) list three principal characteristics of qualitative research:

- Naturalistic
- Descriptive
- Focused on Meaning and Explanation

The qualitative research component of nonexperimental studies is often characterized as naturalistic because researchers insert themselves into an already existing situation and simply observe events that would be taking place even without their presence. Whereas experimental and quasi-experimental studies typically involve quantitative analyses in which researchers create a scenario that will ultimately allow them to measure one group against another, a qualitative study can measure results without any intervention. For example, if the students of a certain teacher consistently perform well above average on standardized tests, and much better than students of other teachers, researchers might observe that teacher's classes over a period of time in order to describe his or her methods.

Nonexperimental studies are potentially extremely valuable despite the lack of control and treatment groups. Description is one of the central strengths of qualitative research. Although quantitative researchers want as much simplicity in their experiment as possible, qualitative research thrives on detail. Finally, the focus of qualitative research is on meaning and explanation; researchers can focus on multiple questions and explore multiple explanations.

Data Collection and Analyses

- Surveys – Surveys, or questionnaires, are a frequently used data collection tool in all forms of research, but particularly so in nonexperimental and quasi-experimental studies. Surveys allow a great deal of flexibility. They can be administered in person, by mail, or over the phone. The questions can be close-ended or open-ended and can also use a Likert scale, which asks respondents to rate the strength of their opinions by using a numerical scale. Surveys may be cross-sectional, or administered at one point in time only, or they may be longitudinal, or administered on multiple occasions in a long-term research study. An interview is also considered a type of survey and has the added advantage of allowing researchers to probe a respondent for more detail than a paper survey could include. Interviews also help ensure that the respondent has a complete understanding of the question. Although surveys can give researchers a great deal of useful information and are normally fairly cost-effective, they do not allow the kind of causal claims that can be made in purely quantitative studies.

- Case studies – Another widely used method of qualitative research, particularly in the social sciences, is the case study. Case study research allows a very close, detailed study of a small number of subjects or situations and puts a premium on understanding the context of a situation. Interviews are often one source of information for case studies. Much like surveys, case studies do not allow causal claims, and there is also the danger that such close observation from a researcher can bias the study.

- Ethnography – Ethnographic research involves the close and prolonged research of a specific community, event, or phenomenon. It relies heavily on description and attention to detail.

- Correlation Studies – A correlation study is an attempt to provide evidence for a relationship between variables. However, unlike in experimental or quasi-experimental studies, researchers make no attempt to manipulate the variables being observed and can make no claims about causal effects.

Implications for Practice and Policy

Although nonexperimental studies lack a true experiment's ability to make causal statements, qualitative data can be coded into response categories with the results analyzed using statistical procedures. New and advanced statistical procedures such as covariance analysis increase the usefulness of quantitative and qualitative data from nonexperiments, and computer programs to analyze these data have been developed. Further, because experiments will never be able to account for all contextual effects in education, researchers and practitioners alike will always be dependent on nonexperimental studies to some degree (Grissmer & Flanagan, 2000).

Other Aspects of Research Design to Look for in Assessing Research Quality: Research Question and Research Findings

In much the same way that a poorly built foundation can doom a house before it is built, a poor research question can doom a research study before it even begins. The question might ask too much or too little, it could be too general or too specific, or it could simply be unverifiable and thus useless to consumers of research. A great deal of attention is normally paid to ensuring that the research methods of a study are sound, but the methods are meaningless if the question cannot be empirically tested or lacks significance (Shavelson & Towne, 2002). It is also important that findings presented are supported by the study's research questions and methodology.

The Research Question

A research question also informs and is firmly connected to the type of study design. Shavelson and Towne (2002) identify three interrelated types of research questions:

- Description – What is happening?
- Cause – Is there a systematic effect?
- Process – Why or how is it happening?

The first type of research question, description, and the third type, process, are questions that link to nonexperimental qualitative methodologies. The second question, cause, is asking specifically about causal relationships or effects and therefore is

connected with experimental and some quasi-experimental quantitative research methods.

When confronted with a research question, consumers of research should look closely at the following three criteria:

- Scope of the question;
- Level of specificity implicit in the question;
- Amount of previous and related research on which the question is based.

Scope of the Question

When researchers encounter a problematic issue in education and decide to conduct a study, one of the first decisions they must make is exactly how much ground should be covered and what exactly should be investigated. It would be naive to expect that a single research study could answer this question: Does providing teachers with more training improve student performance? Too many possible ways of approaching this study are left unchecked by the question. What type of training was provided? On what kind of performance were the students tested? How much more training did the teachers receive than they were normally given? No matter how the study was conducted, researchers would not be able to relate their conclusions back to the question. However, if the study tested the effects of a specific intervention on the performance of high school students in a math class, the study would then become less broad and thus more useful to practitioners.

Specificity of the Question

Along similar lines as the scope of the question is the level of specificity. It is important to note what specifically is being tested, particularly in the field of education where context is paramount. Results of a study depend heavily on the participants and their environment, and so a teacher interested in using a certain intervention with low-level learners may not be able to use results from a study of teaching practices in a GED class. The study may look at only a certain age or age range or may test an intervention only in a certain subject.

Previous Research

Practitioners should be aware of where the study falls on the continuum of previously conducted related research. If similar or identical studies have been performed, the research should build on existing knowledge, offer new conclusions, and invite further investigation on the subject. The more research that has been done on any given topic, the more credibility can be lent to the question.

Of course this is not always the case, and studies should not be discounted simply because they have little or no support from other research. If a study is the first of its kind, however, consumers should pay close attention to whether it has been subjected to peer review and in what form the review was performed. A study would be more credible if professionals from different organizations served as the reviewers than if it was simply reviewed by the president of the organizations publishing the research, for example. We further discuss peer review in the next section.

Research Findings

One of the most challenging tasks facing practitioners when they evaluate research studies has to do with assessing conflicting conclusions between various studies on the same topic. Studies that are replicated by multiple researchers often result in opposing conclusions, and the reasons behind such differences are not always obvious to an untrained eye. Studies on treatments in education often have a larger number of confounding variables and allow less researcher control than those in other fields. Improvements or declines in performance, for example, could be attributed to changes in factors such as teacher quality or student dropout rates. If there is little agreement about a treatment in the literature, most practitioners would understandably be hesitant to use an intervention.

To some extent, in evaluating the quality of the research question and methods of a study, research consumers are already evaluating the quality of the research findings. It is unlikely that researchers who use a well-designed question and well-designed methods would offer findings that somehow made the study not useful. However, it can happen, and consumers should keep several things in mind when judging whether research findings are valid:

- Do the results relate back to the question? This may seem like an obvious point to make, but it is also an important one. Researchers can and often do include a discussion of issues they encountered during the study that were unexpected and relatively unrelated to the research question, but they must also include findings that can be connected to the original issue being studied.

- Do the findings seem *too* conclusive? A good research question will never be answered "yes" or "no." In fact, most research questions result in more questions than they do answers. Even if a study's results appear to be conclusive, there should be an extensive discussion of possible limitations of the study. Researchers should explain any variables or situations that may not have been completely controlled and should offer these possibilities up freely to the public.

- Has the research been subject to peer review? Replication is the only way to confirm research findings. It is possible to get certain results in one isolated study that other researchers are unable to reproduce.

In Summary

With the implementation of the No Child Left Behind statute and particular attention increasingly being paid to scientifically based research in education, educators are already facing the results from new research studies, and more and more of the studies are based on experimental or quasi-experimental research designs. We hope that this paper will serve as a tool to assist consumers in understanding and using what may in many cases be unfamiliar research.

The research design does not in and of itself make a study scientific. When randomization is not possible, other methods can be used that, if done properly, will also yield valid results. Shavelson and Towne (2002) list five criteria demonstrated by a scientific study:

- Must allow direct, empirical investigation of an important question;
- Must account for the context in which the study is carried out;
- Must align with a conceptual framework;

- Must reflect careful and thorough reasoning;
- Must disclose results to encourage debate in the scientific community.

Studies with quasi-experimental or nonexperimental designs can satisfy each of these standards and should not be discounted simply because they were not randomized experiments. But it is important for consumers of research not to base instructional and programmatic design decisions on studies whose methodologies cannot provide sufficient rigor to support the findings presented. The key is for consumers of research to take a critical and objective look at research findings and determine whether the study methodology is appropriate, both in design and implementation, for the research questions being asked and the study findings being presented.

References

Grissmer, D. W., & Flanagan, A. (2000). Moving educational research toward scientific consensus. In D. W. Grissmer & J. M. Ross (Eds.), *Analytic issues in the assessment of student achievement.* Washington, DC: U.S. Department of Education, National Center for Education Statistics.

Kutner, M., Sherman, R., Tibbets, J., & Condelli, J. (1997). *Evaluating professional development: A framework for adult education.* Washington, DC: Pelavin Research Institute.

Marzano, R. J., Pickering, D. J., & Pollock, J. E. (2001). Classroom instruction that works: Research-based strategies for increasing student achievement. Alexandria, VA: *Association for Supervision and Curriculum Development*.

McEwan, E. K., & McEwan, P. J. (2003). *Making sense of research: What's good, what's not, and how to tell the difference.* Thousand Oaks, CA: Sage Publications.

Rittenhouse, T., Campbell, J., & Daltro, M. (2002). *Dressed-down research terms: A glossary for non-researchers*. Rockville, MD: Center for Mental Health Services.

Shavelson, R. J., & Towne, L. (Eds.). (2002). *Scientific research in education.* Washington, DC: National Academy Press.

Slavin, R. E. (2003). A reader's guide to scientifically-based research. *Educational Leadership, 60*(5), 12–16.

Valentin, T. (1997). Understanding quantitative research about adult literacy. Cambridge, MA: National Center for the Study of Adult Literacy and Learning. *Focus on Basics,* 1(A). [Online] Available: http://gseweb.harvard.edu/~ncsall/fob/1997/valen.htm.

Whitehurst, G. J. (2002, June). *Statement of Grover J. Whitehurst, Assistant Secretary for Research and Improvement, before the Senate Committee on Health, Education, Labor, and Pensions.* Washington, DC: U.S. Department of Education. Retrieved from http://www.ed.gov/news/speeches/2002/06/06252002.html.

Whitehurst, G. (2003, April). *The Institute of Education Sciences: New wine, new bottles.* Presentation at the annual meeting of the American Educational Research Association, Chicago, IL.

Appendix
Glossary of Terms

The items in our glossary are taken from the paper *Dressed Down Research Terms: A Glossary for Non-Researchers* written by Rittenhouse, Campbell, and Daltro in 2002 for a project funded by the Center for Mental Health Services. Although the glossary was primarily intended for use by consumers in the field of mental health research, many of the terms are common to research in the social and behavioral sciences in general. We have selected some of the terms that consumers would be most likely to come across when reading education research studies.

Analysis of Variance (ANOVA) – A statistical test showing the effects of an "independent variable" on a "dependent variable"; a technique to determine whether there are "statistically significant" differences of "means" between two or more groups.

Anecdotal evidence – What people say about something; not proven by hard (experimental) research.

Applied research – A kind of study that tries to make sense of the real world and to change what people actually do in the real world.

Assessment – A test or other way of measuring something, such as a person's mental health or goals or needs; often the first test in a series of tests, or a test given before treatment starts.

Attrition – The "drop-out" rate among people who are being studied. People may quit because they want to, or they may not be able to stay in the study group (because of illness, lack of time, moving to another city, etc.), or they may not fit into the study anymore (if they get a job or marry, for example, in a study about single people who are not working).

Benchmark – A standard, test, or point of reference (often a number).

Bias – Something that may lead a researcher to wrong conclusions; for example, mistakes or problems in how the study is planned, or how the information is gathered or looked at. If two different interviewers had different styles that caused people with the same thoughts to give different answers, but the answers were all put together in one pool, there would be a bias. It is impossible to conduct completely bias-free research.

Bivariate analysis – The study of two things (amounts, values, "variables") and how they are connected.

Case study method – The close study of one person, group, process, event, etc. (most often one person). The one chosen (for example, Lauren Slater, who takes Prozac) is seen as like others in a larger group (for example, the larger group of all people taking Prozac) who are not being studied.

Categorical Variable – A piece of information that can be put in a single category, instead of being given a number: for example, the information about whether a person owns a car or about whether the person belongs to a certain race can be put in the category of “yes” or the category of “no.”

Causality – The link between causes and their effects. For example, smoking (the cause) leads to lung cancer (the effect), and studying how often this happens and why would be studying causality. In most research about how people behave, causality can’t be proven, and ideas are tested by whether things (“variables,” amounts) change together.

Chi-square – A statistical test that measures “significance” in the study of “frequency distributions.”

Close-ended questions – Questions that list the possible answers; for example, “multiple-choice” questions or “true-false” questions.

Coding – Putting answers into groups (usually numbered groups), so the answers can be counted and studied more easily.

Cohort Analysis – A study of a group of people who stay in that group over a long time. For example, all people born in 1960 are a cohort; or all students who will graduate from high school in 1999. The study follows this group over time, rather than looking at them once.

Confidence Interval – A number (range) that shows how likely it is that the true amount is inside the listed range of amounts; for example, a 95% confidence interval of 25-45 would mean there is a 95% chance that the right amount (number, score, measurement) is somewhere between 25 and 45.

Confounding Factors – The inability to tell between the separate impacts of two or more factors on a single outcome. For example, one may find it difficult to tell between the separate impacts of genetics and environmental factors on depression.

Construct Validity – The measure of how well the test fits the ideas behind the study and the way the topic has been set out. Usually such a test separates 2 groups that are known to be opposite extremes.

Continuous Variable – Something that has an unlimited number of possible values; for example, height, weight, and age are all continuous because a person’s height, weight, or age could be measured in smaller and smaller fractions between the numbers of the whole inches, pounds, or years.

Control Group – The people being studied who are not getting the treatment or other “intervention”/change that the people in the “experimental” group are getting; for example, in a study testing a medication, the control group would not take the medication.

Correlation - A measure ranging from 0.00 to 1.00, of how well two or more things (“variables”, values, scores, etc.) change together. Both things may get higher at the same time, or lower at the same time, or one may get higher while the other gets lower. For example, saving money and spending money are correlated (inversely), because the more money you save, the less you spend.

Cross-cultural Method (comparative method) – A way of studying different cultural groups (for example, Eskimos and Mennonites) to see how they are the same and how they are different.

Cross-sectional Study – Research that compares people at one time only. Cause and effect can't be seen in this kind of study.

Data - Information taken from the study records, questionnaires, interviews, etc.

Databases - Groups of information recorded in a standardized (set, official) way.

Data Collection - The gathering of information through surveys, tests, interviews, experiments, library records, etc.

Data Processing - Recording, storing, calling up, and analyzing information with a computer program.

Degrees of Freedom (df) - The number of values/amounts that are free to vary in one calculation. Degrees of freedom are used in the formulas that test hypotheses statistically.

Demography - The study of a group of people, including its size, how old different members are, what sex and race different members belong to, how many people are married, how many years they went to school, etc.

Dependent Variable - The "effect" that depends on changes in the "cause" (or "independent variable"). In an experiment, the dependent variable is the one the researcher measures. For example, better sleep might be dependent and a change in medication would be independent.

Descriptive Statistics - A way of sharing information by putting numbers into words so the information is easier to understand.

Descriptive Study - Research that finds out how often and where something (like a race or an age or a behavior) shows up; this kind of study doesn't look at "cause" and "effect," and is not "experimental."

Direct Observation - The study of things you have actually seen, rather than things you have heard about or read about.

Discrete Variables - Separate values or groupings, with no possible values (numbers, measurements) between them. The only choices are separate categories; for example, "male" and "female" are discrete variables. These are also called "nominal variables."

Distribution - The measure of how often something (for example, an age or a hair color) is found in the group being studied; or the range of those measures.

Effectiveness Study - A measure of change after treatment; not an "experimental" study having a "control group."

Ethnography - A kind of study that looks at and describes a society's culture.

Evaluation Research - A study to see whether a program or a project is doing what it set out to do.

Experimental Design - A kind of study that controls the circumstances of the research and measures the results exactly.

Experimental Group - The people who receive the treatment being studied. This group is compared with the “control group,” in which people are as much like the experimental group as possible, except that the control group does not receive the treatment.

External Validity - A measure of how well the results of a study apply to other people in other places.

Face Validity - A measure of whether a study’s results seem to make sense and whether they are clear.

Factor Analysis - A type of study used to find the underlying causes and characteristics of something. The general purpose of this test is to take the information in a large number of “variables” and to link it with a smaller number of “factors” or causes.

Field Research - A kind of study that looks at people in their everyday world, not in a laboratory or other special setting. The everyday world is the “field.” This research is usually not “experimental.”

Focus Group - A group of people who have shared an experience (for example, who have all taken the same medication or who have all been sexually harassed) and who are asked about that experience.

Frequency Distribution - A scale, drawing, or graph that shows how often something (a number, answer, percentage, score) is found in the total pool of information being studied.

Independent Variable - Something that causes change in something else (the “dependent variable”). The independent variable is the one changed by the researcher to see what will happen to the dependent variable(s).

Indicator - A characteristic something has that lets you tell that thing apart from something else. For example, pregnancy is an indicator that a person is female, but having long hair is not.

Inferential Statistics - A method that allows researchers to make judgments about a whole “population” by using examples from a smaller part (a “sample”) of that population.

Internal Validity - A measure of how well a study accounts for and controls all the other differences (that are not related to the study question) among the people being studied. An internally valid study usually requires a “control group” and “random assignment.” In an experiment, this kind of validity means the degree to which changes that are seen in a “dependent variable” can be linked to changes in the “independent variable.”

Interval Scale - A scale with points that are equally distant from each other, but without an absolute zero point; for example, the Celsius temperature scale.

Intervention - A planned change; for example, a new therapy or a new medication; or the act of making this change.

Institutional Review Board (IRB) - The group who looks at the ethical standards of all research that involves studying people.

Likert Scale – A scale to show how a person feels about something; it usually includes a range of possible answers, from “strongly agree” to “strongly disagree,” which each have a number. The total score is found by adding all these numbers.

Longitudinal Research Design - A study lasting a long time (usually years), because the researcher is seeing how time affects the main question of the research.

Matching - Choosing a “control group” (the group who doesn’t receive the treatment or other thing being tested) who is like the “experimental group” (who does receive the treatment); the groups would be alike in gender, age, race, and severity of disability, for example.

Mean (arithmetic) - The average of a group of values (numbers, scores); the number you would get if you added the score of each person, for example, and then divided that by the total number of people.

Measure - A test; or how an amount or a thing is shown.

Median - The exact middle; the point which divides a set of values (numbers, scores, amounts) so that exactly half the values are higher than the point and exactly half are lower.

Mode - The most frequent value (number, score, amount) in a group of values. For example, the mode in the group of “3, 5, 3, 100” is “3.”

Multivariate Analysis - The study of two or more effects (“dependent variables”) at one time.

Non-parametric Statistical Procedures - Tests that don’t need to make strong assumptions about characteristics of the people who take the tests.

Non-response Bias - A research fault based on the people who didn’t agree to be studied, although they were chosen. People who didn’t agree may have been different in other important ways from people who did, and so the study’s results might be true for only part of the chosen group. For example, if the chosen group is depressed people and the more depressed ones were too tired or hopeless to answer a survey, then any answers about the amount of energy or hope in depression would not be giving a full picture.

Open-ended questions - Questions which let people answer in their own words instead of having to choose from set answers like “a” or “b” or “true” or “false.”

Outcome - The way something, often a treatment or a program or a study, turns out; the effect it has on people; or the record or measure of the effects.

Outcome Measure - The measure of a change (usually the difference in scores before and after treatment).

Parameter - Something that sets a group of people apart from other groups.

Parametric Statistical Procedures - Ways to study information that is taken from a group of people who fit a “normal” range like the bell curve.

Pilot Study - A small first study using the same methods that a researcher wants to use for a larger study; the pilot study shows how well those methods work.

Population - The total number, usually of people, in the group being studied. In some studies, the population may be organizations, records, or events instead of people.

Predictive Validity - A measure of how well a test can predict something.

Pre-post Testing - Giving the same test before treatment and just after treatment.

Pretest - A test given to a small group of people to see how well the test works before giving it to more people.

Probability - A measure of how likely something is. For example, probability could be written as "$p<.05$," which means that based on chance alone this thing should happen fewer than 5 times in 100.

Probability Sampling - Also known as "random sampling." Choosing people to be studied, in such a way that each person (or thing, place, etc.) in the total pool has an equal chance of being chosen.

Psychometrics - Psychological tests that are standardized (formal, set); for example, an IQ test.

Qualitative Studies - Research using what people say or write in words, rather than numbers or people's numbered answers; for example, studies based on short answers or personal histories.

Quantitative Studies - Studies of information that people give in numbers or in a way that can be numbered.

Quasi-experimental Design - A study that seems like an "experimental study" and is designed to be almost as powerful a test as if it were experimental, but the people studied are not put into their groups randomly and there is no "comparison" or "control group."

Random Assignment - The process of putting study participants into groups ("experimental" or "control") purely by chance.

Random Sample - A group of people (or animals or things) chosen from a larger group by chance. Sometimes this kind of sampling is done with a table of random numbers, or with a computer giving out random numbers, or by drawing lots.

Range - All the values (amounts, numbers, scores) from lowest to highest; the distance the whole group covers.

Regression Analysis - A way of predicting one value/amount (the "effect" or "dependent variable") from other values/amounts (the "causes" or "independent variables"); predicting the effect by what the cause looks like.

Reliability - A measure of whether the answers or results will be the same if the test or experiment is repeated.

Replication - Repeating a study to check the results; or a study that repeats an earlier one.

Research Design - A plan for gathering and studying information.

Respondent - A person who is being interviewed or studied or who answers a questionnaire.

Response Rate - A number showing how many questionnaires were filled out, usually written as a percentage (of the total questionnaires sent or given).

Sample - A part of a larger group of people. The sample may or may not be chosen by chance. This term can also be a verb, meaning to choose this smaller group.

Scaling - Giving numbers, in order, to information which was in words or ideas; for example, showing a person's opinion by a number from this list: 1) strongly agree; 2) agree; 3) disagree; 4) strongly disagree. Scaling always uses numbers.

Scale - A test; a group of linked questions that can be added together to form a measure of one thing.

Self-selection - A way of choosing the people for a study by letting them set themselves apart from a larger group in some way; for example, by responding to a questionnaire or by going to a program.

Special Populations - Groups of people that can't be studied in the same way and by the same rules as other groups, for some reason.

Stakeholders - People who have a share or an interest in something; for example, people who receive some of the profits of a hospital because they have helped to set up the hospital or have given money to it in the past. Stakeholders can be clients, relatives, professionals, community leaders, agency administrators, volunteers, etc.

Standard Deviation - A measure of how widely the values (amounts, numbers, scores) in a group of values are spread around the "mean" (midpoint). For example, all the scores may be very close to the midpoint, or many of them may be much higher or lower.

Significance - A mathematical test of whether a study's results could be caused by chance or whether they really show what they seem to show.

Survey Research - A type of study that uses phone questions, mailed questions, interviews, or self-completed forms, and that does not use the "experimental method."

T- Test - A statistical test of the difference between two "means."

Trend - A steady change in one direction over time; for example, more and more parents letting their children have later and later bedtimes over several years would be a trend.

Time Series Design - A way of studying what researchers have noticed at set times; for example, studying how many cavities a group of children have every 6 months.

Univariate Analysis - The study of only one thing that might change, not a group of different things that might each change.

Validity - The measure of how well a scale or test shows what it's supposed to show. There are several different types of validity and each type must be tested separately:

Construct Validity - The measure of how well the test fits the ideas behind the study and the way the topic has been set out. Usually such a test separates two groups that are known to be opposite extremes.

Content Validity – The measure of how fully the whole topic of the study is covered by the test. For a test to have content validity, every piece of the topic should also be part of the test. This is sometimes called "Face validity."

Criterion Validity - The measure of how well the test matches an accepted test ("gold standard") outside the study. There are two types of criterion validity:

(a) Concurrent validity – The measure of how well the test being studied and the "gold standard" test measure the same thing at the same time.

(b) Predictive validity - The measure of how well the test being studied predicts some practical result that the "gold standard" will find later.

Variable - Anything that can have different values (be different sizes or amounts) at different times; what is being measured in a study.

Variance - The measure of how wide the range of values (amounts, sizes, or scores written in numbers) is; of how far apart these numbers are. It is a number found by multiplying the "standard deviation" by itself.