Running head: EFFICACY OF INTERVENTIONS IN EDUCATIONAL RESEARCH

Does it Work? A Guide to Investigating the Efficacy of Interventions in Educational Research

Abstract

This paper examines types of research questions posed by educational researchers and identifies intervention research as a type of causal question. Next, research designs for answering causal questions are reviewed, paying attention to the application of lesser used designs that may overcome limitations faced when randomized experimental designs are not feasible or appropriate. The role of the designs is discussed as is the role of non-causal research in education. Finally, a graphical organizer to aid in interpreting existing research and planning future research on educational interventions is presented. A design for a sample study to test the effects of a new math program that may be used as a model for participants who may be considering planning their own research is also presented.

Does it work? A guide to investigating the efficacy of interventions in educational research

In a recent article on Newsweek.com (2008), Peg Tyre commented on the latest study showing that more boys—a lot more boys, nearly double—are being referred to medical professionals for emotional and behavior problems than are girls. That statement is a fact. In the next several paragraphs, Tyre discusses possible explanations for such a disparity. These explanations take the form of various theories about why boys are struggling more than girls. Some attribute the cause to an increase in ADHD in boys or the burgeoning levels of contaminants in our environment. Tyre offers her own theory—that kids are overscheduled and asked to behave like adults at too young of an age; moreover, their unstructured free play time has diminished. She supports her theory with anecdotal evidence, but Tyre is not alone in her thinking. The recent plethora of books with titles such as *The Hurried Child* (Elkind, 2001) and *The Case Against Standardized Testing: Raising the Scores, Ruining the Schools* (Kohn, 2000) show a backlash against schools' increasing reliance on standardized testing as well as the No Child Left Behind Act (U.S. Department of Education, n.d.). But is her theory correct? Tyre makes plausible arguments, but there are other arguments—and, more importantly, conflicting data—that cast doubt on her lay theory. Consider, for instance, the many benefits of Head Start (U.S. Department of Health and Human Services, 2001) and the Abcedarian Project (National Institute of Early Education Research, 2008) for low income children. These kids do not seem to be harmed by earlier exposure to academic rigor. Consider also children in colonial times (Washington Crossing Historic Park, n.d.) or living in indigenous tribes (Sprott, 2002) who were and are expected to perform significant household chores with little playtime at very young ages. There is no evidence that these children suffered negative outcomes from such labor.

So, who do we believe? Are all theories equally valid? Of course not. The scientific method provides a way to evaluate claims of competing theories so we can determine which are better supported than others. We may never "prove" a theory correct, but we can "disprove" enough claims of the theory to render the theory invalid. As educators and education researchers, we share the common goal of trying to find solutions that work to increase students' success in school, and ultimately, in life. Often, we may ask ourselves if what we are doing works. We may obtain anecdotal evidence that it does, but what if it only works in one classroom and not others? Perhaps the activity we love works for reasons unbeknownst to us. It would be helpful to be able to identify the aspects of the activity that are most beneficial for improving students' learning in order to most efficiently meet students' needs. Similarly, how do we know that what is being taught during professional development teacher work days actually works? It would save much time and resources if teachers knew that what they were being told to implement in their classrooms really and truly works.

When one of the authors of this paper was a teacher, she dreamed of having access to a large binder filled with research-supported activities that made sense and promoted student learning. What actually happened, however, was that she was required to attend multiple workshops and inservice trainings in which each presenter claimed that his/her particular method of instruction was really the best way to teach. What was most disconcerting was that the theories underlying the different methods often conflicted with each other, resulting in superficial changed to the curriculum rather than well thought out plans for how to design comprehensive instruction based on these techniques.

As stated in the National Research Council's 2002 report on Scientific Research in Education (Shavelson & Towne, 2002, pp. 1, 12), "No one would think of getting to the Moon

or of wiping out a disease without research. Likewise, one cannot expect reform efforts in education to have significant effects without research-based knowledge to guide them… [T]o address the challenges of, for example, low-performing schools, the 'achievement gap,' and language diversity, educators today require new knowledge to reengineer schools in effective ways. To meet these new demands, rigorous, sustained, scientific research in education is needed." To meet this need and to also address the poor quality of much existing educational research, there must be an increased dissemination of information on and guidelines for effectively conducting research on educational interventions. New guidelines exist (e.g., see Center for Psychology in Schools and Education, 2008), but they are still seldom implemented.

With the recent emphasis in our public schools on accountability, it becomes even more important to make sure that we are investing in curriculum and teaching methods that really work to increase students' academic success. It is our hope in this paper to provide teachers, teacher educators, and educational researchers with the tools to be able to evaluate current research on interventions—in our case, activities that are supposed to work to improve schooling—as well as the tools to be able to conduct high quality research on their own.

Unfortunately, for many years, educational research on interventions was of poor quality and not generalizable to schools because of either its poor design or lack of ecological validity (Shadish, Cook, & Campbell, 2002). In 2002, the IES was created to help strengthen the quality of educational research, particularly research on educational interventions—those specific activities, programs, curriculum changes, or textbooks that purport to increase student achievement and success in school:

> The Education Sciences Reform Act of 2002 established a new organization within the U.S. Department of Education, the Institute of Education Sciences. Our mission is to provide rigorous evidence on which to ground education practice and policy. By identifying what works, what doesn't, and why, we intend to improve the outcomes of education for all students, particularly those at risk of failure. (Whitehurst, n.d.)

Since the formation of the IES and the push for quality educational research, guidelines for conducting such research have been clarified and updated. For example, the recent revision of Cook and Campbell's classic text on quasi-experimentation (1979) by Shadish, Cook, and Campbell (2002) presents an extensive discussion about not only how to conduct high quality educational intervention research, but also the conditions under which different types of research designs might be used. In addition, the Center for Psychology in Schools and Education, an office within the American Psychological Association, just published a useful chart to help educational researchers test intervention effects using multiple methods (2008). Still, such guidelines remain out of reach of many educators and educational researchers. In this paper, we will discuss the types of research questions posed by educational researchers, review research designs created to answer causal questions, and then present a useful graphical organizer for helping educational researchers decide which study design to use. We will conclude with a design for a sample study to test the effects of a new math program using the guidelines just presented.

Types of Research Questions Concerning Education

In this section, we discuss examples of the types of research questions posed by educational researchers and identify intervention research as a particular type of causal question. This list is not comprehensive, but it does address the majority of questions asked by those interested in educational research.

Factual Questions

These questions concern descriptive information about a topic, such as what are the facts about X, or how does A relate to B. An example question might be, "On average, how do U.S. students compare in their math achievement compared to those in other industrialized nations?" Another is "Are SES and achievement related?" These types of questions are often investigated using straightforward quantitative designs, such as survey questionnaires and correlational statistics.

Investigative Questions

Investigative questions concern why things are the way they are, such as why does a particular phenomenon happen. A typical example is the question, "Why do students with low SES tend to do more poorly in school than their more affluent counterparts?" This type of question may be answered using grounded theory (Glaser & Strauss, 1967) or using a predetermined theoretical framework, with the careful collection of evidence to support or disprove key hypotheses of the theory involved. Attachment theory is a good example. Ainsworth (Crain, 2005) started with the idea that some children seemed to be happier, more confident, and more connected to their parents. She investigated this phenomenon at first with ethnographic research techniques where she lived among a tribe in Uganda and studied their

parenting techniques. She developed the idea of secure attachment, then after moving to Baltimore, she tested her theory by designing a Strange Situation for children to see if how they actually reacted upon being separated from their mothers confirmed her hypothesis of how they would react.

Explanatory Questions

These questions concern *how* something works, such as, how does this method or this type of schooling work. An example question is "How does Montessori education benefit students with mental retardation?" This type of question is usually answered with qualitative research that investigates a phenomenon over a period of time, gathering careful descriptive evidence from classroom observations, interviews, etc. and is particularly suited for grounded theory research (Glaser & Strauss).

Intervention Questions

These questions concern whether one treatment works better than other, along the lines of whether X (treatment) is helpful to students. Similarly, we can ask if one treatment is better than another, such as, is X better than Y, or which is better under what conditions. A sample question might be: "Is whole language instruction better than phonics instruction?" These are the types of questions we are addressing in our paper, and they are best answered using experimental methodology, particularly the gold standard in educational research of the Randomized Control Trial (RCT). Due to the limitations and difficulties of executing this design, particularly in school settings, the quasi-experimental design was created to also help answer questions of a causal nature. Used properly, it can be a very good way to test the efficacy of particular school interventions.

In the next section, we review research designs for answering causal questions that are concerned with the efficacy of instructional interventions, paying particular attention to some of the newer designs that may overcome some of the limitations encountered when RCTs cannot be conducted.

Research Designs for Answering Causal Questions in Intervention Research

New methods, treatments, interventions, curriculum, and other strategies are often implemented in classes or schools with the intention of making some positive impact, such as increasing student performance, learning, or skills or decreasing behavior problems. As these new elements are introduced in a classroom or implemented across grades or school-wide, the question of "how effective is X on Y" and more specifically "does X cause Y" often arises. These are valid and important questions to answer. For example, does the implementation of a new teaching method increase student performance in mathematics more than the previous method?

It is important to define what causality is in the context of intervention research. A causal relationship is such that the following occurs:

1) the cause (e.g., new intervention or curriculum) occurs prior to the effect (e.g., student performance in mathematics);

2) there is a relationship between the cause and the effect; and

3) the cause is the only plausible explanation for the effect (Shadish et al., 2002).

There are a number of research designs that allow researchers to study causal relationships, and the most commonly known is an experiment. Additional but less frequently used designs exist,

however, that are valuable in educational research when true manipulation of the treatment (as in a true experiment) is not possible. The following research designs will be reviewed: a) randomized experiment; b) quasi-experiment; c) regression discontinuity; d) propensity score analysis; e) correlational; and f) action research and design experiments.

Randomized Experiments for Intervention Research

What is referred to as an experiment has been more specifically labeled a 'social experiment' (p. 546) by Cook and Shadish (1994) or a 'field experiment' by Kerlinger & Lee (2000, p. 581). A social experiment (shortened to just 'experiment' in this manuscript) occurs outside of a controlled environment (e.g., in a 'real' classroom) which therefore results in less standardization and more enduring treatments as compared to experiments conducted in a controlled laboratory (T. D. Cook & Shadish, 1994). In a randomized experiment, the treatment (e.g., new teaching method or curriculum) is assigned to students (i.e., the participants in the research study) on the basis of chance such as through a coin toss or random number generation (Shadish et al., 2002). When random assignment is performed correctly, the groups created are similar, on average (Shadish et al., 2002). Random experiments, more specifically termed randomized control trials (RCT) (Shadish) or more loosely "true experiment" (Wallen & Frankael, 2001, p. 279), are often known as the "gold standard" (Shadish et al., 2002, p. 13) in educational research. This is because when designed rigorously and systematically, random assignment creates groups that are similar, on average, and any differences between groups can be attributed to the intervention rather than other factors (Shadish et al., 2002; Torgerson & Torgerson, 2008). In addition, when a randomized controlled trial is designed and conducted systematically and rigorously, relatively simple statistical procedures are all that is required to analyze the data (Torgerson & Torgerson, 2008).

In many randomized controlled trials conducted to examine intervention research, it is not possible to randomly assign at the student level but it is possible to randomly assign at a cluster level (e.g., randomly assigning classes rather than individual students to different teaching methods). If the interest is on studying the student, however, randomly assigning at the cluster level introduces analytical issues such as potential homogeneity that exists within the clusters. Although beyond the scope of this paper, statistical procedures such as multilevel modeling are commonly used to address issues related to cluster random assignment (Raudenbush & Bryk, 2002; Shadish et al., 2002).

There are a number of variations of randomized experiments that can be applied in intervention research. The most basic design includes two conditions (e.g., one treatment and one control group) in which students are randomly assigned to one of two groups. The students in the treatment group receive the intervention or treatment (e.g., new teaching method or curriculum). How the control group is defined is left to the researcher's choosing. The control group may be, for example, a group that receives no treatment at all or a number of variations in which the control group receives the treatment at some point or receives a comparison treatment (e.g., the method of instruction that has always been offered or the curriculum that has been used previously) (Shadish et al., 2002). Regardless, both the treatment students and the control students are measured after the intervention. A classic example of this basic design is research conducted on the Salk polio vaccine in which 400,000 children were randomly assigned to receive a placebo or a polio vaccine (Meier, 1972). Other commonly used experimental designs include: 1) pretest-posttest control group; 2) alternative treatments design with pretest; 3) multiple treatments and controls with pretest; 4) factorial; 5) longitudinal; and 6) crossover (Shadish et al., 2002). Discussing each of these designs is beyond the scope of the present

paper. Interested readers are encouraged to review any of a number of excellent sources (e.g., Frankael & Wallen, 2005; Kerlinger & Lee, 2000; Shadish et al., 2002).

Quasi-Experiments in Intervention Research

Although RCTs may be the gold standard in educational research because they most clearly allow the inference of causality, there are many instances in which RCTs are not ethical nor feasible (Rutter, 2007). A quasi-experiment does not include random assignments of students to intervention (Shadish et al., 2002; Wallen & Frankael, 2001). This does not mean that inferences of causality cannot be made from quasi-experiments; however, causality is more difficult and more attention must be paid in designing the study in order to reduce the likelihood of other explanations for what has occurred (beyond that of the intervention) (Shadish et al., 2002).

In a quasi-experiment, the following conditions must still occur: 1) the cause occurs prior to the effect; 2) there is a relationship between the cause and the effect; and 3) the cause is the only plausible explanation for the effect (Shadish et al., 2002). The last condition is the point at which random experiments and quasi-experiments differ. In a quasi-experiment, the last condition is not met through random assignment but can be met by the following. First, potential threats to internal validity (i.e., the probability that something other than the intervention caused the outcome) are identified and examined to determine the likelihood that they may explain the outcome rather than the intervention. Second, design (e.g., additional pretest measurements and more control groups) and statistical (i.e., using statistical procedures to remove confounding of variables) controls are introduced. Third, "coherent pattern matching" (Shadish et al., 2002, p.

105) is introduced in which "a complex prediction is made about a given causal hypothesis that few alternative explanations can match".

There are a number of different types of quasi-experiments including designs that: 1) do not have control groups (e.g., one group pretest-posttest, repeated treatment); 2) have a control group but do not have a pretest (e.g., posttest-only with nonequivalent groups); and 3) have both a control group and a pretest (e.g., untreated control group with dependent pretest and posttest samples, untreated control group with dependent pretest and posttest samples using a double pretest, cohort control group with pretest from each cohort). Interested readers are encouraged to review any of a number of excellent sources for specific details on designing quasi-experimental studies (e.g., Frankael & Wallen, 2005; Kerlinger & Lee, 2000; Shadish et al., 2002). Two valuable but lesser used types of quasi-experiments—regression discontinuity and propensity score analysis— will be discussed next. These designs were selected because they have great applicability in educational research and, of all the quasi-experimental designs, are the closest kin to a randomized experiment.

Regression Discontinuity

Although regression discontinuity was introduced in the late 1950s (Campbell, 1984), with the exception of application to evaluate Title I programs in the mid-1960s (Trochim, 1980), it has been used sparingly in educational research. However, there are many instances in which regression discontinuity is a better option than a quasi-experiment or may increase the power of the test when combined with a randomized experiment (Shadish et al., 2002). Additionally, regression discontinuity is the only natural experiment that can deal with unobserved confounding variables (Rutter, 2007). The regression discontinuity design for intervention

research is really very simple: Students are assigned to a treatment condition based on a cutoff score of an assignment variable (which is at least ordinal in measurement scale and measured prior to intervention) rather than randomization or other assignment (Shadish et al., 2002). This assignment, by cutoff and only cutoff, is a strict rule that must be adhered for regression discontinuity to function properly. Participants on one side of the cutoff score receive the treatment, and participants on the other side of the cutoff score do not receive the treatment (Shadish et al., 2002). A treatment effect is found if the scores of the two groups differ (i.e., show discontinuity) at the cutoff on a scatterplot of participants' scores (Torgerson & Torgerson, 2008).

Regression discontinuity is a robust quasi-experimental design is an excellent approach to take in instances where it is not possible or not ethical to design a randomized controlled trial. In education, it is easy to imagine the vast number of instances in which regression discontinuity may be applied--from instances where students scoring below proficient on the FCAT reading, for example, are assigned to developmental reading classes to instances where students scoring above a cutoff receive a meritorious program (e.g., gifted program) (Shadish et al., 2002). Regression discontinuity designs can also be combined with a randomized control element to strengthen the design (Torgerson & Torgerson, 2008). With the FCAT reading proficiency example this combined process can be illustrated as follows. Students scoring above proficient do not receive developmental reading. Students scoring below proficient are randomly assigned to one of two different types of developmental reading: one is a new, innovative developmental approach and the other is the usual developmental reading approach.

As stated previously, regression discontinuity designs have been used infrequently in education and the social sciences. Only three refereed studies were found by the authors that

have applied regression discontinuity within the past ten years (Bryant, Bryant, Gersten, Scammacca, & Chavez, 2008; Cahan, Greenbaum, Artman, Deluya, & Gappel-Gilon, 2008; Gormley, Gayer, Phillips, & Dawson, 2005). Note that all three studies found within the past ten years have been published since 2004. Due to the stronger standards for research being implemented at the Institute for Education Sciences, regression discontinuity designs should increase as robust alternatives to RCTs; therefore educational researchers and consumers of research should have some familiarity with these designs.

Propensity Score Analysis

Propensity score analysis is also considered a robust quasi-experimental design with non-equivalent groups (Shadish et al., 2002) and is the best approach to take when estimating causal effects from observational data (Rosenbaum & Rubin, 1983; Rubin, 1997). Data is considered observational if it has been collected on students that does not involve manipulation. Introduced by Rosenbaum and Rubin (1983), propensity score analysis is considered an extension of discriminant analysis (Rubin, 1997). The value in propensity score analysis is that it can be used to answer causal questions with observational data in situations where random assignment is not feasible (Rubin, 1997). Observational studies allow for empirical examination of treatment effects, similar to an experiment, however they differ from an experiment in that there is a lack of systematic assignment of students to groups (Dehejia & Wahba, 2002). Researchers that use observational data run the risk of biased data because differences between groups may be due to the treatment or to pre-existing differences between the groups (that could have been balanced out through the random assignment process in a true experiment) (Rosenbaum, 1986).

Propensity score analysis has been used most extensively in medical research (e.g., Connors et al., 1996; Earle et al., 2001; Foody, Cole, Blackstone, & Lauer, 2001; Gunn, Thamilarasan, Watanabe, Blackstone, & Lauer, 2001; Mitra, Schnabel, Neuget, & Heitjan, 2001) and has been slow to migrate to education and the social sciences (Pruzek, 2004). An example of an educational research study where propensity score analysis would be appropriate is examining outcomes of children based on their attendance at a public or private school. In this example, it would likely be difficult to get support of parents to allow their child to be randomly assigned to either a public or private school. Thus a researcher wishing to study this topic is left with conducting a quasi-experiment and matching students on important and relevant covariates. The matching process can become quite cumbersome, however, considering the large number of potentially relevant covariates on which to match. Traditional matching is conducted based on direct matching of covariates whereas propensity score analysis matches on propensity score. In traditional matching, therefore, researchers are limited in the number of covariates on which to match and thus the design of the study is comprised. This often leads to biased results in which differences in the outcome are difficult to attribute to the intervention and may rather be due to covariates that were not considered in the matching process (Rosenbaum, 1986). This bias can then lead to a case of mistaken identity (e.g., false positive or false negative) with the treatment effect (Rosenbaum, 1991).

More specifically, the propensity score is the estimated chance of receiving the treatment based on pretreatment covariates (Rubin, 1997). Given the propensity score, this conditional distribution of the covariates is the same for both treatment and control groups and thus creates a balancing score (Rosenbaum & Rubin, 1983). The beauty in propensity score analysis is the one score is created from the information of the covariates that can be applied in the model rather

than applying multiple covariates in the model--greatly simplifying the model (Rubin, 1997). Unlike other procedures, propensity score analysis does not hinge on parsimony. Rather, all variables that may possible predict the outcome are included (Rutter, 2007). One limitation with the use of propensity score analysis is that the assurance that there will not be systematic differences between groups is based solely on the use of observed covariates. In comparison, random assignment provides the assurance that systematic group differences tend to not exist based on both observed and unobserved covariates. However, sensitivity analysis can be examined to determine the probability that relevant but unobserved covariates were excluded from the model (Rosenbaum, 1991).

Although computing propensity score analysis is a multi-step process, it does not require special statistical software. Additionally, primers (Luellen, Shadish, & Clark, 2005) and how-to tutorials are available (Hahs-Vaughn & Onwuegbuzie, 2006).

Correlational Designs

Correlational studies in intervention research allow researchers to determine the extent to which variables are related without attempting to manipulate any (Frankael & Wallen). Correlations are just that—relationships—and thus causality cannot be inferred from studies that are strictly correlational in nature (Rutter, 2007), although results from correlational studies can suggest causality and the suggested causality is often the spark from which RCTs are created (Wallen & Frankael, 2001). Correlational studies are generally considered less advantageous than quasi-experiments (Rutter, 2007). So why bother with correlational studies? Historically, much has been gained from correlational studies, especially in disciplines where design of RCTs are difficult (Wallen & Frankael, 2001). Possibly the best known example are the multitude of

correlational studies that examined smoking and lung cancer (Wallen & Frankael, 2001). Thus correlational studies are valuable in explanatory studies where clarifying understanding is critical (Wallen & Frankael, 2001). Correlational studies are also beneficial in prediction. For example, predicting college grade point average based on high school grade point average (Wallen & Frankael, 2001). In terms of intervention research, examining the relationship between, for example, a diagnostic measure and a test score may be very beneficial in clarifying the extent to which the diagnostic can be used to predict a test score. New statistical methodologies, such as structural equation modeling and multilevel modeling (e.g., hiearchical linear modeling), are correlational procedures that can be very powerful in examining relationships. Interested readers are encouraged to examine key texts in these areas (e.g., Raudenbush & Bryk, 2002; Schumacker & Lomax, 1996).

Action Research

Action research was first introduced approximately 60 years ago by Kurt Lewin with the intent of seeking solutions to social issues (Lewin, 1946/1948). Action research, or teacher research, is one way to foster meaningful professional development (Cochran & Lytle, 1999), and it can have a powerful impact for the teachers who engage in it (Boles, Kamii, & Troen, 1999; Cochran & Lytle, 1999; Graham, 1998; Hankins, 1998). Teacher researchers indicate that through their investigations they learn about their students (Fecho, 2000), their schools (Herr, 1999), and themselves (Hankins, 1998). Bargal cited eight principles for action research (2008, p. 19):

1. Action research combines a systematic study, sometimes experimental, of a social problem as well as the endeavors to solve it.

2. Action research includes a spiral process of data collection to determine goals, action to implement goals, and assessment of the results of the intervention.
3. Action research requires feedback of the results of intervention to all parties involved in the research.
4. Action research implies continuous cooperation between researchers and practitioners.
5. Action research relies on the principles of group dynamics and is anchored in its change phases. The phases are unfreezing, moving, and refreezing. Decision making is mutual and is carried out in a public way.
6. Action research takes into account issues of values, objectives, and power needs of the parties involved.
7. Action research serves to create knowledge, to formulate principles of intervention, and to develop instruments for selection, intervention, and training.
8. Within the framework of action research, there is an emphasis on the recruitment, training, and support of the change agents.

Although there are cited benefits of teacher research, there are also concerns with effectively supporting teacher researchers and assisting teachers to learn about teacher research (Radencich, 1998).

There are different schools of thought on what action research is or can be as well as how to design action research. Reason and Bradbury (2001) define action research as an idealogy as well as a methodology. The traditional view of action research holds that this form of research is conducted by a practicing teacher within their own classroom with the goal of making instructional changes based upon the results (Little & Rawlinson, 2002)--i.e., taking action based

on the research--although as noted later, this is a narrow perception of what may constitute action research. In addition, qualitative statistical procedures are also often considered the tool of choice for analysis of action research (Clayton et al., 2008), and this is evidenced in a complete chapter devoted to action research in the *Handbook of Qualitative Research* (Kemmis & McTaggart, 2000). However, experiments and quasi-experiments can also be applied in action research (Bargal, 2008) and the application of quantitative methods is recommended depending on the research question addressed in action research (Lankshear & Knobel, 2004).

It is important to note that while action research is often considered a methodology itself, this does not exclude action research studies from being rigorously designed in a systematic fashion (Lankshear & Knobel, 2004). This is one of many myths related to action research that have been dispelled by Lankshear and Knobel (2004) who introduce teacher research more broadly, suggesting that teacher research *is research.* They offer a more global picture of what constitutes research conducted by teachers and emphasize the systematic and rigorous nature of teacher research.

Design Experiments

Design experiments, also termed design studies or teaching experiments, have been embraced in recent decades in education (Gorard, Roberts, & Taylor, 2004) to the extent that an entire issue of *Educational Researcher* was devoted to the topic (2003, *32,* 1). In essence, design experiments allow teachers to examine contextual learning while at the same time designing and creating interventions for the classroom in a systematic way (Gorard et al., 2004). Applying design experiments is not always clear cut: "Design experiments are messier than traditional experiments, because they monitor many dependent variables, characterize the situation

ethnographically, revise the procedures at will, allow participants to interact, develop profiles rather than hypotheses, involve users and practitioners in the design, and generate copius amounts of data of various sorts" (Gorard et al., 2004, p. 581). Before design experiments can be easily incorporated into practice (Gorard et al., 2004), systematic and explicit design experiment models are needed (Kelly & Lesh, 2002).

What is the Role of These Designs in Intervention Research?

The descriptions presented are not only meant to serve as a primer on the designs but also to provide a framework so that discussion on how they can be applied in intervention research is better contextualized. When applied effectively in intervention research, randomized control trials (RCT) will provide the best evidence of causality of all research designs. In addition, relatively simple statistical analysis can be applied to data from RCTs if designed appropriately. However, RCTs may be difficult to apply in many educational settings due to ethical issues or lack of feasibility in randomizing students to the intervention.

In situations where a RCT cannot be applied, quasi-experimental designs may be applied. There are a number of traditional quasi-experimental designs (aka ‘natural experiments’) (e.g., pretest-posttest control group) that may be appropriate for intervention research. Lesser-used quasi-experimental designs that are considered the “next best thing” to RCT include regression discontinuity and propensity score analysis. While both are powerful in providing evidence that may be used for causal inference, both require higher level statistical skills (although no special statistical software is needed). Thus, teachers interested in applying these designs who do not possess the statistical ability needed may want to consider collaborating with university faculty who can provide the statistical expertise. Correlational research may be appropriate in

intervention research when other types of design cannot be applied and/or as a precursor to an experiment in an attempt to explore relationships or determine predictive power of one or more variables.

Action research and design experiments, when applied systematically and designed rigorously using experimental or quasi-experimental methods as well as correlational designs, can provide evidence of causality in intervention research. The key in designing both action research and design experiments for intervention research is being systematic and rigorous. Without this rigor, results become less interpretable and thus evidence from the results less valuable and lacking inference of causality.

The Role of Non-Causal Research Designs in Intervention Research

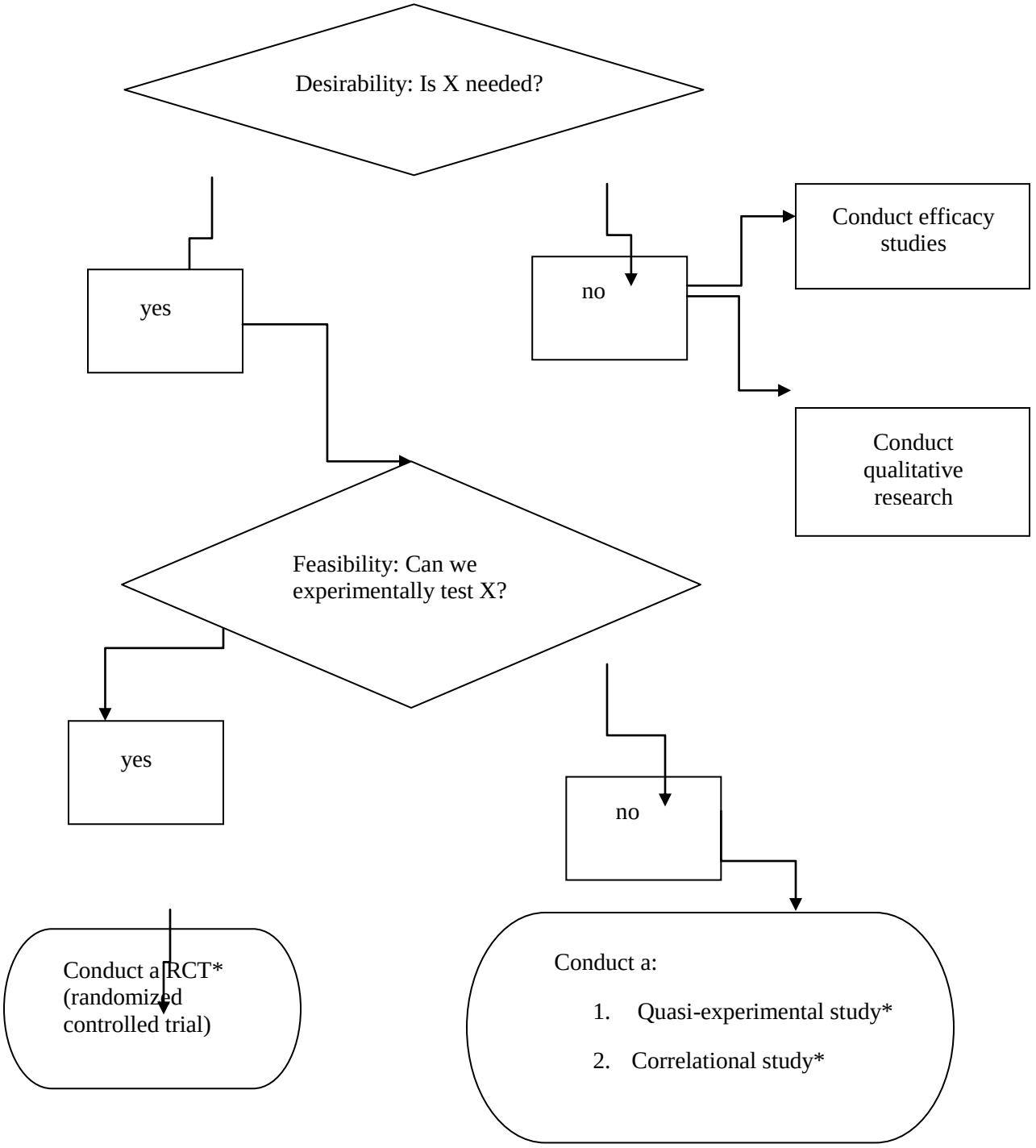
Depending on how the efficacy research is designed, all the designs presented may, in a given situation, be non-causal (even a RCT may fit this category if it is poorly designed). Closer to RCTs are regression discontinuity designs and propensity score analysis and thus the inference of causality becomes clearer, but not perfect, with these designs. Beyond these types of designs, however, there are many research questions that are valuable to examine but do not permit the application of a research design that will allow the inference of causality. For example, theory building is usually initiated by qualitative research such as grounded theory. Other types of qualitative designs such as case studies and phenomenologies provide comprehensive examination of people or events that can help inform a discipline but do not provide inference of causality. Descriptive studies that only describe a situation or group of students may be very valuable in enlightening a research problem but at the same time do not provide evidence of causality. These are just a few examples of non-causal designs. Although results from these

types of designs may not suggest cause and effect when studying whether a particular intervention works, they still hold value if they serve to appropriately answer the research question. Additionally, when designed effectively, they can contribute to understanding the efficacy of an intervention, albeit not causality of such efficacy.

Flowchart for Determining How to Investigate Causal Questions Regarding Educational Interventions

In Figure 1, we present a flowchart that represents a decision-making path a researcher can use when either planning a research study or evaluating a research study that involves intervention research (in other words, does treatment X work?). To begin, we suggest that the first question that needs to be asked is whether X is a desirable treatment. Is it needed? If not, then the path suggests that efficacy studies should be conducted and/or qualitative research studies to determine how to improve X or find where it might be needed. However, if X is found

Figure 1. Does X Work? Flowchart for Determining How to Investigate This Question



**Action research and design experiments may incorporate any of these designs, albeit usually on a relatively smaller scale.*

to be desirable, Figure 1 shows that the next question to ask is whether X is feasible to test experimentally test. If yes, then the solution is to conduct a randomized control trial (RCT) of X against either a control group, another treatment group, or both.

If it is not feasible to conduct an RCT, then we suggest two options to try instead, in order of decreasing validity. The first choice is to conduct a high quality quasi-experimental study where the groups have been carefully selected to match on as many important criteria as possible (e.g., socioeconomic status, age, ability, etc.). This may be accomplished by using one of the lesser used statistical techniques discussed in the second section of our paper, particularly propensity score analysis. In the event of the application of a cutoff score in the research, an additional lesser used technique (regression discontinuity analysis) may be applied. Although complicated statistically, they provide a way of testing causal questions without the need for random assignment into a treatment and control group.

The second option is to conduct a well-designed correlational study, using high quality statistical analyses and measures (such as structural equation modeling or hierarchical linear modeling). These analyses should control for pre-existing differences between groups so that the suggestion of causality results.

With any of these designs, action research or design experiments may be conducted although possibly on a smaller scale, i.e., local studies of the efficacy of a particular treatment. Depending on the design, these studies may be limited in their generalizability to a broader population (i.e., external validity), but they may be very helpful for determining the efficacy of an intervention on a local scale (i.e., internal validity).

We realize that not all researchers may have the resources to apply some of the designs, such as propensity score analysis, in their research. Figure 1, however, serves as a functional heuristic to use when evaluating competing claims about the effectiveness of a particular instructional treatment, something most teachers must do on a regular basis. The remaining options are well within the reach of most educational researchers. Applying randomized controlled trials and quasi-experimental designs are ones that we encourage more researchers and teachers to adopt. Determining the efficacy of an intervention on a local scale is not only practically useful—it can help one's day to day teaching—but it is also important for the broader research community in that it may reveal conditions under which the treatment X works or does not work as well. In the next section, we present an example of a study designed for testing whether a particular intervention works in a local setting using a randomized controlled trial in the context of an action research study as the framework for our design.

Example Study

We present this example study applying a randomized controlled trial (albeit limited randomization at the classroom level) in the context of an action research design because we want to appeal to the broadest possible audience, discussing a design, that, when conducted with appropriate methodological rigor, can be used by both teachers and educational researchers to investigate the efficacy of an instructional intervention at the local level. To begin, let us say that one is interested in a new mathematics curriculum that focuses predominantly on complex problem solving, with less emphasis given to procedural fluency. One might be the actual teacher asked to make this curriculum change or an educational researcher interested in the local impact of this intervention on both teacher efficacy and student achievement. In this illustration, we will assume that we are the teacher who has been asked to make this curriculum change. Let us also

assume that we are teaching at a middle or high school level in that we are teaching two sections of the same class. With administrative approval, one section will be taught with the new curriculum while the second section will be taught with the previously used curriculum. Deciding on which class to administer the new curriculum to may be done on the basis of a coin toss. To determine the comparability of mathematics skills of students prior to initiating the new curriculum, we collect data on the previous year's mathematics portion of our state's standardized assessment as well as their grade received in the mathematics class taken previous to this (assuming all students completed the same mathematics class). So that we are not overburdening students with additional testing, these scores will serve as their pretest scores. However, because these instruments were not designed to specifically measure the skills acquired from exposure to the new curriculum, we may also want to consider administering either a teacher-created test or other assessment that is designed to more directly measure skills that should have been acquired at the conclusion of teaching the new curriculum.

How can we be assured that we are not biasing our own study by teaching 'better' using the new curriculum? One thing we decide to do is to have a colleague who also teaches math observe in our classroom at various times throughout the year to check our fidelity of implementing both the status quo curriculum as well as the new curriculum. We also decide to keep a detailed teaching log in which we reflect on our teaching and track any unusual (and not so unusual) occurrences in both classes. Our lesson plans will further provide documentation of adherence to curriculum. In combination, these three elements (colleague observation, teaching log and lesson plans) will provide some suggestion of our fidelity to the curriculum, whether old or new.

At the conclusion of the semester, we track students' grades. At the end of the academic year, we also collect information on this year's state mathematics assessment. If we also administered an additional pre-assessment, either created by us or that was provided with the curriculum, we would want to administer that same assessment as a posttest at the conclusion of the semester. In terms of analysis, depending on finding comparable groups on the pretest measures, we may be able to apply basic statistics such as an independent *t* test, analysis of variance, or multiple regression. As stated previously, one limitation to this research is that it was not possible to draw a random sample of students. More limiting, however, is that it was not possible to randomly assign at the student level. That limits our ability to generalize to a larger population (i.e., external validity). However, depending on how comparable we find our groups to be prior to initiating the study, we may have relatively good internal validity—the ability to find evidence that our intervention is the reason for the change in mathematics achievement.

Conclusion

The purpose of this paper was to present a conceptual framework that summarized state of the art research designs for investigating causal questions regarding educational interventions. This paper examined examples of the types of research questions posed by educational researchers and identified intervention research as a particular type of causal question. Next, research designs for answering causal questions were reviewed, paying particular attention to and providing examples of the application of lesser used designs (such as regression discontinuity and propensity score analysis) that may overcome some of the limitations faced when randomized experimental designs are not feasible or appropriate. The role of these designs in teaching and research were also discussed as was the role of non-causal research designs in education. Next, a useful graphical organizer to aid in interpreting existing research and planning

future research on educational interventions was presented. Finally, a design for a sample study to test the effects of a new math program that may be used as a model for participants who may be considering planning their own research was presented.

This paper was designed to serve dual purposes: First, to help teachers, teacher educators, district professional development directors and others understand how to judge quality intervention research, and second, to provide this same audience a way to participate and contribute to their own high quality intervention research.

References

Bargal, D. (2008). Action research: A paradigm for achieving social change. *Small Group Research, 39*(1), 17-27.

Boles, K. C., Kamii, M., & Troen, V. (1999, April). *Transformative professional development: Teacher research, inquiry, and the culture of schools.* Paper presented at the American Educational Research Association annual meeting, Montreal, Canada.

Bryant, D. P., Bryant, B. R., Gersten, R., Scammacca, N., & Chavez, M. M. (2008). Mathematics intervention for first- and second-grade students with mathematics difficulties; The effects of Tier 2 intervention delivered as booster lessons. *Remedial and Special Education 29*(1), 20-32.

Cahan, S., Greenbaum, C., Artman, L., Deluya, N., & Gappel-Gilon, Y. (2008). The differential effects of age and first grade schooling on the development of infralogical and logico-mathematical concrete operations. *Cognitive Development, 23*, 258-277.

Campbell, D. T. (1984). Foreword. In W. M. K. Trochim (Ed.), *Research design for program evaluation: The regression discontinuity approach* (pp. 15-43). Beverly Hills, CA: Sage.

Center for Psychology in Schools and Education. (2008). A guide to the use of Randomized Controlled Trials (RCTs) in assessing intervention effects: The promise of multiple methods [Electronic Version]. Retrieved September 8, 2008, from http://www.apa.org/ed/cpse/multmethod08.pdf

Clayton, S., O'Brien, M., Burton, D., Campbell, A., Qualter, A., & Varga-Atkins, T. n. (2008). 'I know it's not proper research, but...': How professionals' understandings of research can frustrate its potential for CPD. *Educational Action Research, 16*(1), 73-84.

Cochran, M., & Lytle, S. L. (1999). The teacher research movement: A decade later. *Educational Researcher, 28*(7), 15-25.

Connors, A. F., Speroff, T., Dawson, N. V., Thomas, C., Harrell, F. E., Wagner, D., et al. (1996). The effectiveness of right heart catheterization in the initial care of critically ill patients. *JAMA, 276*(11), 889-918.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis for field settings*. Boston: Houghton Mifflin.

Cook, T. D., & Shadish, W. R. (1994). Social experiments: Some developments over the past fifteen years. *Annual Review of Psychology, 45*(545-580).

Crain, W. (2005). *Theories of development: Concepts and applications* (5th ed.). Upper Saddle River, NJ: Pearson.

Dehejia, R. H., & Wahba, S. (2002). Propensity score-matching methods for nonexperimental causal studies. *The Review of Economics and Statistics, 84*(1), 151-161.

Earle, C. C., Tsai, J. S., Gelber, R. D., Weinstein, M. C., Neumann, P. J., & Weeks, J. C. (2001). Effectiveness of chemotherapy for advanced lung cancer in the elderly: Instrumental variable and propensity analysis. *Journal of Clinical Oncology, 19*(4), 1064-1070.

Elkind, D. (2001). *The hurried child: Growing up too fast too soon*. Reading, MA: Addison-Wesley.

Fecho, B. (2000). Critical inquiries into language in an urban classroom. *Research in the Teaching of English, 34*(4), 368-395.

Foody, J. M., Cole, C. R., Blackstone, E. H., & Lauer, M. S. (2001). A propensity analysis of cigarette smoking and mortality with consideration of the effects of alcohol. *The American Journal of Cardiology, 87*, 706-711.

Frankael, J. R., & Wallen, N. E. (2005). *How to design and evaluate research in education* (6th ed.): McGraw Hill.

Glaser, B. G., & Strauss, A. (1967). *Discovery of grounded theory: Strategies for qualitative research*. Mill Valley, CA: Sociology Press.

Gorard, S., Roberts, K., & Taylor, C. (2004). What kind of creature is a design experiment? *British Educational Research Journal, 30*(4), 577-590.

Gormley, W. T., Gayer, T., Phillips, D., & Dawson, B. (2005). The effects of universal pre-K on cognitive development. *Developmental Psychology, 41*(6), 872-884.

Graham, P. (1998). Teacher research and collaborative inquiry: Teacher educators and high school English teachers. *Journal of Teacher Education, 49*(4), 2550265.

Gunn, P. A., Thamilarasan, M., Watanabe, J., Blackstone, E. H., & Lauer, M. S. (2001). Aspirin use and all-cause mortality among patients being evaluated for known or suspected coronary artery disease. *JAMA, 286*(10), 1187-1194.

Hahs-Vaughn, D. L., & Onwuegbuzie, A. J. (2006). Estimating and using propensity score analysis with complex samples. *Journal of Experimental Education, 75*(1), 31-65.

Hankins, K. H. (1998). Cacophony to symphony: Memoirs in teacher research. *Harvard Educational Review, 68*(1), 80-95.

Herr, K. (1999). Unearthing the unspeakale: When teacher research and political agendas collide. *Language Arts, 77*(1), 10-15.

Kelly, A., & Lesh, R. (2002). Understanding and explicating the design experiment methodology. *Building Research Capacity, 3*, 1-3.

Kerlinger, F. N., & Lee, H. B. (2000). *Foundations of behavioral research* (4th ed.). United States: Wadsworth Thomson Learning.

Kohn, A. (2000). *The case against standardized testing: Raising the scores, ruining the schools*. Portsmouth, NH: Heinemann.

Lankshear, C., & Knobel, M. (2004). *A handbook for teacher research: From design to implementation*. New York: Open University Press.

Lewin, K. (1946/1948). Action research and minority problems. In G. W. Lewin (Ed.), *Resolving social conflicts* (pp. 201-216). New York: Harper Row.

Little, M., & Rawlinson, D. (2002). *Becoming an action researcher to improve learning in your classroom*. Daytona Beach, FL: Project CENTRAL.

Luellen, J. K., Shadish, W. R., & Clark, M. H. (2005). Propensity scores: An introduction and experimental test. *Evaluation Review, 29*(6), 530-558.

Meier, P. (1972). The biggest public health experiment ever: the 1954 field trial of the Salk poliomyelitis vaccine. In J. M. Tanur, F. Mosteller, W. H. Kruskal, R. F. Link, R. S. Pieters & G. R. Rising (Eds.), *Statistics: A guide to the unknown* (pp. 120-129). San Francisco: Holden-Day.

Mitra, N., Schnabel, F. R., Neuget, A. I., & Heitjan, D. F. (2001). Estimating the effect of an intensive surveillance program on stage of breast carcinoma at diagnosis. *Cancer, 91*(9), 1709-1715.

National Institute of Early Education Research. (2008). A benefit-cost analysis of the Abecedarian Early Childhood Intervention [Electronic Version]. Retrieved September 8, 2008, from http://nieer.org/docs/?DocID=57

Pruzek, R. M. (2004). Applications and graphics for propensity score analysis. Retrieved September 14, 2005, from http://www.albany.edu/~jz7088/documents/04-29-2004/psa%5B1%5D.applications.graphics.2004.pdf

Radencich, M. C. (1998). Planning a teacher research course: Challenges and quandries. *Educational Forum, 62*(3), 265-272.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage.

Reason, P., & Bradbury, H. (Eds.). (2001). *Handbook of action research*. Thousand Oaks, CA: Sage.

Rosenbaum, P. R. (1986). Dropping out of high school in the United States: An observational study. *Journal of Educational Statistics, 11*(3), 207-224.

Rosenbaum, P. R. (1991). Discussing hidden bias in observational studies. *Annals of Internal Medicine, 115*(11), 901-905.

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika, 70*(1), 41-55.

Rubin, D. B. (1997). Estimating causal effects from large data sets using propensity scores. *Annal of Internal Medicine, 127*, 757-763.

Rutter, M. (2007). Proceeding from observed correlation to causal inference: The use of natural experiments. *Perspectives on Psychological Science, 2*(4), 377-395.

Schumacker, R. E., & Lomax, R. G. (1996). *A beginner's guide to structural equation modeling*. Mahwah, N.J.: L. Erlbaum Associates.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houston Mifflin.

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

Sprott, J. E. (2002). *Raising young children in an Alaskan Inupiaq village: The family, cultural, and village environment of rearing*

Westport: CT: Greenwood.

Torgerson, D. J., & Torgerson, C. J. (2008). *Designing randomised triles in health, education and the social sciences*. New York: Palgrave Macmillan.

Trochim, W. M. K. (1980). The regression-discontinuity design in Title I evaluation: Implementation, anlaysis, and variations. Unpublished Doctoral dissertation. Northwestern University.

Tyre, P. (2008, September 8). Struggling school-age boys [Electronic Version]. *Newsweek*. Retrieved September 15, 2008, from http://www.newsweek.com/id/157898

U.S. Department of Education. (n.d.). Four pillars of NCLB [Electronic Version]. Retrieved September 8, 2008, from http://www.ed.gov/nclb/overview/intro/4pillars.html

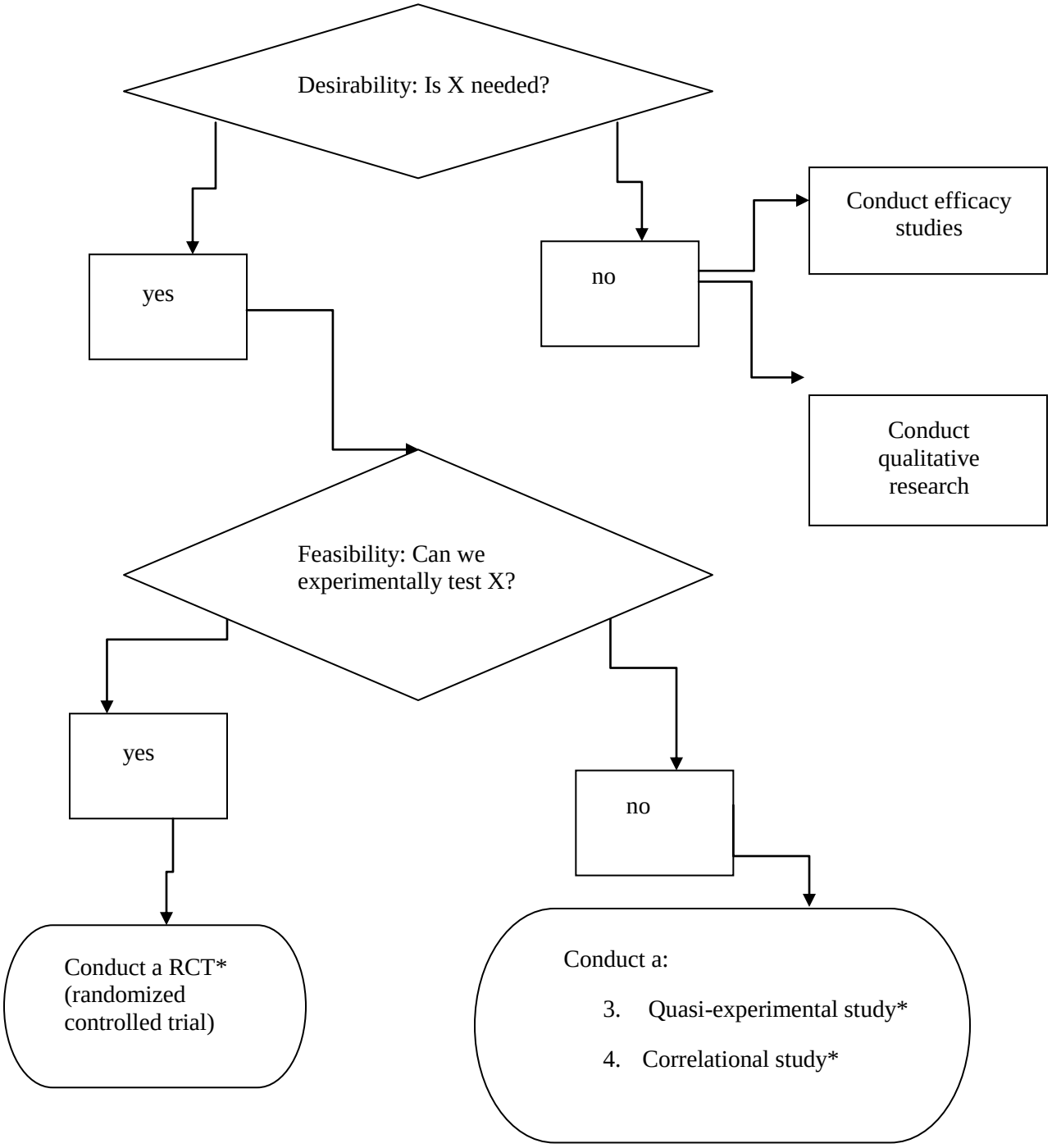
U.S. Department of Health and Human Services. (2001). *Head Start FACES: Longitudinal findings on program performance (Third progress report)*. Retrieved September 8, 2008, from http://www.acf.hhs.gov/programs/opre/hs/faces/reports/perform_3rd_rpt/perform_3rd_rpt.pdf

Wallen, N. E., & Frankael, J. R. (2001). *Educational research: A guide to the process* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Washington Crossing Historic Park. (n.d.). Chores! [Electronic Version]. Retrieved September 8, 2008, from http://www.ushistory.org/WashingtonCrossing/kids/chores.htm

Whitehurst, G. J. (n.d.). Director's statement [Electronic Version]. Retrieved September 8, 2008, from http://ies.ed.gov/director/

Figure 1. Does X Work? Flowchart for Determining How to Investigate This Question



**Action research and design experiments may incorporate any of these designs, albeit usually on a relatively smaller scale.*