# Chapter 3: Diversity of Educational Research

## Published Examples of Research Concepts

**Contents**

Dimensions of Research
    Conflict, Divorce, and Adjustment
Large-Scale Policy Research
    Class Size Reduction
Small-Scale Evaluation Research
    Block Scheduling
References

### Dimensions of Research

Conflict, Divorce, and Adjustment

In a study to illustrate the dimensions of research focused on the influence of divorce on young adolescents' adjustment, Long, Slater, Forehand, and Fauber's (1988) review of the literature on this topic revealed studies showing positive, negative, and neutral influences of divorce on adjustment. Conflicting research findings may be due, in part, to the different types of research used to answer a research question. Long and colleagues (1988) hypothesized that the conflicting findings were linked to one extraneous variable in particular: the level of conflict. Some types of research may be able to control for extraneous variables such as conflict, while others types may not. Experiments, for example, are able to control for extraneous influences through the mechanism of random assignment. Some studies, of course, cannot use random assignment because it is not possible. This was the case in this study, where it would not have been possible to randomly assign students to a divorce group or a control group. Other types of research may single out an extraneous variable statistically (as in a correlational study), build it into the design (as in causal comparative research, illustrated

by this study), or maybe even describe it in a story or describe it by a metaphor (as in a qualitative study).

The design section of a published research report helps us classify the study according to type. The Long and colleagues' (1988) study represents the causal comparative dimension of research. The distinguishing feature of causal comparative research is the classification of subjects into groups based on some attribute, followed by a search for the causes or consequences of differences across the groups. In this causal comparative study, the concern was focused on the consequences of divorce, especially with continuing conflict, on the adjustment of young adolescents. (Investigating the causes of divorce is, for sure, a completely different study.)

A very close cousin to causal comparative research is correlational research. Correlational research, as we've seen, does not study group contrasts. It examines linkages between variables (reflecting preexisting attributes) using statistical techniques that retain subjects' scores at the individual level (no groupings). For example, Long and colleagues (1988) could have investigated the correlational link between conflict scores and adjustment scores. To do this, each student would be scored (or "scaled") on the level of reduced conflict between their divorced parents, say, on a scale of 1 to 20 (from greatly reduced to hardly reduced at all). The adjustment scores would then be correlated with conflict reduction scores to see whether a connection, or relationship, exists between them. If they are related, say, better adjustment is linked to less conflict, then the researcher (as in causal comparative studies) is still uncertain about the basis of the relationship. This is because less conflict could cause better adjustment, better adjustment could cause less conflict, or some third variable, like poverty level, could cause both poor

adjustment and greater conflict.

Long and colleagues (1988) classified their adolescent subjects into three groups, including one from divorced families with high parent conflict, one from divorced families with low parent conflict, and one from "intact" families (the comparison group). Given the causal comparative approach to research, we know that Long and colleagues (1988) grappled with the problem of groups being different not only on the classifying variable but in other important ways as well. Their use of the "intact" group is closely akin to use of a control group in experimental research. This use of the group is necessary to provide a baseline comparison of sorts, for the researchers wanted to know whether the adjustment of children from divorced families was negatively affected in relation to a nondivorced (intact) family.

These researchers used matching as a method to equate the groups so that they were not comparing "apples and oranges." Only when the intact and divorced groups are considered comparable on other important dimensions can the researchers be comfortable concluding that the differences in adjustment are due to the divorce. These researchers matched the groups on age, sex, and family SES (socioeconomic status), or the "Big Three" in educational research. These are common matching variables because they tend to create comparability better than any other set of three variables (i.e., knowing a person's age, sex, and SES probably tells you more about a person than any other three variables). Matching tries to accomplish in causal comparative studies what random assignment does so well in true experimental studies. You can see that it makes little sense to compare children who differ on a divorce dimension and on an SES dimension. If a difference in adjustment were found between divorced, lower SES children and

nondivorced, higher SES children, one could not disentangle the influence of divorce from that of SES (the variables are confounded). Adjustment differences could be linked to either one (or both).

Causal comparative studies, naturally, have to confront head-on the challenges associated with the operational definitions of constructs. In fact, they often have greater challenges if the basis of the group classification itself is related to a construct, as it was in Long and colleagues' (1988) study. The influence of conflict was a primary concern of these researchers, and it was operationally defined by scores on the O'Leary-Porter Scale and the Divorce Conflict Measure, which was specifically designed for their study. (The construct of adjustment, too, needed an operational definition. The anxiety-withdrawal scores and the conduct disorder scores from the Revised Behavior Problem Checklist were used as measures of "internalizing" and "externalizing" problems, respectively.)

The statistical analysis used by Long and colleagues (1988) revealed that the research was clearly quantitative, not qualitative, in nature. Numbers, not words, summarize their findings, and it was the researchers' belief that adjustment is best measured on a numerical continuum. Other researchers, however, might have answered this research question by examining the *qualities* of adjustment, possibly through interviews, interpretations from observations, or therapists' notes. The study was *inferential* in nature, not merely descriptive. This is clear because of the levels of *statistical significance* reported in the analysis (as *p* values). This tells us that the researchers wish to generalize beyond their sample and make statements about relationships that are likely to exist in the population of students like those studied. Without statistical significance (*p* values), researchers are restricting their statements

(descriptions) to the sample itself.

The experimental versus quasi-experimental distinction is not relevant in this study since the study was not experimental in nature. How could the researchers manipulate divorce and ensuing levels of conflict? With regard to the single-subject versus group distinction, we can see that this research is clearly group oriented. And as far as teacher versus traditional research is concerned, this study appears to be an example of traditional research. Teacher research, by contrast, is very local and conducted for the purpose of answering a specific applied problem within a single classroom (or similar personalized context). Teacher research does not seek to generalize beyond the confines of the study, nor is it concerned with a general understanding of a broader phenomenon. It is, though, very "action" oriented. Further, Long and colleagues' (1988) study was rather small scale in its scope, intent on investigating (evaluating) the role that one construct (conflict) might play in the research on divorce and its influence on adjustment. It was clearly not positioned to offer suggestions for large-scale changes in practice or policy.

Finally, Long and colleagues (1988) provided a good illustration of the need for tempered conclusions during the conduct of causal comparative research. After finding that adolescents from divorced (and high conflict) families were functioning at a lower level than the subjects in the other two groups, these researchers reminded us that adolescent adjustment problems can cause parent conflict (not the other way around) and that their study cannot determine which factor might be the *active mechanism* (that term presumably meaning "cause").

In summary, we have seen that this published research was quantitative,

inferential, causal comparative, group-oriented, traditional, and small scale in its scope (and was neither true nor quasi-experimental).

## Large-Scale Policy Research

### Class Size Reduction

In 1996 California began a class size reduction initiative for grades K–3. Costing several billion dollars, the legislation enabled a reduction in class size from about 30 students to 20 or fewer. Stecher, McCaffrey, and Bugliari (2003) noted that the value of large-scale class size reduction efforts is still unproven, and 30 years of research on the relationship between class size to student performance has yielded "mixed results." (Research findings in education are often a function of differences in research methodology and orientation, in part explained by the seven dimensions described in Chapter 3.) Stecher and colleagues (2003), however, did conclude from their review of the research literature that controlled experiments on the effects of class size reduction have "tipped the policy scales firmly in favor of smaller classes." They wrote, "Can these effects be achieved on a large scale? The experience of California offers important insights into class size reduction as a statewide policy. The size and complexity of initiating a class size reduction program in the nation's most populous state and the diversity of California's classroom represents an important, real-world test of the effectiveness of [class size reduction] as a broad-based policy" (Introduction section, para. 3).

The amount of data available for large-scale policy research is staggering in California alone. (Beginning in 1998, California students in Grades 2 to 11 completed the

Stanford Achievement Test Series, or SAT-9, each spring.) To ease their task somewhat, researchers Stecher and colleagues (2003) used the SAT-9 reading, math, and language *scale scores* (as opposed to the raw scores, percentile ranks, or normal curve equivalent scores), since the scale scores are comparable across ranges and equated across grade levels (enabling cross-grade comparisons). Their final sample consisted of nearly 2,500 California elementary schools that permitted a comparison of grades K–3 over time as the class size reduction was implemented. (Their samples are referred to as *cohorts*, or groups studied over time because they share a similar characteristic.) Over 100,000 students were divided into cohorts, depending on when during their grade level progression the class size reduction began. The researchers noted that their school-level analyses, albeit complex, were "less susceptible to confounding from external sources" and were "able to control for student mobility by only including students who attended the same school [during class size reduction] from kindergarten through second or third grade" (Caveats section, para.1).

Despite other problems related to control (the California class size reduction, after all, was not a true experiment), Stecher and colleagues (2003) could only compare with confidence a 1-year difference in exposure, that is, from the first-grade larger classes to subsequent smaller classes. They also noted that class size reduction co-occurred with other significant policy and program changes, such as new state standards and curricula, changes in grade-level promotion policies, new accountability systems that offered large rewards for increases in test scores, and the elimination of traditional bilingual education programs. Problems such as these are common with large-scale policy research because the research designs that permit such large-scale data collection and analyses are rarely, if

ever, well-controlled experiments. The researchers also stated that the accountability system had "created a high-stakes atmosphere that may lead to changes in test scores that are independent of actual changes in achievement" (Caveats section, para. 4). (Chapter 8 examines in greater detail problems like these that threaten the validity of conclusions. Also, Chapter 7 focuses on the meaningfulness, or lack thereof, of measures used in educational research.) Stecher and colleagues (2003) realized that both external sources of influence and problems with measurement can greatly distort research findings.

What did Stecher and colleagues (2003) find in their complex, albeit compromised, data? They found that class size reduction had no relationship to achievement (a "null finding"), at least when comparisons were restricted to a 1-year exposure difference (2 years of exposure versus 3 years of exposure) given the problems related to control in their massive data set. The researchers offered a "cautious interpretation" (and advised against a "pessimistic" conclusion) by noting their findings were consistent with two possible inferences: "a) reduced size classes have no effect, or b) two, three or four years of exposure do have a positive effect" (Caveats section, para. 6). The later inference must await further research because the data set, despite its size, did not permit controlled comparisons between larger and smaller classes over 2 or more consecutive years. This example show that the enormity of data collection does not compensate for inherent control problems.

## Small-Scale Evaluation Research

### Block Scheduling

Researchers Lewis, Cobb, Winokur, Leech, Viney, and White (2003) reviewed

the published research on block scheduling (90- to120-minute class periods) in use at many secondary schools. They made this familiar judgment: "It is difficult to produce any consistent conclusions from the recently published literature on block scheduling as most researchers disagree about the positive and negative effects of [block scheduling]" (Conclusion section, para 1). They also reported that the literature is consistent in its inconsistency! Lewis and colleagues (2003) decided to evaluate the achievement effects of a block scheduling program at one junior high school using a similar junior high school that used a traditional schedule as the comparison group. Specifically, they asked two questions: What influence does block scheduling have on language arts achievement? Might that influence be related to student gender and prior achievement levels?

To answer these questions, they selected two similar schools in their district, one that used block scheduling and one that did not. Students in the ninth-grade language arts courses at the two schools were matched on gender and prior achievement levels in order to make them comparable. A total of about 120 students were used in those comparisons. (This small-scale, local research—two schools and one district—is contrasted with the large-scale, statewide research done by Stecher and colleagues described above, which involved several thousand schools and over 100,000 students.) The primary dependent variable (described more fully in Chapter 5) was a standardized achievement measure in language arts.

The researchers' results supported the implementation of block scheduling since students with blocked schedules scored significantly higher on the achievement test than did the students with traditional schedules. No differences were found between males and

females. There was, however, clear evidence that lower-achieving students benefitted more from the blocking. The influence of block scheduling, at least in this two-school comparison in language arts, depended on the prior achievement level of students. This is known as an *interaction.* The fact that the influence of the scheduling variable *depended* on students' prior achievement levels offers a clue why research in this area in not consistent: The influence of block scheduling depends on one (or many) student characteristics. The characteristics of the sample, therefore, may influence what type of result is found. The difficult topic of interaction (and its interpretation) is covered in more detail in Chapter 10.

## References

Lewis, C., Cobb, R. B., Winokur, M., Leech, N., Viney, M., & White, W. (2003). The effects of full and alternative day block scheduling on language arts and science achievement in a junior high school. *Education Policy Analysis Archives, 41*(11). Retrieved August 3, 2005, from http://epaa.asu.edu/epaa/v11n41/

Long, N., Slater, E., Forehand, R., & Fauber, R. (1988). Continued high or reduced interparental conflict following divorce: Relation to young adolescent adjustment. *Journal of Consulting and Clinical Psychology, 56*(3), 467–469.

Stecher, B. M, McCaffrey, D. F., & Bugliari, D. (2003). The relationship between exposure to class size reduction and student achievement in California. *Education Policy Analysis Archies, 11*(40). Retrieved August 3, 2005, from http:epaa.asu.edu/epaa//v11n40/