Primary, Secondary, and Meta-Analysis of Research[1]

GENE V GLASS
Laboratory of Educational Research
University of Colorado

My subject is data analysis at three levels. *Primary analysis* is the original analysis of data in a research study. It is what one typically imagines as the application of statistical methods.

Secondary analysis is the re-analysis of data for the purpose of answering the original research question with better statistical techniques, or answering new questions with old data. Secondary analysis is an important feature of the research and evaluation enterprise. Tom Cook (1974) at Northwestern University has written about its purposes and methods. Some of our best methodologists have pursued secondary analyses in such grand style that its importance has eclipsed that of the primary analysis. We can cite with pride some state of the art documents: the Mosteller-Moynihan secondary analysis of the Coleman study; the Campbell-Erlebacher analysis of the Ohio-Westinghouse Headstart evaluation; and the Elashoff-Snow secondary analysis of Pygmalion in the Classroom, to name three.

About all that can effectively be done to insure that secondary analyses of important studies are carried out is to see that the data from the original studies are preserved and that secondary analyses are funded. The preservation of original data could improve. Last month, one of our graduate students, Karl White, spent 15 hours and made 30 phone calls attempting to obtain from the government a copy of the data tapes for the Coleman study only to learn in the end that they had been irretrievably filed in unmarked tape cannisters with some 2,000 other unmarked data tapes. Tom Cook remarked in an Annual Meeting symposium on secondary analysis that you can get the data if you have chutzpah or if you're sociometrically well-connected. The whole business is too important to be treated so casually. On the other extreme, one can point with satisfaction to the ready availability to any researcher of the data tapes from Project TALENT or the National Assessment of Educational Progress.

Others are advancing the practice of secondary analysis. My major interest currently is in what we have come to call—not for want of a less pretentious name—the *meta-analysis* of research. The term is a bit grand, but it is precise, and apt, and in the spirit of "meta-mathematics," "meta-psychology," and "meta-evaluation." Meta-analysis refers to the analysis of analyses. I use it to refer to the statistical analysis of a large collection of analysis results from individual studies for the purpose of integrating the findings. It connotes a rigorous alternative to the casual, narrative discussions of research studies which typify our attempts to make sense of the rapidly expanding research literature.

The need for the meta-analysis of research is clear. The literature on dozens of topics in education is growing at an astounding rate. In five years time, researchers can produce literally hundreds of studies on IQ and creativity, or impulsive vs. reflective cognitive styles, or any other topic.

In education, the findings are fragile; they vary in confusing irregularity across contexts, classes of subjects, and countless other factors. Where ten studies might suffice to resolve a matter in biology, ten studies on computer assisted instruction or reading may fail to show the same pattern of results twice. This is particularly true of those questions that are more properly referred to as outcome evaluation than analytic research. Research on attention in learning or concept formation may sometimes

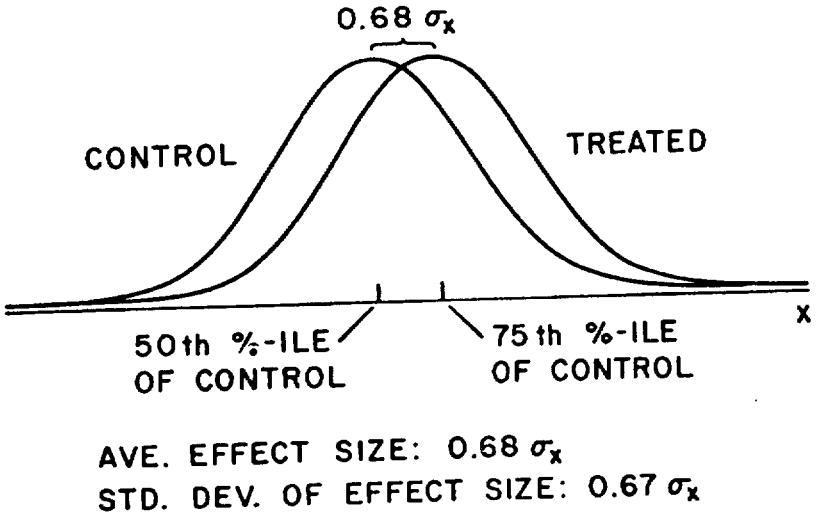

Figure 1. *Normal curves illustrating the aggregate effect of psychotherapy in relation to untreated control groups. (Data based on 833 effect size measures from 375 studies, representing about 40,000 treated and untreated subjects.)*

progress along a systematic course; studies are designed from the findings of previous studies and researchers have a sense of what is known and what must be asked next. Outside the laboratory, however, inquiry is less cumulative and evolutionary. Five hundred studies on class size or ability grouping can accumulate; they will defy simple summary. Their meaning can no more be grasped in our traditional narrative, discursive review than one can grasp the sense of 500 test scores without the aid of techniques for organizing, depicting, and interrelating data.

A common method of integrating several studies with inconsistent findings is to carp on the design or analysis deficiencies of all but a few studies—those remaining frequently being one's own work or that of one's students or friends—and then advance the one or two "acceptable" studies as the truth of the matter. This approach takes design and analysis too seriously, in my opinion. I don't condone a poor job of either; but I also recognize that a study with a half dozen design and analysis flaws may still be valid. Most research criticism I read—and some that I've written—is airy speculation, unbefitting an empirical science. It is an empirical question whether relatively poorly designed studies give results significantly at variance with those of the best designed studies; my experience over the past two years with a body of literature on which I will report in a few minutes leads me to wonder whether well-designed and poorly-designed experiments give very different findings. At any rate, I believe the difference to be so small that to integrate research results by eliminating the "poorly done" studies is to discard a vast amount of important data.

Ken Boulding made one of his off-hand observations that sticks with you, first seeming flip, then serious but wrong, and finally impressive for the simple truth it contains. "Knowledge exists in minds, not in books," he wrote. Before what has been found can be used, before it can persuade skeptics, influence policy, affect practice, it must be known. Someone must organize it, integrate it, extract the message. A hundred dissertations are mute. Someone must read them and discover what they say. And we have reached the point in our field where new methods of discovering knowledge in "findings" or "results" are needed. The armchair literature review in which one cites a couple dozen studies from the obvious journals can't do justice to the voluminous literature of educational research that we now confront.

With the poet's uncanny prescience, T. S. Eliot asked in his poem, "The Rock," "[W]here is the knowledge we have lost in information?" We are inundated with information. The ERIC system fills over two million document requests yearly. Some have termed our predicament "the misinformation explosion." I assess it differently; we face an abundance of information. Our problem is to find the knowledge in the information. We need methods for the orderly summarization of studies so that knowledge can be extracted from the myriad individual researches.

At the same time, a reorientation of attitudes and values is needed. Maurice Goldhaber, former director of the Brookhaven Laboratories, would remind his non-teaching staff that "a good review is the moral equivalent of teaching." In our field, a good review is the intellectual equivalent of original research. Lewis M. Branscomb, former director of the National Bureau of Standards, wrote in a *Science* editorial last year that "when professional advancement and peer recognition are so heavily oriented toward original discovery, and research funding is largely restricted to original . . .research, it is hard to motivate a scientist to write scholarly reviews."

In educational research, we need more scholarly effort concentrated on the problem of finding the knowledge that lies untapped in completed research studies. We are too heavily invested in pedestrian reviewing where verbal synopses of studies are strung out in dizzying lists. The best minds are needed to integrate the staggering number of individual studies. This endeavor deserves higher priority now than adding a new experiment or survey to the pile.

Those who manage financial support of educational research and

evaluation must realize that the integration of research findings is a scholarly undertaking requiring costly literature searches, extensive data analyses, and time to reflect and write. Branscomb went on to write in the same editorial that ". . .federal science policy seems to make support for review scholarship the stepchild of research support.... Big money, fortunately, still goes to original research—the fun part every scientist likes best. Support for review. . .languishes."

The problems of meta-analysis that we face in an applied-evaluative field are uniquely important. They have seldom been rigorously examined.

Dick Light and Paul Smith at Harvard addressed some technical problems in their 1971 *Harvard Educational Review* paper entitled "Accumulating Evidence: Procedures for Resolving Contradictions among Different Research Studies." Their cluster approach is a sensible solution to some problems of meta-analysis when one has raw data available and can meet their statistical assumptions. Our own techniques are working at quite a different level, however. We don't rely on availability of original data, which are as evanescent as the morning dew; and we take a more liberal position than Light and Smith with respect to the criteria studies must meet before they can be compared.

Scholars have been moving toward meta-analysis out of necessity. I recall reading, while a graduate student, Astin & Ross's (1960) review of the experimental literature on the effects of glutamic acid on intelligence of retarded children. Twenty or thirty experiments at the time gave conflicting results. The earlier studies tended to show that glutamic acid raised IQs; the later studies didn't. When they crosstabulated negative vs. positive results with whether the experiments were well or poorly designed, the association was clear: the seemingly "positive" findings came from the poorly designed experiments; good experiments showed no effects.

Wilbur Schramm (1962) attempted to integrate over four hundred controlled studies of the effects

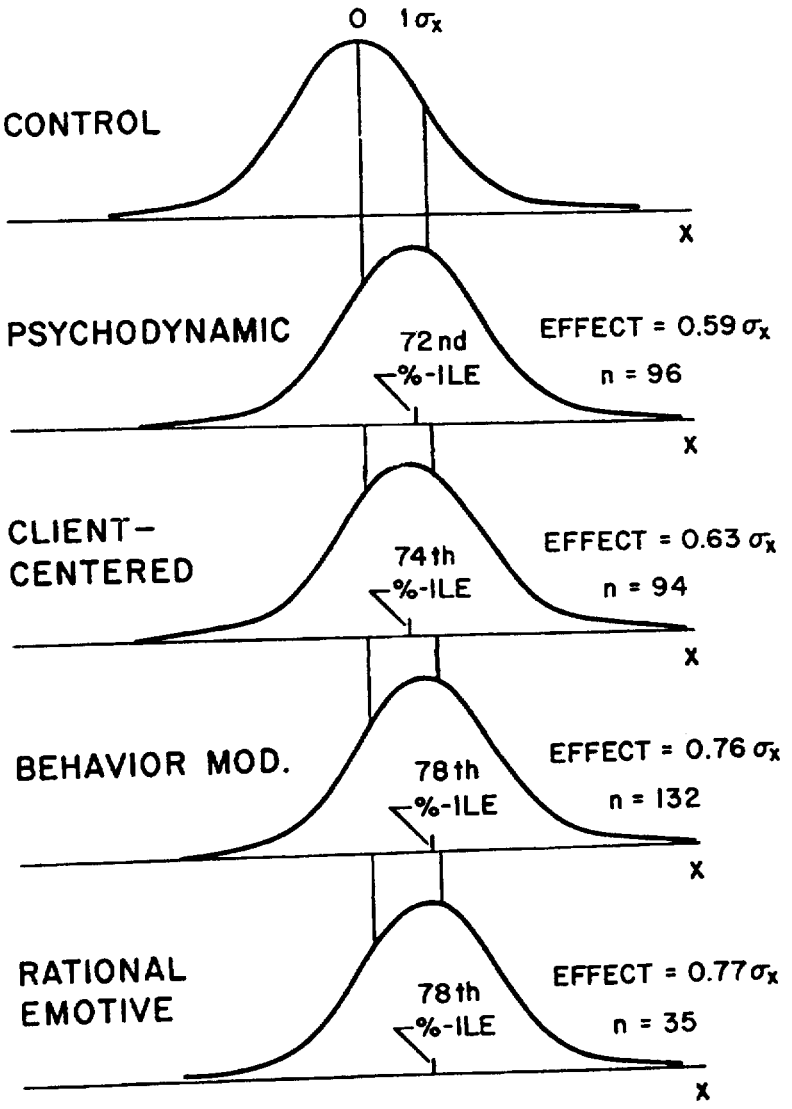

Figure 2. *Normal curves illustrating the effects of four types of psychotherapy in relation to untreated control groups.*

of television instruction. He classified each study as showing either statistically significant or non-significant results, and then crosstabulated this statistical significance factor with several properties of the study: age of the pupils, subject taught, etc. The relationships which Schramm found greatly clarify this large body of literature.

Other researchers have attempted similar integrations of literature on a slightly smaller scale. Bracht (1970) studied one hundred and eight experiments bearing on the aptitude-treatment interaction problem; Dunkin and Biddle (1974) made an admirable attempt to integrate hundreds of studies of teaching methods and behaviors; Cecil Clark (1971) organized the concept learning literature; Jamison, Suppes, and Wells (1974) reviewed studies on the evaluation of instructional media, including television, radio, and programmed or computer-as-

sisted instruction. Recently, Gregg Jackson (1975) put together many research studies on the effects of retention-in-grade and tried to tease the policy implications out of the findings. Sudman and Bradburn (1974) integrated the voluminous literature on response biases in surveys, examining some 900 studies and eventually synthesizing findings from over 300 of them.

These are praiseworthy attempts to cope with large and perplexing bodies of literature. But the methodologies we have applied have been too weak for the complexity of the problem. Measurement of the outcomes of the studies have typically been dichotomous: statistically significant vs. non-significant. Few properties of the studies have been related to outcomes, and examination of relationships has made use of simple two-factor crosstabulations instead of more versatile multivariate techniques. The methodology in widest use is the *voting method* of tabulating study results. Statistically significant vs. non-significant findings are classified by one, or perhaps two, attributes of the studies. There is little to recommend this attack on the problem. It is biased in favor of large-sample studies that may show only weak findings. It is not suited to the task of answering the important questions of *how large* an effect a particular treatment produces, or among several effective treatments which is most effective. The "vote taking" approach frequently produces only perplexing results.

In an AERA symposium on secondary analysis, Marshall Smith remarked that we don't need more reviews of this type. They undermine confidence in the field, not because the field has failed to establish reliable findings, but because the typical research review is pursued with insufficient ingenuity to reveal what has been found. The typical reviewer concludes that the research is in horrible shape; sometimes one gets results, sometimes one doesn't. Then the call is sounded for better research designs, better measures, better statistical methods—in short, a plaintive wish that things were not so complicated as they are. But the perfect, definitive study is a will o' the wisp. No educational problem admits a single answer. We study a system and a process too complex and too interactive to give up its secrets so easily. The variance in our findings of studies is essential, largely irreducible. It should be viewed as something to be studied in its own right, not something that can be eliminated with "tighter" designs or sharper measures. When a hundred studies can be arranged to yield a half-dozen answers to a question, we can feel confident that we are nearer the truth. Most of us were trained to analyze complex relationships among variables in the primary analysis of research data. But at the higher level, where variance, nonuniformity and uncertainty are no less evident, we too often substitute literary exposition for quantitative rigor. The proper integration of research requires the same statistical methods that are applied in primary data analysis. To coin the term "meta-analysis"—in a profession where cute neologisms abound—is not to pretend to any great insight or discovery, but rather to draw attention to the need to think differently than we do about reviewing and integrating research. It can be productively thought of as a problem in data analysis.

Figure 3. *Normal curves illustrating the comparative effects of behavioral and non-behavioral psychotherapies; effect sizes derived only from studies in which the two therapy types were simultaneously compared to a control group.*

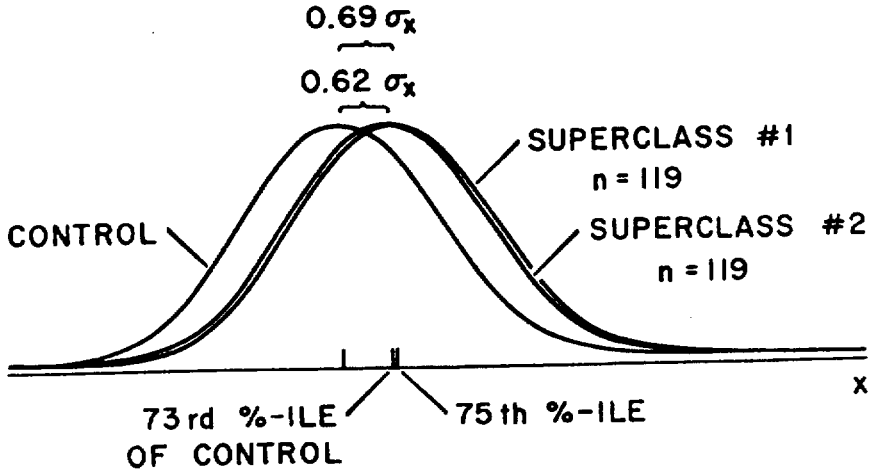

Illustration of Meta-Analysis on Psychotherapy Outcome Research

To illustrate meta-analysis on an important question, my colleague, Dr. Mary Lee Smith, and I set out to integrate the outcome evaluation literature in psychotherapy and counseling. Through an extensive literature search, nearly 400 controlled evaluations of the effects of psychotherapy were found. Each study was described in quantitative, or quasi-quantitative terms in several ways. Most important was the "effect size" of therapy: the mean difference on the outcome variable between treated and untreated subjects divided by the within group standard deviation. Thus, a study could be described as showing a .5, or .75, or -.25 standard deviation effect of therapy. Since some studies measured outcomes on more than one variable or at more than one time, the number of effect size measures exceeds the number of studies. In all, there were over 800 measures of effect size from the 375 studies. Other features of the study described by variables included the duration of the therapy (hours), the experience of the therapist (years), the diagnosis of the subjects (neurotic, psychotic), the type of therapy (Freudian, behavior modification, client-centered, etc.); the organization of therapy (individual, group), and the type of outcome measured (anxiety, self-concept, school/work achievement, physiological stress, etc.). The properties and findings of

the studies thus quantified, their aggregate meanings could then be sought through numerous statistical analyses. The findings of this meta-analysis are varied and multiple, and will be presented in detail elsewhere. Some of the more general and pertinent findings can be summarized briefly, however.

Table 1
Average Correlation Between SES and Achievement For Different Kinds of SES Measure*

SES Measure Consists of Indicators of/	Average r_{xy}	SES Measure Consists of Indicators of	Average r_{xy}
Income (only)	.315 (19)	Income & Education	.230 (36)
Education (only)	.185 (116)	Income & Occupation	.332 (15)
Occupation (only)	.201 (65)	Education & Occupation	.328 (20)
		All Three	.318 (27)

*Number of coefficients averaged in parentheses.

At the most general level, the findings can be summarized as in Figure 1. The average of the over 800 effect size measures is 0.68 σ_x, i.e., on the average, the therapy group mean was about two-thirds standard deviation above the control group mean on the outcome variable. The 375 studies with 40,000 treated and untreated subjects averaged slightly under twenty hours therapy, by therapists with two-and-a-half years experience, and outcomes were measured about four months after therapy. Thus, therapy of any type under these average conditions can be expected to move the typical client from the 50th to the 75th percentile of the untreated population.

The effect sizes of four different types of therapy are compared in Figure 2. The five normal curves represent the typical four treated populations in relation to the untreated control subjects. For example, about 100 effect size measures from outcome evaluations of psychodynamic therapies (loosely based on psychoanalytic principles) averaged about six-tenths standard deviation. The major impression one forms from Figure 2 is that the four types of therapy are not greatly different in their average impact.

The comparisons in Figure 2 were made without attempting to control for differences among therapies in the types of problem dealt with, the experience of the therapists, the duration of the therapy, etc. The best control of these additional factors is achieved by sorting out for special analysis all studies in which the conditions are strictly comparable for the different therapies. Such subclassification was possible when the ten types of therapy we distinguished were aggregated into two "superclasses" of therapy, viz., behavioral therapies (Superclass #1) and non-behavioral therapies (Superclass #2). To achieve a controlled comparison of the relative effectiveness of these two superclasses, all those studies were sought in which a behavioral and a non-behavioral therapy were simultaneously compared to a control group. Thus, for these studies, the effects of behavioral and nonbehavioral therapies could be compared where there was equivalence of type of complaint, duration of therapy, type of outcome measure, etc. The results are depicted in Figure 3. The findings are startling. There is only a trivial .07 σ_x superiority of the behavioral over the non-behavioral therapies. For all the superiority claimed by one camp or the other, for all the attention lovingly squandered on this style of therapy versus that style, the available evidence shows essentially no difference in the average impact of each class of therapy. What might one find in a similar comparison of the effects of "open" and "traditional" schooling?

Illustration of Meta-Analysis on SES and Achievement Correlation

In studying the relationship between socioeconomic status and school achievement, White (1976) collected over 600 correlation coefficients from published and unpublished literature. He subjected the coefficients to extensive analyses to determine how their magnitude was related to varying definitions of SES, different types of achievement, age of the subjects, etc. White found that the 636 available correlations of SES and achievement averaged .25 with a standard deviation of about .20 and positive skew; this means SES and achievement correlation is below what is generally believed to be the strength of association of the two variables.

White's meta-analysis of the SES and achievement correlation revealed many interesting trends. The correlation diminished as students got older, r decreasing from about .25 at the primary grades to around .15 late in high school. SES correlated higher with verbal than math achievement (.24 vs. .19 for 174 and 128 coefficients, respectively). When White classified the SES and achievement correlations by the type of SES measure employed, the patterns of Table 1 emerged. SES measured as income correlated more highly with achievement than either SES measured by the education of the parents or the occupational level of the head of household. Combining SES indicators resulted in higher correlations. White found several reliable trends in the collection of 600 coefficients that should help methodologists designing studies and sociologists constructing models of the schooling-social system.

Problems Needing Meta-Analysis

Several problems in education hold promise for statistical meta-analysis. The reading research literature has been abstracted systematically for over 50 years, but integrations of this huge literature are rare. A meta-analysis would be indicated on such questions as the relative effectiveness of phonics and "look-say" teaching approaches or of the effectiveness of different teaching orthographies. Reading is a prime area for research integration because of the standardization of outcome measures: reading speed, comprehension, and attitude constitute the basic outcomes. The research literature on class size and its

relationship to achievement is huge and has never been rigorously integrated. Likewise, the literatures on programmed instruction and instructional television—though a bit dated—are sizeable and, although Schram tallied the significance tests, the studies have not been thoroughly exploited. On more topical issues, research abounds in unpublished reports on the effects of school integration, the effectiveness of computer assisted instruction, the cognitive and affective outcomes of modern math curricula, and dozens of other interesting topics.

Conclusion

As educational researchers, we find ourselves in the mildly embarrassing position of knowing less than we have proven. The proofs reside in a vast literature that is often superciliously scorned and insufficiently respected. Extracting knowledge from accumulated studies is a complex and important methodological problem to which I commend your attention.

Note

[1]The research reported herin was supported by a grant from the Spencer Foundation, Chicago, Illinois. The author and Dr. Mary Lee Smith are co-principal investigators. This paper is the text of the author's presidential address to the 1976 Annual Meeting of the American Educational Research Association, San Francisco, April 21, 1976. The address as presented included additional findings of meta-analysis of psychotherapy outcome research which will be incorporated into another publication.

References

Astin, A. and Ross, S. Glutamic acid and human intelligence. *Psychological Bulletin*, 1960, *57*, 429-434.

Bracht, G.H. Experimental factors related to aptitude treatment interactions. *Review of Educational Research*, 1970, *40*, 627-645.

Campbell, D.T. and Erlebacher, A.E. "How regression artifacts in quasi-experimental evaluations can mistakenly make compensatory education look harmful." In *Compensatory education: A national debate*. Hellmuth, J. (Ed.). The Disadvantaged Child, Vol. 3, New York: Brunner/Mazel, 1970.

Clark, D.C. Teaching concepts in the classroom: A set of teaching prescriptions derived from experimental research. *Journal of Educational Psychology*, 1971, *62*, 253-278.

Cook, T.D. The potential and limitations of secondary evaluations. Chapter 6, pp. 155-234 in Apple, M.W., Subkoviak, H.S. and Lufler, J.R. (Eds.), *Educational evaluation: Analysis and responsibility*. Berkeley: McCutchan, 1974.

Dunkin, M. and Biddle, B. *The study of teaching*. New York: Holt, Rinehart and Winston, 1974.

Elashoff, J.D. and Snow, R.E. (Eds.) *Pygmalion reconsidered*. Worthington, Ohio: Charles A. Jones, 1971.

Jackson, G.B. The research evidence on the effects of grade retention. *Review of Educational Research*, 1975, *45*, 613-635.

Jamison, D., Suppes, P., and Wells, S. The effectiveness of alternative instructional media: A survey. *Review of Educational Research*, 1974, *44*, 1-67.

Light, R.J. and Smith, P.V. Accumulating evidence: Procedures for resolving contradictions among different research studies. *Harvard Educational Review*, 1971, *41*, 429-471.

Mosteller, F.M. and Moynihan, D.P. (Eds.) *On equality of educational opportunity*. New York: Vintage Books, 1972.

Schramm, W. Learning from instructional television. *Review of Educational Research*, 1962, *32*, 156-167.

Sudman, S. and Bradburn, N.M. *Response effects in surveys: A review and synthesis*. Chicago: Aldine Publishing Co., 1974.

White, K.R. *The relationship between socioeconomic status and academic achievement*. PhD thesis, University of Colorado, 1976.