RESEARCH ON AUDIO-TUTORIAL INSTRUCTION: A Meta-Analysis of Comparative Studies

James A. Kulik, Chen-Lin C. Kulik, and Peter A. Cohen, *The University of Michigan*

The present article describes a statistical synthesis of results from 48 comparative studies of an innovative method of college teaching, Postlethwait's Audio-Tutorial or A-T approach. The analysis showed that in general A-T instruction has a significant but small overall effect on student achievement in college courses, and it has no significant effect on student course evaluations or on course completions. Findings were similar for well-designed and less-well-designed studies included in this analysis, and they were also similar for studies carried out at different types of schools and in different subject areas. Results reported in journals, however, were more favorable to A-T than results found in dissertations and other unpublished reports.

Key words: meta-analysis; autoinstructional programs; audio-tutorial instruction

The educational literature of recent years reflects a growing interest in individualized instruction—teaching that takes into account the different backgrounds and aptitudes of learners. During the 1940s and 1950s, for example, the Education Index listed only about four or five articles each year on individualized instruction; the average number jumped to about 35 per year during the 1960s; and in the 1970s, well over 100 articles have been appearing each year (Kozak, 1974). Today, many educators are confident that individualized teaching will play an even greater role in education in the future. K. Patricia Cross (1976), for example, has predicted an instructional revolution in which our society will move beyond its goal of education for all to an ideal of education for each—or individualized instruction.

Among the systems of individualized instruction developed for col-

Address reprint requests to Dr. James A. Kulik, Center for Research on Learning and Teaching, The University of Michigan, 109 East Madison Street, Ann Arbor, MI 48109.

Research in Higher Education Vol. 11, No. 4, 1979
© 1979 Agathon Press, Inc. 0361-0365/79/080321-21$01.50

lege classrooms, few have received as much attention as Postlethwait's audio-tutorial or A-T approach (Postlethwait, Novak, & Murray, 1972). The A-T method dates back to 1961, when biologist Samuel Postlethwait began developing audiotapes and other visual and manipulative materials for remedial instruction in his introductory botany course at Purdue University. When Postlethwait's initial efforts proved successful, he decided to convert his entire course to an audio-tutorial approach. The revised course had three major components: independent study sessions, in which students learned from audiotapes and other media in self-instructional carrels; general assembly sessions, held each week, which were used for guest lectures, long films, and major examinations; and weekly integrated quiz sessions, which were held for groups consisting of between six and ten students and an instructor. In 1969, Postlethwait and his colleagues developed an additional approach to audio-tutorial instruction, mini-courses, or self-contained instructional units that provided even greater individualization of the amount, nature, and sequencing of instruction.

In the years since Postlethwait's initial work, many educators adopted his methods. Their hope, like Postlethwait's, was that a multisensory approach would meet the educational needs of students of low and average skill, while helping brighter students move more quickly through college courses. Some teachers replaced conventional laboratories or recitation sections with audio-tutorial laboratories; some devised total audio-tutorial courses that included general assembly, integrated quiz, and independent study sessions; and some used mini-courses. Whatever the approach, audio-tutorial instruction gained a firm foothold in higher education. Today, A-T courses are offered at junior and senior colleges, in the humanities and sciences, and in large and small institutions.

Postlethwait and his colleagues were concerned about evaluation of their innovations from the start. In 1962 Postlethwait set up an experimental section, which received all instruction by programmed audiotape, and he required students in this section to take the same semester examinations given to the conventionally taught group. On the semester examination, Postlethwait reported, this experimental section did just as well as the conventional group, but no better. Many of the teachers who adopted Postlethwait's teaching method have also followed his lead and compared results in their A-T classes with results of conventional instruction.

Three reviews summarized results of these comparisons of A-T and conventional instruction. In a 1975 review of research on A-T instruction, Mintzes described the results of six comparative studies. These studies reported on student achievement in A-T and conventional courses offered at the college level. Fisher and MacWhinney (1976) reported findings from 44 comparisons of student achievement in A-T and conventional courses. Fisher and MacWhinney's review covered a more diverse group of studies than Mintzes' survey. Included in Fisher and MacWhinney's survey were studies of A-T and of other audiovisual methods; studies at the college, high school, and elementary school level; studies using final examination scores as a criterion of achievement, as well as studies using instructor-assigned grades, weekly quiz scores, and a combination of these criteria; and studies varying in experimental adequacy. Finally, in a survey of studies of instructional technology in college teaching, Kulik and Jaksa (1977) described findings in 24 comparative studies of audio-tutorial instruction at the college level.

These reviewers reached different conclusions about the effectiveness of A-T instruction. Mintzes (1975), for example, suggested that the results of comparative studies appeared to be inconclusive and even contradictory. He reported that three of the six studies he reviewed favored audio-tutorial instruction, two studies found no differences between instructional methods, and one study favored conventional instruction. Fisher and MacWhinney's (1976) conclusions were highly favorable to A-T. Eighteen of the 44 studies they located found significantly higher student achievement in sections using audiovisual techniques; 25 studies reported no significant differences; and one study found a significant difference favoring the lecture method. These authors also reported that the affective response toward A-T instruction was favorable. Kulik and Jaksa (1977) reported that nine studies reported significantly higher final examination scores in the A-T sections, two studies found conventional instruction clearly superior, and 13 studies reported no significant differences in achievement in the comparison groups. They concluded that although A-T sometimes led to improved student learning, on the average the improvements it produced were small ones.

Each of these reviews used "box scores" to summarize the results of comparative studies. Although this approach provides a general overview of an area, it has a number of limitations (Glass, 1976). First, although a box-score tells how often one approach is better or worse than another, it does not tell how much better or worse. A box score may show that an innovative method beats a traditional method in 25 to 30 studies, but, in Glass's words, it does not say whether it wins "by a nose or in a walkaway." Second, box scores are of little help to the investigator trying to find out which characteristics distinguish studies that produce substantial effects from those with negligible results. The studies tallied in a box score all have distinctive characteristics, and results usually vary in confusing ways from study to study. To try to gather the pattern of findings from a box score is like trying to grasp the sense of hundreds of test scores without using statistical techniques to organize, depict, and interpret the data.

Glass (1976) proposed the method of "meta-analysis" as an alternative to this box-score type of review. Meta-analysis is simply the analysis of analyses or, more formally, the statistical analysis of a large collection of results from individual studies for the purpose of integrated findings. Researchers who carry out meta-analyses first locate studies of an issue by clearly specified procedures. They then characterize the outcomes and features of these studies in quantitative or quasi-quantitative terms. Finally, meta-analysts use multivariate techniques to describe findings and relate characteristics of the studies to outcomes.

In his presidential address to the American Educational Research Association, Glass (1976) described the application of this method to outcome research on psychotherapy and counseling. In the years since then, a number of other researchers have used his method to synthesize results of psychological and social research. Recent reports of the use of meta-analysis examined effects in the following areas: elementary school science curricula (Bredderman, 1979); class size and achievement (Glass & Smith, 1978); home and school environment and school learning (Haertel & Walberg, 1979; Iverson & Walberg, 1979; Uguroglu & Walberg, 1979); gender differences in nonverbal communication (Hall, 1978); individualized instruction in mathematics (Hartley, 1977); television and social behavior (Hearold, 1979); personalized and computer-based college teaching (Kulik, Kulik, & Cohen, 1979a, 1979b); advance organizers (Luiten, Ames, & Ackerson, 1979); drug therapy for psychological disorders (Miller, 1979); open vs. traditional education (Peterson, 1979); experimenter effects (Rosenthal, 1976); and socioeconomic status and academic achievement (White, 1979).

This article presents a further example of the use of meta-analysis of research findings. It is designed to answer the types of questions commonly asked by meta-analysts. How effective does this innovative method prove to be in the typical comparative study? Is it especially effective for certain types of outcomes or certain types of students? Under which conditions does it appear to be most effective? This article also focuses on several questions that have not been addressed in other meta-analyses. How well do different measures of effect size agree when applied to the same data? Do studies that show strong effects for one type of outcome tend to show strong effects for other outcomes? And, finally, can meta-analysis resolve differences in conclusions reached by traditional reviewers of research?

METHOD AND PROCEDURES

Data collection for the present meta-analysis began with a systematic computer search of three library data bases: *Psychological Abstracts, Comprehensive Dissertation Abstracts,* and the data base of educational materials from the Educational Resources Information Center. We added further research reports to our list by branching from bibliographies located in the original search. Finally, we located a few additional reports in recent issues of major disciplinary and interdisciplinary journals that feature work on audio-tutorial instruction.

In all, we located results from 48 studies of A-T instruction and conventional teaching, described in 47 different reports. To be included in our sample, studies had to satisfy three criteria. First, a study had to take place in an actual college course. We did not include laboratory analogues of college teaching in our sample. Second, the duration of the study had to be reasonably long—i.e., more than an hour or two of A-T in a one-semester course. Third, the study had to be free from obviously crippling methodological flaws—e.g., treatment groups that clearly differed in aptitude or a criterion test that was unfairly "taught" to one of the comparison groups.

In addition, we established guidelines that maximized independence among studies and that ensured that the same studies were not counted twice in the analysis. When several papers reported the same comparison, we used the most complete report for our analysis. When the same comparison was carried out in the same course at the same institution for one or more terms, we used the data from the most recent term. When an instructional outcome was measured on several instruments in a single paper, we pooled the results from the instruments to obtain a composite measure.

Study outcomes

The next step in the meta-analysis was to express outcomes of each study in quantitative terms. First, we described the effect of A-T on achievement as measured on a final or major examination. As our index of achievement effect, we used the average examination score (expressed as a percentage) in an A-T class minus the average in the comparable conventional class. Second, we measured the A-T effect on course completion. Our measure here was the difference in withdrawal rates for A-T and conventional classes, where withdrawal rate was the percentage of students initially enrolled who failed to complete a course in a term.

The third major outcome we examined was student course satisfaction. Quantifying the effect of A-T on course satisfaction measures presented some difficulties. To measure course satisfaction, most researchers examined degree of endorsement of items on course evaluation questionnaires. Different investigators, however, used different rating scales to obtain student reactions to instruction. One researcher might ask: ''On a 7-point scale, how would you rate the quality of this course? (1 = poor . . . 7 = excellent).'' Another might use the item: ''Overall, I consider this an excellent course (5 = strongly agree . . . 1 = strongly disagree).'' We had to decide when differently phrased items should be considered equivalent, and we had to convert ratings to a common metric. We first developed four lists of model rating items to cover four major aspects of instruction: overall quality, overall learning, overall enjoyment, and amount of work. We decided to include in our analysis results on any item that appeared in one of the lists. We finally converted all ratings to a 5-point scale, where 5 represented the highest rating (i.e., high quality, high enjoyment, much work, etc.) and 1 represented the lowest possible rating.

To make our study more comparable to earlier meta-analyses, we also calculated Cohen's (1969) and Glass's (1976) deviation-unit measures of effect size for these instructional outcomes. Cohen first introduced these ''pure'' measures of effect size about a decade ago, and since then they have become a basic tool in meta-analysis. In their analyses of interpersonal perception, for example, Rosenthal (1976) and Hall (1978) used Cohen's statistic *d,* defined as the difference between the means of the two groups being compared divided by the standard deviation common to the two populations. In their meta-analysis of outcomes of psychotherapy, Smith and Glass (1977) used a similar statistic, *ES,* the difference between experimental and control groups divided by the standard deviation of the control group.

To make our study more comparable to traditional reviews, we also examined the direction and significance of differences in outcomes of A-T and conventional teaching. On the basis of results, we classified each outcome as: (1) favoring conventional instruction and statistically significant; (2) favoring conventional instruction but not statistically significant; (3) favoring A-T but not statistically significant; and (4) favoring A-T and statistically significant.

Agreement among measures of effect size

How well do various measures of effect size agree? When applied to the same data set, do different measures produce the same results? Will findings of researchers who measure examination improvement in percentage points, for example, agree with results of those who use deviation-unit measures such as Cohen's *d* or Glass's *ES?* And how well do such continuous measures of effect size agree with a simple four-category scale based on direction and significance of differences?

A total of 42 of the 48 studies located for this meta-analysis contained data on examination performance. For 35 of the studies, we were able to calculate examination differences in percentage points between audio-tutorial and conventional classes; for 28 studies, we calculated Cohen's *d;* for 22 studies, we computed Glass's *ES;* and we classified each of the 42 studies into one of four categories reflecting the direction and statistical significance of the difference between A-T and conventional instruction.

Table 1 presents the intercorrelation matrix for these four indices of effect on achievement. It is obvious that the measures correlate very strongly when applied to the same data set. Two implications of these high intercorrelations are worth noting. First, it is possible to write regression equations that will predict with a high degree of accuracy from one kind of measure of effect to another. Such regression equations can be used to "plug" missing data on specific effect size measures. If, for example, a study does not report final examination averages in percentage terms but does report data from which Cohen's *d* can be calculated, Cohen's *d* can be used to predict with a high degree of accuracy the number of percentage points that separated experimental and control groups on a final examination. In the present analysis, we used this procedure to fill in missing observations in studies with an incomplete report of results. The second implication is that careful reviewers should report similar patterns of results for a given body of studies, even if the reviewers use different indices of effect size.

Agreement among different types of outcomes

If a study yields a large effect size for one type of outcome measure, is it likely also to show large effect sizes for other types of outcomes? This is not the same question as we answered above. There our con-

TABLE 1. Intercorrelations of Four Measures of Achievement Effect

	Differences in %-scores	Cohen's *d*	Glass's *ES*
Difference in %-scores			
Cohen's *d*	.94		
Glass's *ES*	.95	.99	
Four-category scale	.86	.84	.83

cern was a single type of outcome measured in different ways; here we were concerned with different types of outcomes measured in the same way. To answer the question, we correlated effect size measures for three different instructional outcomes—achievement, withdrawal, and student satisfaction. None of the intercorrelations was statistically significant. Achievement effect, for example, correlated –.33 with rating effect and –.32 with withdrawal effect. The implication of this result for the present analysis was clear. Calculating a single average effect size for all three types of outcomes would be a mistake. Achievement outcomes, attitudinal outcomes, and effects on course completion are different and separate matters.

OVERALL EFFECTS

One of the major goals in meta-analysis is to reach overall conclusions about the magnitude of overall effects. In this section of the article, we consider in turn the size of A-T effects on student achievement, course completion, and student ratings.

Student achievement

The A-T class performed at a higher level than the conventional class in 29 of the 42 studies of examination performance; the remaining 13 studies favored conventional instruction. A total of 15 of the 42 studies reported statistically significant differences between teaching methods. Eleven of these 15 studies favored A-T, and only four studies favored conventional instruction. If no overall generalization about the effect of A-T were possible, one would expect about half the cases to favor A-T, and half to favor conventional teaching. Instead, a majority favored A-T.

Continuous measures of effect size permit a more sensitive test of the influence of A-T on examination performance. The average difference between A-T and conventional class examination averages was 1.6 percentage points, and the standard deviation of this difference was 5.0. The average examination score was 68.5 in the typical A-T class; the average was 66.9 in the typical conventional class. It is statistically unlikely that a difference of this size would be found if there were no overall differences in effectiveness of A-T and conventional teaching. We were able, therefore, to reject the null hypothesis of no effect of A-T on student achievement.

Cohen's *d* for these data was .2, and Glass's *ES* was also .2. Thus, the effect of A-T in a typical class was to raise student achievement by about .2 standard deviation units. This implies that the typical student in an A-T class was performing at the 58th percentile on final examinations, whereas the typical control student achieved at the 50th percentile. Cohen described effects of this magnitude as small. When $d = .2$, treatment group membership accounts for only 1% of the variance in a trait in the population under study, and treatment effects are ordinarily too small to be observed without special measuring procedures. The difference in average height between 15- and 16-year-old girls, for example, is a difference of this magnitude.

Cohen contrasted small effects to medium and large ones. Medium size effects ($d = .5$) are large enough to be visible to the naked eye. One would typically notice, for example, the differences in height between 14- and 18-year-old girls. In our meta-analysis, a medium-size effect is a difference of 5 percentage points or more on a final examination. When $d = .8$, effects are large. In this analysis, large effects are differences of more than 8 percentage points on a final examination. Examples of large effect sizes from other areas are IQ differences between holders of the Ph.D. degree and typical college freshmen, or between college graduates and persons with only a 50-50 chance of passing in an academic curriculum.

Although the effect of A-T in the typical study was small, effect sizes varied from study to study. Figure 1 presents a distribution of effect sizes for the 42 studies. The figure shows that about 20% of these studies reported a medium or large effect in favor of A-T; about 70% of the studies found small or trivial effects; and about 10% of the studies reported moderate or large effects in favor of conventional instruction.

Course completion

Twenty-two of the 48 studies examined the effect of A-T on course completion. In nine studies withdrawal rate was higher in the A-T class; in ten studies it was higher in the conventional class; and in three studies withdrawal rate was identical in A-T and conventional sections. The difference in withdrawal rates was statistically significant in only four studies. In three of these cases withdrawal rate was significantly higher in the A-T class, and in one case it was significantly higher in the conventional class. Under a null hypothesis of no overall effect of instructional method on course completion, one would expect withdrawal rates to be higher in A-T classes about half the time. The obtained results do not differ significantly from this expectation.

With withdrawal rate treated as a continuous variable, we were able to perform a more sensitive test of the hypothesis of no difference in course withdrawals in the two kinds of classes. But even with this procedure we could not reject the hypothesis of no difference in course

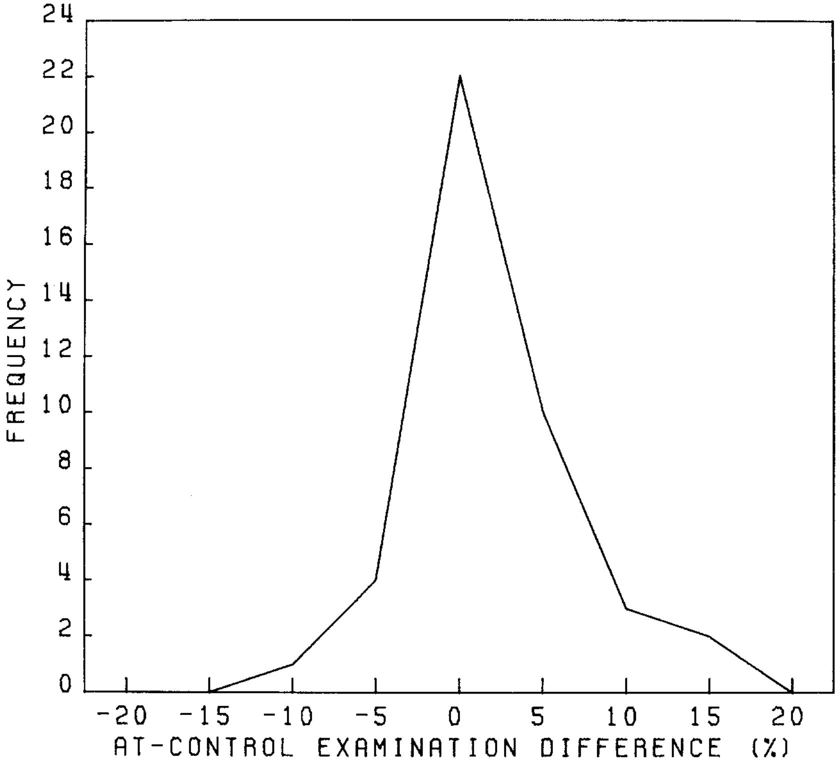

FIGURE 1. Distribution of differences in examination averages for 42 A-T and conventional classes (A-T = audio-tutorial instruction)

withdrawals as a function of teaching method. In the typical comparison, the withdrawal rate for the A-T class was about 2% higher than the rate in the conventional class. The average A-T withdrawal rate was 19%; the average rate in the conventional classes was 17%. The average effect size can also be expressed as *h* (Cohen, 1969), a measure of effect size that is comparable to Cohen's *d*. For these withdrawal data, *h* equaled 0.1, a trivial effect.

Like effects on achievement, A-T effects on course completion varied from study to study. Figure 2 presents a distribution of differences in withdrawal rates for the 22 studies. The figure shows that in the great majority of cases (in 19 [or 86%] of all studies) A-T effects were small or trivial. In one study, A-T reduced course withdrawals by a moderate amount, and in two studies A-T contributed at least to a moderate degree to an increase in student withdrawals.

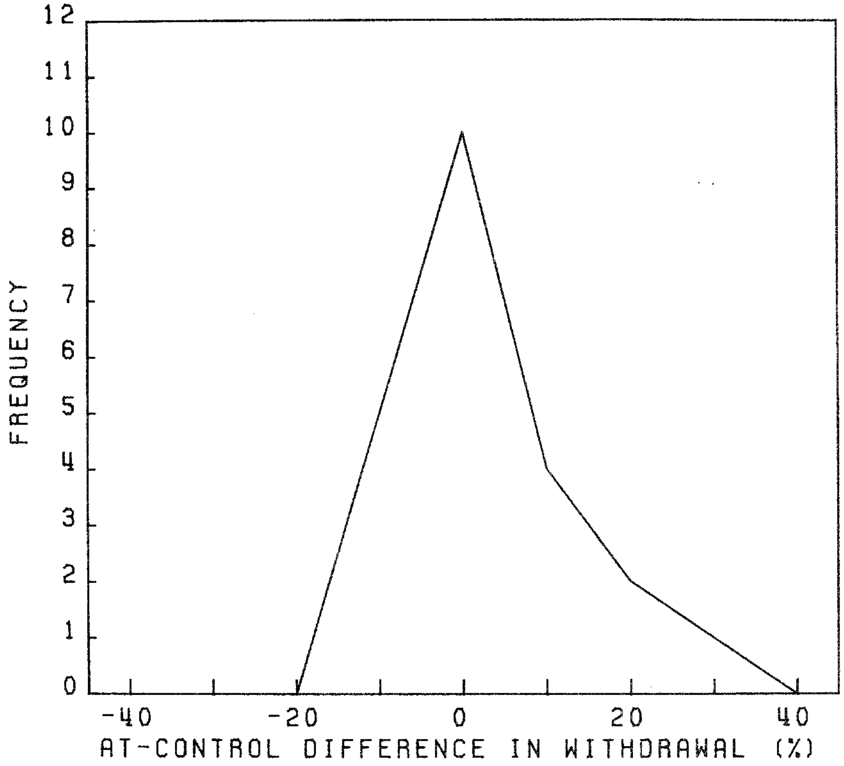

FIGURE 2. Distribution of differences in withdrawal rates for 22 A-T and conventional classes (A-T = audio-tutorial instruction)

Student ratings

Only six of the 48 studies located in our search of the literature contained student rating data. A-T ratings were higher than conventional ratings in three of the six studies that secured ratings of overall quality, and conventional ratings were higher in the remaining three cases. In most of these studies, the difference between A-T and conventional ratings was not large enough to be considered statistically reliable. Only one study showed a statistically reliable difference in favor of A-T instruction, and one study showed a statistically significant difference in favor of conventional teaching. These results provide no statistical support for the notion that students respond more favorably to A-T classes than they do to conventional classes.

Once again we used continuous measures of effect size for a more powerful test of the influence of A-T on student ratings. The average difference in quality rating was .26 on a five-point scale going from 5 (the highest rating) to 1 (the lowest rating). In the typical course, the A-T rating of course quality was 3.56, and the conventional rating was 3.30. This difference was small and insignificant. It corresponded to a Cohen's *d* of .12.

Ratings by students of how much they learned, how much they enjoyed the course, and how hard they worked were available from even fewer studies. Four studies provided ratings on amount learned, four on course enjoyment, and two on workload. Differences on these dimensions, like differences in ratings of course quality, were small and insignificant.

APTITUDE-TREATMENT INTERACTION

A number of the studies of A-T and conventional teaching were multifactorial studies in which an investigator examined the effects on achievement of both teaching method and student aptitude. These aptitude-treatment studies were carried out to determine whether aptitude plays as strong a role in A-T classes as in conventional classes. Bloom's (1968) mastery model suggests that there will be a lower correlation between aptitude and achievement in individualized classes than there is in conventional classes. Because individualized classes give students the time and instruction they need, the model suggests, both high- and low-aptitude students should reach high levels of performance.

A total of twelve studies reported aptitude-achievement correlations separately for A-T and conventional classes. Student aptitude was measured in somewhat different ways in these studies. In five of the studies, a standardized measure provided aptitude scores. Standardized measures used included the Scholastic Aptitude Test, the Natural Science test of the American College Testing program, and the Nelson Biology Test. Six other studies measured aptitude on an instructor-prepared pre-test. And finally, in one study, student grade-point average at entry to a course provided the aptitude measure.

We used meta-analysis to organize the results of these twelve studies. In five of the studies, the correlation between aptitude and achievement was higher in the A-T section, and in seven cases, it was higher in the conventional section. None of the studies reported a significant difference in aptitude-achievement correlations. Finally, the average correlation coefficient in the A-T classes was .36, and the average correlation in the conventional section was very similar, .39. All these data are consistent with a null hypothesis of no effect of instructional method on aptitude-achievement correlations. Contrary to mastery model predictions, ability has as much influence on student achievement in A-T classes as it has in conventional classes.

STUDY CHARACTERISTICS AND EFFECT SIZES

The effects of A-T were clearer in some studies than in others. One of our major goals was to discover factors that might explain this variation in effects. We wanted to know whether studies that reported strong effects differed systematically from those which produced weak effects.

Stepwise multiple regression provided a tool for relating study characteristics to study outcomes. The dependent variables in the regression analysis were percentage differences in final examination scores and percentage differences in withdrawal rates. We would have liked to analyze relationships between study characteristics and effects on student ratings, but the number of studies with student rating data was too small for regression analysis.

The independent variables in the regression analysis came from a set of 14 variables used to describe characteristics of the studies in our sample. Five of these variables described *methodological features* of the studies. These variables covered both internal and external threats to validity (Campbell & Stanley, 1963; Bracht & Glass, 1968), and included method of subject assignment, instructor effects, historical effects, bias in scoring the criterion, and bias in constructing the criterion. Eight other variables described *ecological* conditions under which innovative and conventional instruction were compared. These conditions included the character of the experimental treatment, the duration of time in which the innovative approach was used, the subject matter of the class, the grade level, and so on. The final variable described a *publication feature* of the study—whether the comparison was reported in a published article or in an unpublished paper or dissertation. Two of us independently coded each study on each variable. After discussing any disagreements between raters, we made final decisions about the placement of each study on each variable. The 14 variables, the coding categories for each, and the number of studies in each category appear in Table 2.

As a first step in the regression analysis, we examined the characteristics of the 42 studies with achievement data. It was immediately clear that several study characteristics could not explain the variation in student achievement outcomes since there was little variation on these characteristics in the achievement studies. Almost all the achievement studies controlled for possible historical effects and for bias in scoring the criterion; almost all studies involved semester-long

TABLE 2. Categories for Describing Studies and Number of Studies in Each Category

Coding Categories	Number of Comparisons
Methodological features	
Random assignment of comparison groups	
1. No	25
2. Yes	23
Control for instructor effect	
1. Different instructors	23
2. Same instructor	21
Control for historical effect	
1. Different semesters	7
2. Same semester	40
Control for scoring bias in criterion	
1. Non-objective test	3
2. Objective test	31
Control for author bias in criterion	
1. Institution developed test	21
2. Commercial standardized test	10
Ecological conditions	
Character of treatment	
1. Modified A-T of Postlethwait	25
2. Original A-T of Postlethwait	23
Requirement for mastery	
1. No	45
2. Yes	3
Duration of treatment	
1. Fraction of a semester	5
2. Whole semester	43
Content emphasis on "hard" science	
1. "Soft" science	8
2. "Hard" science	40
Content emphasis on "pure" knowledge	
1. Applied	12
2. Pure	36
Content emphasis on "life" studies	
1. Non-life	16
2. Life	32
Course level	
1. Introductory	42
2. Other	6
University setting	
1. Comprehensive, liberal arts or community college	27
2. Doctorate-granting institution	20

Table 2 ***(Continued)***

Coding Categories	Number of Comparisons
Publication features	
Source of study	
1. Unpublished	31
2. Published	17

comparisons; and almost all lacked an explicit mastery requirement. Because of lack of variation in these four variables—historical control, scoring bias, duration, and mastery requirement—they could not have contributed systematically to variation in study outcomes, and they were left out of further analyses.

Only one of the ten independent variables included in the regression analysis correlated significantly with achievement effect size. This variable was manner of publication, and even its correlation with achievement effect size, .31, was modest. This correlation, however, indicated that studies published in journals reported a more favorable effect of A-T than studies in dissertations and unpublished papers. In a typical journal article, the A-T examination average was 3.8 percentage points higher than the examination average in the conventional class; in the typical study reported in a dissertation or unpublished paper, the achievement difference was 0.6 percentage points.

Other study features were less highly related to effect size. Correlations between effect size and the remaining nine variables were low in magnitude, and none could be considered significantly different from zero. To investigate the possibility that a combination of variables might predict effect sizes more accurately than a single predictor, we also carried out a stepwise multiple regression analysis with generous limits for inclusion of predictor variables. Results of the analysis were clear-cut. Once publication history was taken into account, none of the variables was significantly related to effect size.

A second regression analysis examined the relationship between study characteristics and withdrawal outcomes. Again, we first examined the characteristics of the 22 studies with withdrawal data. Some factors could be eliminated immediately as possible causes of variation in withdrawal results from study to study. All of the comparisons reporting withdrawal rates lasted one whole semester, and all courses reporting withdrawal rates were introductory courses. Almost all the courses with withdrawal data lacked an explicit mastery requirement. We therefore eliminated these three study characteristics from regression analysis. The results of the regression analyses were straightforward. None of the 11 study characteristics was significantly related to size of withdrawal effect. We were unable therefore to find either a single variable or a combination of variables that would distinguish between studies showing strong and weak A-T effects on withdrawal rate.

DISCUSSION AND CONCLUSIONS

Meta-analysis is still a new method for synthesizing research results, and meta-analysts do not yet agree on all aspects of its use. New applications of the method, therefore, still contribute to our knowledge of its potential and its limits. The present application contributes at least two important points to meta-analytic methodology.

First, this analysis suggests that different measures of effect size relate strongly to one another when applied to the same data set. The correlation between Cohen's (1969) and Glass's (1976) deviation-unit measures of effect size, for example, was nearly unity. Correlations between these deviation-unit measures and our concrete measures of effect size were also very high. Even a simple four-point scale classifying studies according to direction and significance of differences correlated highly with other effect size indices. Such results do not imply that all measures of effect size are equally precise, but they do suggest that careful research syntheses will yield similar patterns of findings, no matter what statistical approach they use. We should probably be suspicious of meta-analyses which reach conclusions about experimental treatments very different from those provided by other types of research syntheses.

Second, our analysis suggests that results of experimental treatments are not unidimensional. A number of studies included in our analysis examined multiple outcomes of instruction, and the different types of outcomes were not significantly correlated in these studies. Studies which reported a strong A-T effect on course completion, for instance, were as likely to report a positive as a negative effect on student achievement. Studies reporting a negative impact on student ratings might report a negative or a positive impact on achievement. The implication of such findings seems clear enough to us. Meta-analysts should resist the temptation to calculate a single average effect size for all types of outcomes. The deviation-unit measures of effect size developed by Cohen (1969) and Glass (1976) allow meta-analysts to calculate a single average size for a wide variety of outcomes, and Glass's (1976) initial paper on meta-analysis provides a striking example of this practice. Our results suggest that meta-analysts may not yet be justified in reaching global conclusions like Smith and Glass's: "On the average, the typical therapy client is better off than 75% of untreated individuals" (1977, p. 304). Instead, meta-analysts may have to specify *in what respects* treated individuals are better off.

This analysis also produced a number of findings about audiotutorial instruction. First, we found a small but significant overall effect of A-T on student achievement. Second, A-T had little effect on withdrawal rate or on course evaluations. Withdrawal rates and student ratings were about the same in A-T and conventional classes. Third, aptitude and achievement were as highly correlated in A-T classes as they were in other courses. Contrary to mastery model predictions, A-T does not seem to reduce the influence of aptitude on student achievement. Finally, in our analysis, characteristics of studies were not strongly related to study outcomes. Findings were very similar for well-designed and less well-designed studies and for studies carried out at different types of schools and in different subject areas. Results reported in journals, however, were more favorable to A-T than results found in dissertations and unpublished studies.

Many researchers who compared A-T and conventional teaching concluded that this approach was at least as effective as conventional teaching. Our analysis supports this view. In its effect on student achievement, A-T is like a long list of alternatives to the lecture method—discussion, supervised self-study, instructional video, programmed instruction—that are roughly its equivalent in effectiveness (Dubin & Taveggia, 1968). Although this is a satisfactory record of effectiveness in some ways, it is a modest record in at least one respect. The other major method of individualized college teaching introduced during the 1960s—Keller's Personalized System of Instruction or PSI (Keller, 1968)—has produced much more dramatic results. In most comparative studies, PSI made a substantial contribution to examination performance and also contributed significantly to student ratings of course quality (Kulik, Kulik, & Cohen, 1979a). A-T's record of effectiveness is less impressive.

The failure to find strong support for the attribute-treatment interaction predicted by Bloom's mastery model of school learning was not unexpected. In our previous work on PSI, we also found that aptitude-achievement correlations were nearly identical in conventional and PSI classes. Individualized instruction, in which students are free to vary the time and manner of learning, does not seem to narrow the gap between gifted and disadvantaged learners. The finding that published articles present a more favorable view of innovative programs than unpublished articles was also predictable. Earlier meta-analyses also reported that results printed in journal articles differed somewhat from those found in unpublished papers (Smith & Glass, 1977; Hartley, 1977). In general, however, study features do not explain much of the variation in study outcomes. We also found this to be true in our meta-analysis of research on personalized instruction (Kulik, Kulik, & Cohen, 1979a).

Our final concern is the amount of agreement between conclusions drawn in our meta-analysis and conclusions from traditional reviews of audiotutorial instruction. In the earliest published review of A-T instruction, Mintzes (1975) drew the conclusion that results of comparative studies of A-T were contradictory and inconclusive. The pattern of results reported by Mintzes was very similar to our own pattern of findings, but Mintzes located too few studies to draw firm conclusions. Had he been able to locate more studies, Mintzes might have reached conclusions like our own.

Our conclusions are also consistent with those reached in an earlier review (Kulik & Jaksa, 1977). In a survey of 24 studies of A-T at the college level, Kulik and Jaksa concluded that A-T sometimes produced favorable effects but that these effects were small in size. They cited a 3% increment as a typical gain due to A-T. Their figure was based on the ratio of final examination average in A-T classes to final examination average in conventional classes. In the present study, we found an average difference of 1.6 percentage points between final examination score in A-T and conventional classes. The final examination average in A-T classes (68.5) divided by the average score in conventional classes (66.9) was equal to 102.4%, very similar to the figure reported by Kulik and Jaksa.

Fisher and MacWhinney's (1976) conclusions deserve more comment. First, these authors pointed out that results from a large heterogeneous group of studies, although favoring A-T, were somewhat mixed. Fisher and MacWhinney reported that 18 out of 44 studies favored A-T, one study favored conventional teaching, and 25 studies produced nonsignificant results. Our overall findings were similar. We reported that 11 out of 42 studies favored A-T, four studies favored conventional teaching, and 27 studies reported no significant differences between teaching methods. Our distribution does not differ significantly from that of Fisher and MacWhinney.

Fisher and MacWhinney, however, divided their total group of 44 studies into those with major design faults and those without design flaws. Results of the two groups of studies were strikingly different. Most of Fisher and MacWhinney's well-designed studies favored A-T, and almost all of those with design flaws found no significant differences between methods or favored conventional teaching. Since well-designed studies seemed almost uniformly in favor of A-T, Fisher and MacWhiney concluded that A-T instruction was very effective in improving student achievement. We found, on the other hand, no significant correlation between design characteristics of studies and study outcomes. In addition, we were unable to distinguish Fisher and MacWhinney's well-designed studies from poorly designed studies. We were unable, in fact, to include some of Fisher and MacWhinney's well-designed studies in our analysis since these studies seemed to us to contain major, crippling design flaws. Other studies that Fisher and MacWhinney classified as well-designed seemed at best distantly related to Postlethwait's A-T. On the other hand, some of the studies that Fisher and MacWhinney classified as poorly designed seemed good enough to include in our meta-analysis. Thus, our evaluation of studies differed strikingly from Fisher and MacWhinney's, and we differed, therefore, in our overall conclusions about A-T.

Finally, Fisher and MacWhinney reported that in almost every study they located students expressed strong positive attitudes toward A-T. Our analysis showed that students reacted similarly to A-T and conventional classes. Here, the discrepancy in conclusions may be easier to resolve. Fisher and MacWhinney drew their conclusions from data collected without the benefit of a control group. Our conclusion was based solely on comparative studies that used a control group. We found that students gave favorable ratings to their A-T classes *and* to their conventionally taught classes—another example of the tendency, well-known in studies of student ratings, for statistically average classes to be rated ''above average'' by students. In our judgment, currently available data suggest that A-T does not lead to higher or lower course ratings than conventional teaching methods.

ACKNOWLEDGMENTS

This study was supported by National Science Foundation Grant No. SED 77-18566.

A complete list of studies used in the analyses described in this article is available from the first author at the Center for Research on Learning and Teaching, 109 East Madison Street, Ann Arbor, Michigan 48109.

REFERENCES

Bloom, B. S. Learning for mastery. *Evaluation Comment,* 1968, *1*(2, Whole No. 2).

Bracht, G. H., & Glass, G. V. The external validity of experiments. *American Educational Research Journal,* 1968, *5,* 437-474.

Bredderman, T. A meta-analysis of elementary school science process curricula. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, April 1979.

Campbell, D. T., & Stanley, J. C. Experimental and quasi-experimental designs for research on teaching. In N. L. Gage (Ed.), *Handbook of research on teaching.* Chicago: Rand-McNally, 1963.

Cohen, J. *Statistical power analysis for the behavioral sciences.* New York: Academic Press, 1969.

Cross, K. P. *Accent on learning.* San Francisco: Jossey-Bass, 1976.

Dubin, R., & Taveggia, T. C. *The teaching-learning paradox.* Eugene: University of Oregon Press, 1968.

Fisher, K. M., & MacWhinney, B. AV Autotutorial instruction: A review of evaluative research. *Audio-Visual Communications Review,* 1976, *24,* 229-261.

Glass, G. V. Primary, secondary, and meta-analysis of research. *Educational Researcher,* 1976, *10,* 3-8.

Glass, G. V., & Smith, M. L. Meta-analysis of research on the relationship of class size and achievement. Boulder: Laboratory of Educational Research, University of Colorado, September 1978.

Haertel, G., & Walberg, H. J. Synthesis of research on achievement and social environment of the classroom. Symposium paper presented at the annual meeting of the American Educational Research Association, San Francisco, April 1979.

Hall, J. A. Gender effects in decoding nonverbal cues. *Psychological Bulletin,* 1978, *85,* 845-857.

Hartley, S. S. Meta-analysis of the effects of individually paced instruction in mathematics. Unpublished doctoral dissertation, University of Colorado, 1977.

Hearold, S. Meta-analysis of the effects of television on social behavior. Symposium paper presented at the annual meeting of the American Educational Research Association, San Francisco, April 1979.

Iverson, B. K., & Walberg, H. J. Integrating research findings: The relationship of home environment to school achievement. Symposium paper presented at the annual meeting of the American Education Research Association, San Francisco, April 1979.

Keller, F. S. "Good-bye, teacher . . ." *Journal of Applied Behavior Analysis,* 1968, *1,* 79-89.

Kozak, M. R. A critical analysis of individualized instruction since 1944. Unpublished doctoral dissertation, Texas A & M University, 1974.

Kulik, J. A., & Jaksa, P. PSI and other educational technologies in college teaching. *Educational Technology,* 1977, *17,* 12-19.

Kulik, J. A., Kulik, C.-L. C., & Cohen, P. A. A meta-analysis of outcome studies of Keller's Personalized System of Instruction. *American Psychologist,* 1979a, *34,* 307-318.

Kulik, J. A., Kulik, C.-L. C., & Cohen, P. A. Effectiveness of computer-based college teaching. Ann Arbor, Mich.: Center for Research on Learning and Teaching, The University of Michigan, October 1979b.

Luiten, J., Ames, W., & Ackerson, G. The advance organizer: A review of research using Glass' technique of meta-analysis. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, April 1979.

Miller, T. I. The effects of drug therapy on psychological disorders. Symposium paper presented at the annual meeting of the American Educational Research Association, April 1979.

Mintzes, J. J. The A-T approach 14 years later: A review of recent research. *Journal of College Science Teaching,* March 1975, 247-252.

Peterson, P. L. Direct instruction reconsidered. In P. L. Peterson & H. J. Walberg (Eds.), *Research on teaching.* Berkeley, Calif.: McCutchan Publishing Corp., 1979, pp. 57-69.

Postlethwait, S. N., Novak, J., & Murray H. T., Jr. *The audio-tutorial approach to learning.* Minneapolis: Burgess Publishing Co., 1972.

Rosenthal, R. *Experimenter effects in behavioral research.* New York: Irvington, 1976.

Smith, M. L., & Glass, G. V. Meta-analysis of psychotherapy outcome studies. *American Psychologist,* 1977, *32,* 752-760.

Uguroglu, M., & Walberg, H. J. Research synthesis: Motivation and school achievement. Symposium paper presented at the annual meeting of the American Educational Research Association, San Francisco, April 1979.

White, K. R. The relationship between socioeconomic status and academic achievement. Symposium paper presented at the annual meeting of the American Educational Research Association, San Francisco, April 1979.