Educational Evaluation and Policy Analysis
Vol. 1 No. 1 1979

Meta-Analysis of Research on Class Size and Achievement

GENE V GLASS
MARY LEE SMITH
Laboratory of Educational Research
University of Colorado

THERE IS NO POINT IN RECORDING THE obvious about class size: that teachers worry about it more than nearly anything else, that administrators want to increase it, that it is economically important, and the like. The problem with class size is the research. It is unclear. It has variously been read as supporting larger classes, supporting smaller classes, and supporting nothing but the need for better research. Review after review of the topic has dissolved into cynical despair or epistemological confusion. The notion is wide-spread among educators and researchers that class size bears no relationship to achievement. It is a dead issue in the minds of most instructional researchers. To return to the class-size literature in search of defensible interpretations and conclusions strikes many as fruitless. The endeavor is surrounded by a faint aroma of Chippendale, which it resembles in other respects: unwieldy and antique.

One could document the confusion in previous reviews of research on the class-size and achievement relationship. It would be simple to quote Reviewer X claiming that large classes are better, Reviewer Y to the effect that small classes are better, and Reviewer Z that neither is better. But to do so would only embarrass others and add nothing to one's appreciation of the complexity of the research. The problems with previous reviews of the class-size literature are several: (1) literature searches were haphazard and often overly selective; dissertations were avoided, as a rule, and few reviewers sought out large archives of pertinent data; (2) reviews were typically narrative and discursive; the multiplicity of findings cannot be absorbed without quantitative methods of reviewing; (3) reviewers that attempted quantitative integration of findings made several mistakes: They used crude classifications of class sizes; they took "statistical significance" of differences far too seriously; and they lacked sufficiently sophisticated techniques of integrating results.

In the research reported here, an attempt was made to correct these shortcomings and determine if the huge research literature on class size and achievement really was hopelessly confusing or if its message was merely buried in myriad results waiting to be coaxed out with more advanced methods of research integration.

THE LITERATURE SEARCH

The search for class-size studies was carried out in three places: (1) document retrieval and abstracting resources; (2) previous reviews of the class-size literature; and (3) the bibliographies of studies once

Based on a longer report: Glass, G. V. and Smith, M. L. *Meta-analysis of Research on the Relationship of Class size and Achievement,* produced under a grant (NO. OB-NIE-G-78-0103) from the National Institute of Education to the project "Class Size and Instruction" of the Far West Laboratory for Educational Research and Development, Dr. L. S. Cahen, Project Director.

We wish to acknowledge the continuing help and support of Drs. Cahen and Nikola N. Filby of the Far West Lab and Joseph Vaughan and Virginia Koehler of NIE. Some support for the work reported here was contributed through the Visiting Scholars Program of the Center for the Study of Evaluation, University of California, Los Angeles under NIE Grant No. OB-NIE-G-78-0213.

Available through Dr. Cahen, Far West Lab.

found. The ERIC system and *Dissertation Abstracts* were searched completely on the key words "size," "class size," and "tutoring." The dissertation literature was covered as far back as 1900, and the fugitive educational research literature was covered from the mid-1960s to 1978. Of the many hundreds of doctoral dissertations scanned in *Dissertation Abstracts,* about 30 microfilm copies were purchased. About a dozen of these dissertations were incorporated; the remainder dealt with nonachievement and process variables that will be covered in subsequent work. The journal literature on class size was located in the traditional way; one or two current reviews of the research were found—the Ryan and Greenfield (1975) review was particularly comprehensive and helpful—the articles cited were located, and the articles cited in these articles were located in turn.

Approximately 300 documents were obtained and read, and it was found that 150 of them contained no usable data, i.e., no data whatsoever were reported on the comparison of small and large-class achievement. About 70 studies examined the relationship of class size to nonachievement outcomes and classroom process variables. Approximately 80 studies on the class-size and achievement relationship were included in this analysis.

It is difficult to estimate what portion of the existing literature was captured by this search. Even though the corpus of 80 studies exceeds by 50% the most extensive reviews published to date—and these reviews are narrative and inconclusive—it is conceivable that less than half of all studies that exist on the topic were found. Some studies (credited to school districts) could not be located even after several phone calls and letters. Other studies were surely missed because of odd or nondescript titles. The dissertation search was conducted on key words such as "size," "class size," and "tutoring" but the words must appear in titles to be registered in the index to *Dissertation Abstracts.* (Fortunately, the ERIC system uses key words based on the contents of a paper and not titles alone.) Several studies found in the journal literature by branching off existing bibliographies had neither "size" nor "class size" in the title, evidence enough that several dissertations were missed because their titles lacked the key words. Still another complication concerns the use of class size as an incidental variable in studies focused on other issues. There are probably many such studies, and only a few of the most visible ones were located.

THE TEXTURE OF THE LITERATURE

In what follows in this integrative analysis, one can easily lose touch with precisely what kinds of research are being integrated. The statistics and graphs that represent the findings of this meta-analysis of class-size research will seem far removed from the original studies themselves. And, in a very real sense, what will be done for the sake of arriving at general conclusions places the reader in benign jeopardy of losing qualitative and personal familiarity with the research. In this section, the general texture of the class-size literature will be described, and a few studies typical of various eras will be reported.

The research on class size and its relationship to achievement falls into four stages: the pre-experimental era (1895–1920); the primitive experimental era (1920–1940); the large-group technology era (1950–1970); and the individualization era (1970-present). The boundaries of the eras are not impenetrable, and even today an atavistic throwback to the nineteenth century will appear in a doctoral thesis. At each new stage, the sophistication of research methodology increased, and the question of class size and its effect on achievement was examined with different motives. One discerns in the narration accompanying the numbers the cult of efficiency of the early part of this century, the rising birth rate of the post-war 1940s, the advent of teaching technology in the 1960s, and most recently the teacher labor movement combined with declining enrollments. What was said about the data changed as new interpretations served emerging purposes, even when the data changed little themselves.

The first empirical study on educational processes and their effects on achievement included an examination of the class-size question (Rice, 1902). No strong relationship of class size to attainment was observed. But unfortunately, Rice reported virtually no numbers, and it is impossible

to determine now whether the relationship Rice found was genuinely small or whether it was moderately large, but only seemed small to Rice, who may have expected much more. Rice's study was followed by several similar analyses on new data collected between 1900 and 1920. These studies are typified by their rugged nonexperimental logic. A study by Cornman (1909) can serve as an example.

Cornman examined the promotion records for January 1909, in District No. 6, Philadelphia. Before the days of "social promotion," the passage from one grade to the next higher indicated adequate achievement at the lower grade. Cornman categorized classes into three groups: under 40 pupils, 40 to 49, and 50 or more. The rate of promotion was calculated for each class-size category. At grade 3, 88% of 400 pupils in classes of 40 or fewer were promoted, 85% of 1,300 pupils in classes size 40 to 49 were promoted, and 81% of 640 pupils were promoted in classes of over 50 pupils. Cornman also investigated "satisfactory conduct" ratings by teachers in classes of different sizes. The discussion of results showed little sensitivity to questions of experimental control; such concerns were doubtless not wide-spread at the time.

Beginning in the early 1920s, the class-size and achievement question was approached with better methods. Studies began to appear that used matching of pupils in large and small classes on ability and achievement; content and methods were standardized in the two classes; occasionally the same teachers taught classes of both sizes. Tope, Groom and Beeson (1924) studied the relationship between class size and achievement in grammar and English at the high-school level in Grand Junction, Colorado. In the Fall of 1922, three English classes of 44, 34, and 20 pupils were formed. Their Terman Group Test IQs were nearly identical at the first, second, and third quartiles. "After thoroughly establishing our classes, our method of conducting the experiment was merely to proceed with the year's work in the usual way, except that we found it necessary to depend rather more than usual on test grades, because the number of pupils in the large class made it impossible for each pupil to make many daily recitations each period" (Tope et al., 1924, p. 127). The experiment was run for 9 weeks. Then the Starch Grammar Test and Kirby Grammar Test were administered along with some specially designed classroom tests on clauses. The findings slightly favored the two smaller classes over the class of 44.

In the 1940s, class-size research went dormant when educational researchers went to war. It was revived along with the rest of the field in the 1950s and 1960s. Researchers seemed intent on demonstrating, particularly at the college level, that lecture classes could be doubled or tripled in size without loss of effectiveness. At about the same time, massive empirical studies of education were undertaken to inform national education policy: the Coleman study of equality of educational opportunity (1966); Project TALENT; the International Assessment of Education in mathematics and reading; and surveys of government-funded programs of compensatory education (Title I). These large empirical studies typically included, as incidental features, data on the relationship of class size and achievement. The study by Nelson (1959) is representative of the first kind of study to appear in the 1950s and 1960s; the Coleman (1966) study is like many studies of the second type.

In 1959, Nelson reported on a study of large-group college instruction. Four instructors were involved, each teaching one large and one small section of elementary economics. The pupils in each instructor's classes were matched on major (e.g., business, engineering), level (freshman, sophomore), and sex. The course was taught 3 hours a week for a semester. The class-sizes compared were 20 vs. 138, 16 vs. 141, 20 vs. 94, 20 vs. 90, 17 vs. 109, 17 vs. 94, 19 vs. 85. A common final examination was administered to all 14 classes. Achievement outcomes were adjusted by covarying on students' prior grade-point average. The means favored the larger classes by three one-thousandths standard deviation!

The Coleman study is famous. Tens of thousands of pupils in grades 1, 3, 6, 9, and 12 were surveyed. Achievement tests were administered and "school resources" were measured at the level of the school, e.g., teachers' experience, use of special programs. Among these resource variables was pupil/instructor ratio. The P/I ratio was correlated with pupil achievement. The correlations were generally negative. When Mayeske (undated) partialed out three or four other variables which might

have obliterated these correlations, the r's remained consistently negative.

The research relevant to class size that appeared in the 1970s showed a concern for establishing the benefits of individualization. Experiments were performed that involved radically reduced instructional group sizes, one teacher with two or three pupils. Studies of individual pupils taught by computer or machine have also become common; they were not considered in this integrative analysis since the particular concern here is with the processes of human instruction. (For a meta-analysis of tutoring and computer-assisted instruction in mathematics that produced surprising findings, see Hartley, 1977.) An experiment typical of studies of radically reduced group size was conducted by Bausell, Moody, and Walze (1972). Pupils in grades 4 and 5 were randomly assigned to receive either individual tutoring on exponential arithmetic for 1 hour across 2 days or instruction by randomly comparable teachers for the same amount of time in a class of 25 pupils. Instruction was a part of an on-going school program. A test designed to cover only the content of the instruction was administered to all pupils. Pupils in "class-size 1" scored approximately one-half standard deviation above pupils in classes of 25 on the achievement tests.

METHODS

In this section, the methods are described by which the studies were coded and the quantitative findings integrated.

DEFINING THE FIELD

The problem of this meta-analysis is to determine what the available research proves about the relationship of class size to achievement. Drawing boundaries around this topic was simple compared to the difficultues encountered in defining psychotherapy, for example (Smith & Glass, 1977). Conventional definitions of "achievement" seem scarcely to have changed over 80 years; and "class size" is relatively easily described and quantified.

CODING CHARACTERISTICS OF STUDIES

The quantification of characteristics of studies permits the eventual statistical description of how properties of studies affect the principal findings. Such questions can be addressed as "How does the class size and achievement relationship vary as a function of age of pupils?" or "How does it vary between reading and math instruction?" The first step in coding studies is to identify those properties of studies that might interact with the relationship between class size and achievement. There is no systematic and logical procedure for taking this step. One simply reads a few studies from the literature of interest, talks with experts, and then makes a best guess; modifications can always be made later if needed. The best guesses as to which conditions might mediate the relationship fell into five broad categories: Study Identification, Instruction, Classroom Demographics, Study Conditions, and Outcome Variable. About 25 specific items fell into these categories. Some were more fruitful than others; several items were seldom reported in the research publications. A coding sheet was devised onto which the information about each study could be transcribed. A single study might fill several coding sheets, depending on how many different class sizes were compared in pairs, how many different achievement tests were reported, whether data were reported separately for different ages of IQs, and so forth.

The major items of the coding sheet are reported below:

IDENTIFICATION:

1) *Year.* This item was included to check on whether there is a time trend in the class size and achievement relationship.
2) *Source of Data.* Whether from a journal, book, thesis, or unpublished source.

INSTRUCTION:

3) *Subject.* The subject taught (reading, math, etc.) was recorded.
4) *Duration of Instruction.* The amount of teaching was recorded in hours and in weeks.
5) *No. of Pupils.* The numbers of pupils on which the small- and large-class achievement means were based were recorded. This number was not the same as the "class size" since there might be several small or large classes used in the study.
6) *No. of Instructional Groups.* (See #5 above.)
7) *No. of Instructors.* (See #5 above.)

8) *Pupil/Instructor Ratio.* This measure is the measure of class size. One teacher with a group of 30 counts as a P/I ratio of 30; two teachers in a class of 30 gives a P/I of 15.

CLASSROOM DEMOGRAPHICS:

9) *Pupil Ability.* Average IQ of the pupils was estimated when not reported; three broad categories were used: IQ $\leq$ 90; 90 < IQ < 110; IQ $\geq$ 110.

10) *Ages and Average Age.* These two variables permitted discriminating instances in which all pupils were of one age from studies in which pupils of several ages were represented and the average age was used to describe their level since data were not reported separately. This variable was used to distinguish data from elementary and secondary school levels.

STUDY CONDITIONS:

11) *Assignment of Pupils and Teachers to Groups.* The assignment of pupils and teachers to classes of different sizes was described as either "random," "matched," "repeated measures," or "uncontrolled." These variables were important in describing the degree of experimental control exercised in the study. "Random" is obvious; "matched" refers to attempts to equate small and large classes by other than random means on pretests of achievement or ability; "repeated measures" refers to using either the same pupils or teacher in both small and large classes, e.g., 10 pupils might be taught alone and then in a group of 40 and their achievement compared; "uncontrolled" should be obvious.

OUTCOME VARIABLE:

12) *Type of Achievement Measure.* Outcomes were measured by standardized achievement tests, specially designed ad hoc tests, or teachers' assessments of achievement. The latter two categories were grouped.

13) *Quantification of Outcomes.* In some instances, a degree of experimental control could be attained by expressing achievement as gains from pretest to posttest or covariance adjusting posttest means for pretest differences. If this was done, it was noted.

QUANTIFYING OUTCOMES

A simple statistic is desired that describes the relationship between class size and achievement as determined by a study. No matter how many class sizes are compared, the data can be reduced to some number of paired comparisons, a smaller class against a larger class. Certain differences in the findings must be attended to if the findings are later to be integrated. The most obvious differences involve the actual sizes of "smaller" and "larger" classes and the scale properties of the achievement measure. The actual class sizes compared must be preserved and become an essential part of the descriptive measure. The measurement scale properties can be handled by standardizing all mean differences in achievement by dividing by the within group standard deviation (a method that is complete and discards no information at all under the assumption of normal distributions). The eventual measure of relationship seems straightforward and unobjectionable:

$$\Delta_{S-L} = \frac{\bar{X}_S - \bar{X}_L}{\hat{\sigma}},$$

where:

$\bar{X}_S$ is the estimated mean achievement of the *smaller* class which contains S pupils;

$\bar{X}_L$ is the estimated mean achievement of the *larger* class which contains L pupils; and

$\hat{\sigma}$ is the estimated within-class standard deviation, assumed to be homogeneous across the two classes.

As a first approximation to studying the class-size and achievement relationship, it is considered irrelevant that the particular types of achievement that lie behind the variable X are quite different knowledges and skills measured in quite different ways.

If distributional assumptions about X are needed to add meaning to particular values of Δ_{S-L}, normality will be assumed. For example, suppose $\Delta_{S-L} = +1$. Then assuming normal distributions within classes, the average pupil in the smaller class scores at the 84th percentile of the

larger class. These interpretations are occasionally helpful, but seldom critical, and our investment in the normality assumption is not great. It would be no surprise nor any concern if the assumption proved to be more or less wrong, and it's probably not far off in most instances.

CALCULATING Δ_{S-L}

Reports of research frequently omit such basic descriptive measures as means and standard deviations. This omission frequently complicates the calculation of Δ_{S-L}, but seldom obviates it. Transformations of commonly reported statistics (*t*, *F*, etc.) into Δ's can be derived (Glass, 1978). A special problem in calculation of Δ_{S-L} concerns studies in which class size is correlated with achievement across many classrooms (e.g., Coleman, 1966). In these instances, Δ_{S-L} was calculated as follows. The distribution of class-sizes was determined by assuming normality and noting the mean and standard deviation. The regression coefficient was calculated for the regression of achievement (assumed to be calculated on a unit-normal scale) onto class-size *via* $\hat{\beta} = r_{A,CS}/\hat{\sigma}_{CS}$. Then the class sizes at the 25th and 75th percentiles, assuming normality, were determined. These became the "smaller" and "larger" classes. Finally, the achievement in these classes was determined via the formula $\hat{\beta}(X - \overline{X})$ where X is "class size." The value of Δ_{S-L} is then readily calculated. Some studies involved only a dichotomous achievement measure (e.g., "promoted (to the next grade) vs. not promoted"). Proportions thus derived were transformed into metric information and then into values of Δ_{S-L} by means of the probit transformation (see Glass, 1978).

DESCRIBING THE CLASS-SIZE AND ACHIEVEMENT RELATIONSHIP

There exist several alternative statistical techniques for integrating a large set of Δ_{S-L}'s so as to describe the aggregated findings on the class-size and achievement relationship. A large, square matrix could be constructed in which the rows and columns are class sizes and the cell entries are average values of Δ_{S-L}; nearly equal values of average deltas could be connected by lines to form "iso-deltas" in much the manner as economic equilibrium curves are used to depict three variable relationships. Or, a variation of psychometric scaling could be employed: a square matrix of class sizes could be constructed for which each cell entry would be the proportion of times the row class size gave achievement greater than the column class size. This matrix could be scaled by means of Thurstone's Law of Comparative Judgment, which would locate the class sizes along an achievement continuum. (This method was used and the results were reasonably satisfactory; the results appear in the longer report referenced in the footnote on the title page.) Finally, regression equations could be constructed in which Δ_{S-L} is partitioned into a weighted linear combination of *S* and *L* and functions thereof and error. There is much to recommend this latter procedure, and the technique eventually employed is a variation of it.

The regression model selected accounted for variation in Δ_{S-L} by means of *S*, S^2, and *L*. Obviously, something more than a simple linear function of *S* and *L* was needed, otherwise a unit increase in class size would have a constant effect regardless of the starting class size *S*; and the S^2 term seemed as capable of filling the need as any other. The size differential between the larger and smaller class, *L*-*S*, was used in place of *L* for convenience. Thus, the Δ_{S-L} values were used to fit the following model:

$$\Delta_{S-L} = \beta_0 + \beta_1 S + \beta_2 S^2 + \beta_3(L - S) + \epsilon \qquad (1)$$

The regression of Δ_{S-L} onto *S*, S^2, and *L* can only be depicted in three or more dimensions. This restriction is a severe problem when many readers of the findings will consider it punishment enough to be forced to interpret a simple achievement-by-class-size graph. For ease of understanding, then, we wished to represent a complex regression surface in only two dimensions: achievement and class size. Such a representation was made possible by imposing a condition on the set of all Δ_{S-L} that lie on the regression surface. (The condition and methods that underlie the derivation of the single curve from the surface are too detailed and complex to develop here; they can be found in the

long report referenced in the footnote to this paper.) All of the curves below that depict the relationship of class size and achievement were derived by imposing this restriction on the regression surface. Finally, for descriptive purposes, the metric of percentile ranks was chosen over the metric of z-scores; thus the curve z was transformed into a curve of percentile ranks by assuming a normal distribution of achievement.

COMMENT ON STATISTICAL INFERENCE

In the analyses that follow, ordinary matters of statistical inference have been ignored. The application of usual interval estimation procedures or statistical tests makes little sense for two reasons. The data base is laced with a complicated structure of interdependent observations; several comparisons arise from a single study when more than two class sizes are compared, and there is no sensible way to reduce each study to one observation. Even if a study involves comparing only two class sizes, there might have been comparisons of reading and math achievement. It makes far less sense to average these than to let each separately entered in the data base. The data bases of most meta-analyses are complex nested and multilevel arrangements. The methods of analyzing them fully await a full explication; methodological work on these problems has been launched in promising directions (Burstein, 1978). Secondly, randomization is absent from the data set in any form that would make probabilistic models based on it applicable. To the extent that one might care to infer to populations of pupils, the sample size is so large that significance tests would be an empty pro form ritual. To the extent one might wish to infer to populations of studies, it must be recognized that the studies included have in no way been sampled from any conceivable population. Error and instability of various odd sorts exist in the data set; how they should be dealt with is not at all apparent.

FINDINGS

The report of findings falls into two broad categories: (1) description of the data base and (2) regression analyses relating achievement and class size.

DESCRIPTION OF THE DATA BASE

In all, 77 different studies were read, coded, and analyzed. These studies yielded a total of 725 Δ's. The comparisons are based on data from a total of nearly 900,000 pupils spanning 70 years research in more than a dozen countries. (The entire set of data is reproduced in the longer report identified in the footnote on the first page of this paper.)

The total body of evidence can be described partly in quantitative terms through use of frequency distributions of characteristics of the studies. These tabulations will be presented in terms of Δ's rather than studies. The descriptive data do not only communicate an understanding of the evidence upon which the conclusions rest; they point to the relatively over-studied and under-studied aspects of the topic and can help guide future research on class size and achievement.

In Table 1 appears the frequency distribution of Δ's by *year in which the study appeared.* It is clear from Table 1 that class-size research was an active early topic in educational research, was largely abandoned for 30 years after 1930, and has been resurrected in the last 15 years.

In Table 2 appear data on the publication source from which the comparisons were drawn. Although published journal

TABLE 1
Class-Size Comparisons (Δ) by Year of Study

Year	No. of Δ's	%	Cumulative %
1900–1909	22	3.0	3.0
1910–1919	184	25.4	28.4
1920–1929	138	19.0	47.4
1930–1939	47	6.5	53.9
1940–1949	1	0.0	53.9
1950–1959	62	8.6	62.5
1960–1969	150	20.8	83.3
1970–1979	121	16.7	100.0
	725	100.0	

TABLE 2
Class-Size Comparisons (Δ) by Publication Source

Source	No. of Δ's	%
Journal	474	65.4
Book	114	15.7
Thesis	60	8.3
Unpublished	77	10.6
	725	100.0

articles are the major source of data, about 20% of the data were found in theses and unpublished reports—both of which have not been well covered in previous reviews.

In Table 3 appear the frequencies of comparisons categorized by the school subject taught in the study. Nearly half of the comparison came from studies in which elementary school pupils were taught all subjects in classes of varying sizes. There is surprisingly little work on reading alone; however, the 342 "all subjects combined" comparisons typically include reading as an important element.

In Table 4 are reported the numbers of hours of instruction given in the classes being compared. The range is enormous, from a single hour for a very small scale tutoring study, to 9,000 hours, representing 5 years of elementary school instruction. The "hours of instruction" distribution shows three modes: 50, 180, and 900 hours. These times correspond to a 3 credit-hour semester-long course, a 5 credit-hour year-long course, and a year of teaching 5 hours per day. The literature does not lack studies conducted over significant intervals of time. The average duration is 536 hours with a standard deviation of 1,033 hours and a skewness of 5.58.

In Table 5 appears the distribution of comparisons for various ages of pupils. Research is spread fairly evenly across the elementary and secondary grades. The first 4 years of school are only slightly underrepresented. The average age represented in the 725 comparisons is 12.3 years with a standard deviation of 4.0 years.

The next few items of information concern the experimental validity of the comparisons, i.e., the incidence of various experimental controls and ex post facto adjustments. In Table 6, the comparisons are tabulated by the type of assignment of pupils to the different size classes. The type of assignment labeled "repeated measures" refers to the use of the same group of pupils in both a small and a large class and the comparison of their achievement in the two classes. Each of the first three types of assignment represents reasonably good attempts at eliminating gross inadequacies in design; these three conditions account for slightly more than half of all the comparisons. Even though half of the comparisons involved comparing naturally constituted and nonequivalent large and small classes, some of them were based on ex post facto statistical adjustments for preexisting differences. So the data are not half worthless; indeed whether the experimental inadequacies are important mediators of findings is an empirical fact—rather than an a priori

TABLE 3
Class-Size Comparisons (Δ) by Subject of Instruction

Subject Taught	No. of Δ's	%
All Subjects Combined (i.e., elementary school classes)	343	47.2
Reading	39	5.4
Mathematics	84	11.6
Language	144	19.9
Psychology	23	3.2
Natural/Physical Sciences	28	3.9
Social Sciences and History	40	5.5
All Others	25	3.4
	725	100.0

TABLE 4
Class-Size Comparisons (Δ) by Hours of Instruction

Hours Instruction	No. of Δ's	%	Cumulative Percent
1-10	26	3.6	4.5
11-20	40	5.5	11.4
21-40	40	5.5	18.4
41-60	50	6.9	27.0
61-100	30	4.1	32.2
101-150	23	3.2	36.2
151-200	126	17.4	58.1
201-300	17	2.3	61.0
301-400	3	0.4	61.5
401-500	30	4.1	66.7
501-800	37	5.1	73.1
801-1000	132	18.3	96.0
3600	18	2.6	99.1
9000	5	0.8	100.0
Unknown	148	20.4	
	725	100.0	

TABLE 5
Class-Size Comparisons (Δ) by Age of Pupils

Age	No. of Δ's	%	Cumulative %
5-6	56	7.7	7.7
7-8	55	7.6	15.3
9-10	198	27.3	42.6
11-12	98	13.5	56.1
13-14	81	11.1	67.2
15-16	109	15.0	82.2
17-18	108	14.9	97.1
19 and older	20	2.8	100.0
	725	100.0	

TABLE 6
Class-Comparisons (Δ) by Assignment of Pupils to the Small and Large Classes

Type of Assignment	No. of Δ's	%
Random	110	15.2
Matched	235	32.4
"Repeated Measures"	18	2.5
Uncontrolled	362	49.9
	725	100.0

judgment—which will be examined in detail later in this report.

Many studies attempted to control for the initial nonequivalence of small and large classes by correcting the achievement dependent variable, either by calculating simple gain-scores or by covariance adjusting means. We hasten to point out that an uncorrected dependent variable does not necessarily indicate a comparison of poor quality. Corrections might be quite irrelevant in a study that matched or randomly assigned pupils to classes.

Finally, the comparisons can be described by whether achievement was measured with a "standardized test" (i.e., a published test for a national market) or an ad hoc instrument designed specifically to measure achievement in the immediate context of the instruction given (see Table 7).

In Table 8 appears the joint distribution of smaller and larger class sizes on which the 725 Δ's are based. For example, six Δ's derive from comparisons of group sizes 1 and 3. The table contains only 550 entries instead of 725, since comparisons would not be recorded in this tabulation if S and L were contained within the same broad category (e.g., if $S = 18$ and $L = 22$). Such comparisons were incorporated in all subsequent analyses, but the need to keep Table 8 down to a reasonable size precluded the classification of all 725 Δ's. It is apparent in Table 8 which size comparisons have been relatively overstudied and which have been neglected. The dearth of comparisons of instructional group sizes in the range from 2 to 10 pupils is particularly apparent.

REGRESSION ANALYSES

The dependent variable, Δ_{S-L}, in the regression analyses had the following statistical properties:

Properties of Distribution of Δ_{S-L}

a) $N = 725$.
b) Mean = .088; Median = .050.
c) 40% of the Δ_{S-L} were negative; 60%, positive.
d) Standard deviation = 0.401.
e) Range: −1.98 to 2.54.
f) Skewness = 1.151; Kurtosis = 7.461

On the average, the 725 Δ_{S-L}'s were positive, i.e., over all comparisons available—regardless of the class sizes compared—the results favored the smaller class by about a tenth of a standard deviation in achievement. This finding is not too interesting, however, since it disregards the sizes of the classes being compared. One interesting feature of the Δ's is that only 60% of them are positive, i.e., favor the smaller class in achievement. This is so, even though every effort was made in compiling the data base to include studies spanning the full range of class sizes from individual tutorials to huge lectures. One suspects that the odds of observing a positive Δ_{S-L} in the typical class-size range so often studied (e.g., 15 to 40) are even smaller, perhaps as low as 55% to 45%.

In these rough estimates, one of the fundamental problems is revealed that has made the class-size literature so difficult for reviewers. If the relationship one seeks has only 55 to 45 odds of appearing and one looks for it without all the tools of statistical analyses that can be mustered, the chances of finding it are small. One need not wonder why narrative reviews of a dozen or two studies produced little but confusion.

To make sense of the class size and achievement relationship, one must account for the magnitude of the Δ's and their variance in terms of the actual sizes of the smaller and larger classes. These are the purposes of the regression analyses. In the remainder of this section, such regression analyses are reported for the entire

TABLE 7
Class-Comparisons (Δ) by Type of Achievement Measure

Type of Achievement Measure	No. of Δ's	%
Standardized Test	318	43.9
Ad Hoc Measure	407	56.1
	725	100.0

TABLE 8

Joint Distribution of Smaller and Larger Class-sizes in the Comparisons Δ_{S-L}

Smaller Class-size	Larger Class-size: 1	2	3	4-5	6-10	11-16	17-23	24-34	≥35
1	—	1	6	1	3	7	1	34	0
2		—	0	1	0	0	1	0	0
3			—	0	0	0	0	6	0
4-5				—	0	0	1	2	0
6-10					—	8	0	5	2
11-16						—	19	44	27
17-23							—	78	106
24-34								—	197
≥35									—

data set and for the data set stratified on several important characteristics of the studies (e.g., age of pupils, validity of the study).

1. *REGRESSION ANALYSIS FOR ENTIRE DATA SET*

The model $\Delta_{S-L} = \beta_0 + \beta_1 S + \beta_2 S^2 + \beta_3 (L - S) + \epsilon$ was fit by least-squares for the 725 points. The results were as follows:

Variables	Mean	St. Dev.
Independent:		
S, size of smaller class	23.243	11.463
S^2	671.446	603.463
$L - S$, difference between large and small class	19.906	20.671
Dependent: Δ_{S-L}	0.088	0.401

Correlations

	S	S^2	$L - S$	Δ
S	1	.932	.004	−.271
S^2		1	.011	−.135
$L - S$			1	.047

Regression Analysis

Multiple R = .426

Source of Variation	df	MS
Regression	3	6.684
Residual	721	.132

$\hat{\beta}_0 = .57072$ $\hat{\beta}_1 = -.03860$ $\hat{\beta}_2 = .00059$ $\hat{\beta}_3 = .00082$

The regression equation for estimating Δ_{S-L} is

$$\hat{\Delta}_{S-L} = .57072 - .03860S + .00059S^2 + .00082(L - S)$$

Based on the entire data set, the following table of standardized comparisons for selected class sizes can be constructed:

Small Class Size	Large Class Size	Standardized Differential Achievement, Δ_{S-L}
1	40	.565
10	40	.268
20	40	.051
30	40	−.048
1	25	.552
5	25	.409
10	25	.256
15	25	.133
20	25	.039

These data show that the difference in achievement between class-size 1, i.e., individual instruction, and class-size 40 is more than one-half standard deviation. The difference between class-size 20 and class-size 40 is only about five hundredths standard deviation. Class-size differences at the low end of the scale have quite important effects on achievement; differences at the high end have little effect.

The curved regression surface can be reduced to a single line curve in a plane by imposing a consistency condition on the regression surface. On this curve the difference between achievement in class-sizes 1 and 40 is .551 + .009 = .560. The curve is presented in Figure 1. The ordinate is represented by a standard score metric; the zero point is arbitrarily fixed at a class-size of 30.

In Figure 2, the curve in Figure 1 is translated into a metric of percentile ranks

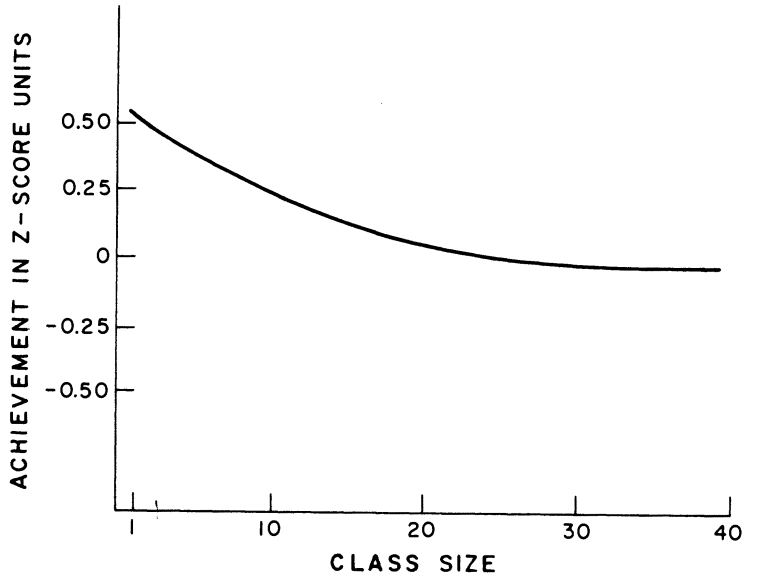

FIGURE 1. Consistent regression line for achievement (in z-score units) onto class size.

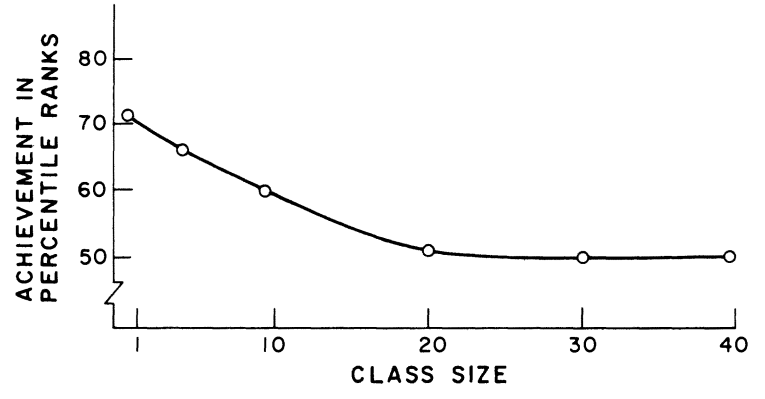

FIGURE 2. Consistent regression line for achievement (percentile ranks) onto class size (all data).

on the ordinate by assuming a normal distribution of achievement. There it can be seen that the difference in average performance from class-size 1 to class-size 40 is from above the 70th percentile to just below the 50th. There is nearly a 10 percentile rank difference between instructional groups of sizes 10 and 20 pupils.

2. *REGRESSION ANALYSES FOR SUBSECTIONS OF THE DATA*

Regression analyses were performed for many smaller portions of the entire data set in an attempt to determine which characteristics of the studies might mediate the size of the class size and achievement relationship. More than a dozen factors were employed in splitting the data base: year of study, subject taught, age of pupils, IQ, type of test, etc. Few of these characteristics were systematically related to the strength of the class size and achievement correlation. Among those factors of discrimination that produced virtually identical regression lines were "source of data," "subject taught," duration of instruction," "pupil IQ," and "type of achievement measure." From among these few characteristics that appeared to interact with the relationship, three stand out as particularly interesting: year of the study, level of schooling (elementary vs. secondary), and internal validity of the study. The complete regression analyses

will be reported below for the latter two characteristics. Details of the "year of study" analyses will not be reported here; suffice it to note that there is no correlation between class size and achievement in those studies carried out before 1940 and a strong relationship favoring smaller classes in post-1960 studies. Two in the two eras differ in many respects, most notably in terms of the sophistication of both experimental design and measurement.

Elementary vs. Secondary. The curvilinear regression model in (2) was fit separately for pupils of age 11 years or younger (elementary) and 12 years or older (secondary). The summary statistics and solutions are as follows:

Elementary (N = 342)

Variables		Mean	St. Dev.
Independent:	S	22.836	11.758
	S^2	659.345	556.750
	$L - S$	13.915	8.311
Dependent:	Δ	0.092	0.256

Correlations

	S	S^2	$L - S$	Δ
S	1	.951	−.377	−.343
S^2		1	−.345	−.215
$L - S$			1	.241

Regression Analysis—Elementary Grades

Multiple R = .505

Source of Variation	df	MS
Regression	3	1.898
Residual	338	.049

$\hat{\beta}_0 = .38503 \quad \hat{\beta}_1 = -.02995 \quad \hat{\beta}_2 = .00052 \quad \hat{\beta}_3 = .00344$

$\hat{\Delta}_{S-L} = .38503 - .02995S + .00052S^2 + .00344(L - S)$

Regression Analysis—Secondary Grades

Multiple R = .439

Source of Variation	df	MS
Regression	3	5.667
Residual	345	0.207

$\hat{\beta}_0 = .75539 \quad \hat{\beta}_1 = -.05024 \quad \hat{\beta}_2 = .00071 \quad \hat{\beta}_3 = .00111$

$\hat{\Delta}_{S-L} = .75539 - .05024S + .00071S^2 + .00111(L - S)$

Some particularly interesting values of Δ on the two regression surfaces are listed below:

Smaller Class Size	Larger Class Size	Δ, Standardized Differential Achievement Elementary	Secondary
1	40	.490	.749
10	40	.241	.357
20	40	.063	.057
30	40	−.011	−.102
1	10	.387	.716
3	10	.324	.619
5	10	.265	.527

The class size and achievement relationship seems consistently stronger in the secondary grades than in the elementary

Secondary (N = 349)

Variables		Mean	St. Dev.
Independent:	S	23.642	11.168
	S^2	683.304	646.598
	$L - S$	25.777	26.641
Dependent:	Δ	0.085	0.504

Correlations

	S	S^2	$L - S$	Δ
S	1	.924	.112	−.259
S^2		1	.098	−.106
$L - S$			1	.024

grades. This interaction is also seen in Figure 3 where the consistent two-dimensional curves are drawn. The ordinate scale in Figure 3 is percentile ranks.

Well-Controlled vs. Poorly-Controlled Studies. The comparisons were distinguished on the basis of degree of experimental control exercised in the study. Although many features of experimental control could have been noted and analyzed, the method of assignment of pupils to classes of different sizes proved to be the most important. Over 100 Δ's came from studies in which pupils were assigned at random to larger and smaller classes; over 300 comparisons were "uncontrolled," i.e., naturally constituted larger and smaller classes were compared. The summary statistics and solutions of the regression models are as follows:

Poorly-Controlled (N = 334)

Variables		Mean	St. Dev.
Independent:	S	26.895	10.923
	S^2	842.302	667.164
	$L - S$	15.210	11.671
Dependent:	Δ	0.051	0.261

Correlations

	S	S^2	$L - S$	Δ
S	1	.957	−.081	−.034
S^2		1	−.066	−.011
$L - S$			1	.172

Well-Controlled (N = 108)

Variables		Mean	St. Dev.
Independent:	S	11.732	10.228
	S^2	241.269	287.327
	$L - S$	17.889	12.767
Dependent:	Δ	0.401	0.554

Correlations

	S	S^2	$L - S$	Δ
S	1	.951	−.062	−.549
S^2		1	−.018	−.451
$L - S$			1	.241

Regression Analysis—Poorly-Controlled Studies

Multiple R = .187

Source of Variation	df	MS
Regression	3	0.263
Residual	330	0.066

$\hat{\beta}_0 = .07399 \quad \hat{\beta}_1 = -.00587 \quad \hat{\beta}_2 = .00009 \quad \hat{\beta}_3 = .00376$

Regression Analysis—Well-Controlled Studies

Multiple R = .621

Source of Variation	df	MS
Regression	3	4.226
Residual	104	0.194

$\hat{\beta}_0 = .69488 \quad \hat{\beta}_1 = -.06334 \quad \hat{\beta}_2 = .00128 \quad \hat{\beta}_3 = .00783$

The curves in Figure 4 show large differences in the class-size and achievement relationship depending on whether pupil assignment was random or uncontrolled. This finding contrasts sharply with similar analyses of the association between experimental design quality and effects in the field of psychotherapy (Smith & Glass, 1977). The difference is probably due to the magnitude of the effects that are the object of the research in the two fields. The typical psychotherapy effect (therapy vs. control group) is between three-quarters and a full standard deviation (Smith, Glass, & Miller, 1979); the typical class-size study was seeking to establish an effect of less than one-tenth standard deviation. It is little surprise, then, that in one field experimental design quality proves critical, and in another field it does not.

In an area of research where the quality of methodology interacts with the findings of studies, the results of the best designed

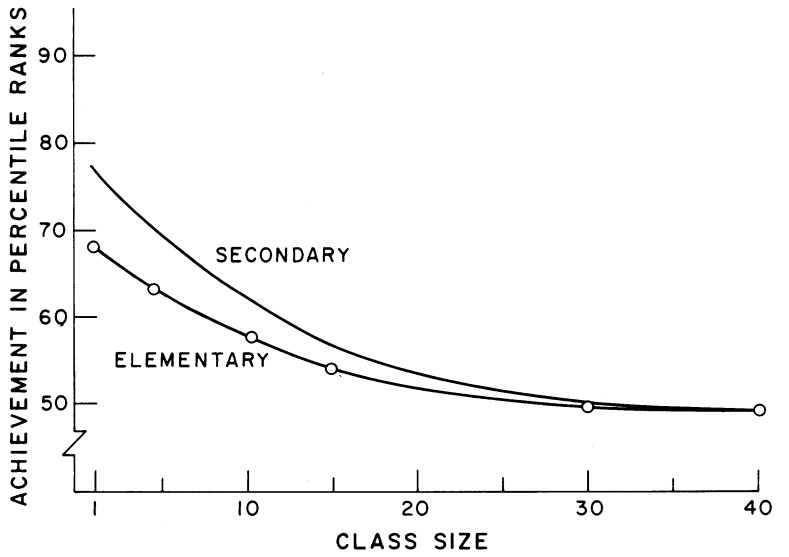

FIGURE 3. Consistent regression lines for the regression of achievement (expressed in percentile ranks) onto class size for elementary and secondary grades.

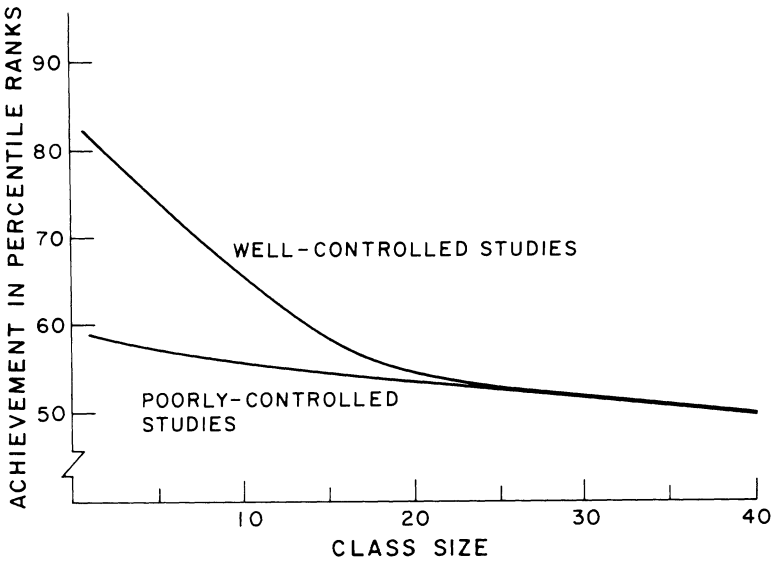

FIGURE 4. Consistent regression lines for the regression of achievement (expressed in percentile ranks) onto class size for studies that were well-controlled and poorly-controlled in the assignment of pupils to classes.

studies should be given more weight in drawing conclusions. The curve for the well-controlled studies in Figure 4, then, is probably the best representation of the class-size and achievement relationship.

Concern was expressed by several persons who examined the preliminary analyses that the curve for the well-controlled studies in Figure 4 might depend excessively on the 20 or 30 comparisons of very small class sizes (e.g., one and two up to five) in the data base. When all those comparisons for which $S = 1$ were removed, the curve in Figure 4 for well-controlled studies was even steeper than that shown; this finding is contrary to the claim that tutoring studies skewed the curve unnaturally. When all comparisons for which S was less than 6 were removed, the curve for well-controlled studies became less steep; however, it still rose from the 50th percentile at size 40 to the 60th at size 10, the 67th at size 5 and the 74th at size 1.

CONCLUSIONS

Research on class size and achievement is a particularly complex body of findings to integrate and understand. The integration of this literature has required more sophisticated analysis than has previously been applied to the problem. The meta-analysis of the research reported here has drawn heavily on precise quantitative description and analysis. A clear and strong relationship between class size and achievement has emerged. The relationship seems slightly stronger at the secondary grades than the elementary grades; but it does not differ appreciably across different school subjects, levels of pupil IQ, or several other obvious demographic features of classrooms. The relationship is seen most clearly in well-controlled studies in which pupils were randomly assigned to classes of different sizes. Taking all findings of this meta-analysis into account, it is safe to say that between class-sizes of 40 pupils and one pupil lie more than 30 percentile ranks of achievement. The difference in achievement resulting from instruction in groups of 20 pupils and groups of 10 can be larger than 10 percentile ranks in the central regions of the distribution. There is little doubt that, other things equal, more is learned in smaller classes.

REFERENCES

Bausell, R. B., Moody, W. B., & Walze, F. N. A factorial study of tutoring versus classroom instruction. *American Educational Research Journal, 9*, 1972, 591–598.

Burstein, L. The role of levels of analysis in the specification of educational effects. Los Angeles: Graduate School of Education, UCLA, 1978.

Coleman, J. S. et al. *Equality of educational*

opportunity. Washington: U.S. Government Funding Office, 1966.

Cornman, O. P. Size of classes and school progress. *The Psychological Clinic, 3,* 1909, 206–212.

Glass, G. V Integrating findings: The meta-analysis of research. *Review of Research in Education,* 1978, *5,* 351–379.

Hartley, S. S. Meta-analysis of the effects of individually paced instruction in mathematics. Ph.D. Thesis, University of Colorado, 1977.

Mayeske, G. W. *A study of one nation's schools.* U.S. Office of Education, (undated).

Rice, J. M. Educational research: A test in arithmetic. *The Forum,* 1902, *34,* 281–297.

Ryan, D. W., & Greenfield, T. B. *The Class Size Question.* Toronto, Ontario: The Ministry of Education, 1975.

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

Smith, M. L., Glass, G. V, & Miller, T. I. *The benefits of psychotherapy.* Baltimore: The Johns Hopkins University Press, 1979.

Tope, R. E., Groom, E., & Beeson, M. F. Size of class and school efficiency. *Journal of Educational Research,* 1924, *9,* 126–132.