DOCUMENT RESUME

ED 348 174 PS 020 898

AUTHOR	Slavin, Robert E.
TITLE	Ability Grouping and Student Achievement in Elementary Schools: A Best-Evidence Synthesis. Report No. 1.
INSTITUTION	Center for Research on Elementary and Middle Schools, Baltimore, MD.
SPONS AGENCY	Office of Educational Research and Improvement (ED), Washington, D.C.
PUB DATE	Jun 86
CONTRACT	OERI-G-86-0006
NOTE	127p.
PUB TYPE	Information Analyses (070)
EDRS PRICE	MF01/PC06 Plus Postage.
DESCRIPTORS	*Ability Grouping; *Academic Achievement; *Educational Practices; Elementary Education; *Elementary School Students; Literature Reviews; Mathematics Instruction; Meta Analysis; Reading Instruction
IDENTIFIERS	Best Evidence Synthesis

ABSTRACT

This report reviews research on the effects of between- and within-class ability grouping on the achievement of elementary school students. The review technique, known as "best-evidence synthesis," combines features of meta-analytic and narrative reviews. Overall, evidence does not support assignment of students to self-contained classes according to ability, but grouping plans involving cross-grade assignment for selected subjects can increase student achievement. Research particularly supports the Joplin Plan, cross-grade ability grouping for reading only, and forms of nongraded programs involving multiple groupings for different subjects. Within-class ability grouping in mathematics is also found to be instructionally effective. Ability grouping is held to be maximally effective: (1) when it is done only for one or two subjects, with students remaining in heterogeneous classes most of the day; (2) when it greatly reduces student heterogeneity in a specific skill; (3) when group assignments are frequently reassessed; and (4) when teachers vary the level and pace of instruction according to students' needs. (An 18-page reference list is appended). (Author/RH)

Center for Research On Elementary & Middle Schools

Report No. 1

June 1986

ABILITY GROUPING AND STUDENT ACHIEVEMENT IN ELEMENTARY SCHOOLS: A BEST-EVIDENCE SYNTHESIS

Robert E. Slavin

Center Staff

Edward L. McDill, Co-Director
James M. McPartland, Co-Director

Karl L. Alexander
Henry J. Becker
Jomills H. Braddock II
Renee B. Castaneda
Barbara S. Colton
Diane B. Diggs
Doris R. Entwisle
Joyce L. Epstein
Anna Marie Farnish
Denise C. Gottfredson
Gary D. Gottfredson
Edward J. Harsch
Brigette B. Hinte
John H. Hollifield

Lois G. Hybl
Marva J. Jeffery
Nancy L. Karweit
Melvin L. Kohn
Mary S. Leighton
Barbara M. Luebbe
Nancy A. Madden
Barbara E. McHugh
Laura B. Rice
Karen C. Salinas
Dorothy C. Sauer
Robert J. Stevens
Shi-Chang Wu

Center Liaison

Rene Gonzalez, Office of Educational Research and Improvement

National Advisory Board

Patricia A. Bauch, Catholic University of America
Jere Brophy, Michigan State University
Jeanne S. Chall, Harvard University
James S. Coleman, University of Chicago
Edgar G. Epps, University of Chicago
Barbara Heyns, New York University
Michael W. Kirst, Chair, Stanford University
Rebecca McAndrew, West Baltimore Middle School
Jeffrey Schneider, National Education Association

ABILITY GROUPING AND STUDENT ACHIEVEMENT IN ELEMENTARY SCHOOLS: A BEST-EVIDENCE SYNTHESIS

Grant No. OERI-G-86-0006

Robert E. Slavin

Report No. 1

June 1986

Published by the Center for Research on Elementary and Middle Schools, supported as a national research and development center by funds from the Office of Educational Research and Improvement, U.S. Department of Education. The opinions expressed in this publication do not necessarily reflect the position or policy of the OERI, and no official endorsement should be inferred.

Center for Research on Elementary and Middle Schools
The Johns Hopkins University
3505 North Charles Street
Baltimore MD 21218

Printed and assembled by:
VSP Industries
2440 West Belvedere Avenue
Baltimore MD 21215

The Center

The mission of the Center for Research on Elementary and Middle Schools is to produce useful knowledge about how elementary and middle schools can foster growth in students' learning and development, to develop and evaluate practical methods for improving the effectiveness of elementary and middle schools based on existing and new research findings, and to develop and evaluate specific strategies to help schools implement effective research-based school and classroom practices.

The Center conducts its research in three program areas: (1) Elementary Schools; (2) Middle Schools, and (3) School Improvement.

The Elementary School Program

This program works from a strong existing research base to develop, evaluate, and disseminate effective elementary school and classroom practices; synthesizes current knowledge; and analyzes survey and descriptive data to expand the knowledge base in effective elementary education.

The Middle School Program

This program's research links current knowledge about early adolescence as a stage of human development to school organization and classroom policies and practices for effective middle schools. The major task is to establish a research base to identify specific problem areas and promising practices in middle schools that will contribute to effective policy decisions and the development of effective school and classroom practices.

School Improvement Program

This program focuses on improving the organizational performance of schools in adopting and adapting innovations and developing school capacity for change.

This report, prepared by the Elementary School Program, synthesizes research on ability grouping in elementary schools to identify grouping practices that promote student achievement.

Ability Grouping and Student Achievement in Elementary Schools: A Best-Evidence Synthesis

Abstract

This article reviews research on the effects of between- and within-class ability grouping on the achievement of elementary school students. The review technique, best-evidence synthesis, combines features of meta-analytic and narrative reviews. Overall, evidence does not support assignment of students to self-contained classes according to ability (median ES = .00), but grouping plans involving cross-grade assignment for selected subjects can increase student achievement. Research particularly supports the Joplin Plan, cross-grade ability grouping for reading only (median ES = +.45) and forms of nongraded programs involving multiple groupings for different subjects (median ES = +.29). Within-class ability grouping in mathematics is also found to be instructionally effective (median ES = +.34). Ability grouping is held to be maximally effective when it is done only for one or two subjects, with students remaining in heterogeneous classes most of the day; when it greatly reduces student heterogeneity in a specific skill; when group assignments are frequently reassessed; and when teachers vary the level and pace of instruction according to students' needs.

Acknowledgments

This work benefitted from the assistance of John Waters in locating references and from the helpful comments of Nancy Karweit, Nancy Madden, and Robert Stevens.

Table of Contents

Introduction.. 1

Comprehensive Ability Grouping In the Elementary School.............................. 10

Research on Comprehensive Ability Grouping In the Elementary School

Ability Grouped Class Assignment 22

Regrouping for Reading and Mathematics 32

Joplin and Nongraded Plans 37

Comprehensive Nongraded Plans 47

Within-Class Ability Grouping 53

Discussion 60

Conclusions 74

References 79

Ability Grouping and Student Achievement in Elementary Schools: A Best-Evidence Synthesis

Ability grouping is one of the oldest and most controversial issues in educational psychology. Hundreds of studies have examined the effects of various forms of between-class ability grouping (e.g., tracking, streaming) and within-class ability grouping (e.g., reading, math groups). By 1930, Miller and Otto had already located twenty experimental studies on ability grouping, and Martin (1927) listed eighty-three "selected references" on the topic.

Scores of reviews of the between-class ability grouping literature have been written. Almost without exception, reviews from the 1920's to the present have come to the same general conclusion: that between-class ability grouping has few if any benefits for student achievement. Recently, meta-analyses on ability grouping in elementary (C-L. Kulik and J. Kulik, 1984) and in secondary schools (Kulik & Kulik, 1982) have claimed small positive achievement effects of between-class ability grouping, with high achievers gaining the most from the practice.

Despite a half-century of widespread agreement (among researchers, at least) that between-class ability grouping is of little value in enhancing student achievement, the practice is nearly universal in some form in secondary schools and very common in elementary schools. Recent data are lacking, but over time most teachers at all levels have reported both using and believing in some kind of ability grouping (e.g., NEA, 1968; Wilson & Schmits, 1978). Yet in recent years many districts have begun to reexamine ability grouping, often out of a concern that students low in socioeconomic status, in particular minority students, are discriminated against by being disproportionately placed in low tracks. In fact, ability grouping has become a major issue in many ongoing desegration cases (e.g., Hobson vs. Hansen, 1967) where the plaintiffs have argued that ability grouping is used as a means of resegregating Black and Hispanic students within ostensibly integrated schools (see McPartland, 1968).

Although many reviews of ability grouping have been written, the most recent comprehensive reviews in this area were written more than sixteen years ago (e.g., Borg, 1965; Findley & Bryan, 1971; Heathers, 1969; NEA, 1968). More recent reviews (e.g., Esposito, 1973; Good & Marshall, 1984; Persell, 1977) have referred to the earlier reviews rather than synthesizing the original evidence. The Kuliks' meta-analyses have extracted effect-size data from large numbers of primary studies, but have done little beyond this to explore the substantive and methodological issues underlying these effects (see Slavin, 1984a).

The present paper reviews the literature on ability grouping in elementary schools from the vantage point of the 1980's. It uses a review strategy called "Best-Evidence Synthesis" (Slavin, 1986), a method which incorporates the best features of both meta-analytic and traditional narrative review. The main elements of a best evidence synthesis are as follows:

- Clearly specified, defensible *a priori* criteria for inclusion of studies are established.
- All published and unpublished studies which meet criteria are located and included.
- Where possible, effects sizes for included studies are computed. Effect size is operationalized as the mean of all experimental-control differences on related measures divided by their standard deviations.
- When effect sizes cannot be computed, effects of studies which meet inclusion criteria are characterized as positive, negative, or zero rather than excluded.
- Apart from computation of effect size and use of well-specified inclusion criteria, best-evidence syntheses are identical to traditional narrative reviews. Individual studies and methodological and substantive issues are discussed in the detail typical of the best narrative reviews.

The present paper is the first application of best-evidence synthesis. Since both meta-analytic and traditional narrative reviews exist in the area of ability grouping, this paper allows for a clear contrast between the methods and conclusions of best-evidence syntheses as opposed to meta-analytic and narrative reviews on the same topic.

What is Ability Grouping?

One important problem in discussing "ability grouping" is that the term has many meanings. Several quite different programs or policies go under this heading. In general, ability grouping implies some means of grouping students for instruction by ability or achievement so as to reduce their heterogeneity. However, various grouping plans differ in ways likely to have a considerable impact on the outcomes of grouping. Some common forms of ability grouping are described below.

Ability Grouped Class Assignment. In this plan, students are assigned on the basis of ability or achievement to one self-contained class (usually at the elementary level) or to one class which moves together from teacher to teacher, as in block scheduling in junior high schools.

Curriculum Tracking. A special form of ability grouped class assignment unique to the secondary level is curriculum tracking, assignment of students by ability or achievement to tracks, such as college preparatory, general, or vocational. In secondary schools using such groupings, students may take all courses within their track, or may have some heterogeneously grouped classes. Typically, ability grouping within tracks is not done.

Specialized secondary schools (e.g., schools for the gifted, vocational schools) might be considered one form of curriculum tracking. In Europe, different levels of secondary schools serve a similar tracking function. For example, in West Germany students planning to attend the university go to the gymnasium, less highly skilled students attend the realschule, and students preparing for vocations attend the hauptschule.

Regrouping for Reading or Mathematics (Ability Grouping for Selected Subjects). Often, students are assigned to heterogeneous homeroom classes for part or most of the day, but are "regrouped" according to achievement level for one or more subjects. In the elementary grades, regrouping is often done for reading (and occasionally mathematics), where all students at a particular grade level have reading scheduled at the same time and are resorted from their heterogeneous homerooms into classes that are relatively homogeneous in reading level. When regrouping for reading is done across grade levels, this is called the "Joplin Plan" (see below).

In secondary schools students are often ability grouped for some subjects (e.g., mathematics) but not for others (e.g., social studies). Ability grouping for selected subjects in secondary schools may involve having higher- and lower-achieving sections of the same course, or may involve assigning students to different courses, as when ninth graders are assigned either to Algebra I or to General Mathematics.

Joplin Plan. One special form of regrouping for reading is called the Joplin Plan (Floyd, 1954), in which students are assigned to heterogeneous classes most of the day but are regrouped for reading across grade lines. For example, a reading class at the fifth grade, first semester reading level might include high-achieving fourth graders, average achieving fifth graders, and low-achieving sixth graders. Reading group assignments are frequently reviewed, so that students may be reassigned to a different reading class if the performance warrants it.

One important consequence of cross-grade grouping and flexible assignment is that reading classes contain only one or at most two reading groups, increasing the amount of time available for direct instruction over that typical of reading classes containing three or more reading classes. The Joplin Plan was principally an innovation of the late 1950's and early 1960's, after which time interest in cross-grade grouping turned more toward nongraded plans (see below).

Nongraded Plans. The term "nongraded" or "ungraded" refers to a variety of related grouping plans. In its original conception (Goodlad and Anderson, 1963), nongraded programs are ones in which grade-level designations are entirely removed, and students are placed in flexible groups according to their performance level, not their age. Full-scale nongraded plans might use team teaching, individualized instruction, learning centers, and other means of accommodating student differences in all academic subjects. Students in nongraded programs might complete the primary cycle (grades 1-3) in two years, or may take four years to do so. The curriculum in each subject may be divided into levels (e.g., nine or twelve levels for the primary grades) through which students progress at their own rates, picking up each year where they left off the previous year. This "continuous progress" aspect of nongrading give students a feeling that they are always moving forward; for example, rather than being assigned to the low reading group each year, a low achieving student simply progresses from level to level at a slower rate.

Some of the nongraded programs evaluated in the 1960's and early '70's did use the flexible, complex grouping arrangements envisioned by Goodlad and Anderson (1963), while others did not. For example, several programs described by their authors as "nongraded" or "ungraded" were in fact virtually indistinguishable from the Joplin Plan. That is, students were assigned to heterogeneous classes most of the day but regrouped across grade lines for reading. One study (Morris, 1969) used "nongraded" to refer to a program in which students were regrouped for reading and math within grade levels, while another (Tobin, 1966) used "ungraded" to refer to a traditionally organized reading program in which high achievers were allowed to work on basals above their nominal grade level.

Special Classes for High Achievers. In many elementary and secondary schools, gifted, talented or otherwise superior students may be assigned to a special class for part or all of their school day, while other students remain in relatively heterogeneous classes.

Special Classes for Low Achievers. One of the most common forms of "ability grouping" is the assignment of students with learning problems to special or remedial classes for part or all of their school day.

Within-Class Ability Grouping. Regardless of the use or non-use of ability grouping of classes, most elementary teachers use some form of within-class ability grouping. The most common form of within-class ability grouping is the use of reading groups, where teachers assign students to one of a small number of groups (usually three) on the basis of reading level. These groups work on different materials at rates unique to their needs and abilities. Similar methods are often used in mathematics, where there may be two or more math groups operating at different levels and rates.

In another common form of within-class ability grouping in elementary mathematics, the teacher presents a lesson to the class as a whole, and afterwards, while the students are working problems, the teacher provides enrichment or extension to a high-achieving group, remediation or re-explanation to low achievers, and something in between to average achievers.

Group-paced mastery learning (Bloom, 1976) may be seen as one form of flexible within-class ability grouping, in that students are grouped after each lesson into "masters" and "non-masters" groups on the basis of a formative test. Non-masters receive corrective instruction while masters do enrichment activities. Finally, individualized or continuous-progress instruction may be seen as extreme forms of ability grouping, as each student may be in a unique "ability group" of one.

Theoretical Advantages and Disadvantages of Ability Grouping

Similar lists of advantages and disadvantages of ability grouping have been given by theorists and reviewers for more than fifty years (see, for example, Billett, 1932; Borg, 1965; Esposito, 1971; Findley & Bryan, 1970; Good & Marshall, 1984; Heathers, 1969; NEA, 1968; Miller & Otto, 1930). Ability grouping is supposed to increase student achievement primarily by reducing the heterogeneity of the class or instructional group, making it more possible for the teacher to provide instruction that is neither too easy nor too hard for most students. Ability grouping is assumed to allow the teacher to increase the pace and level of instruction for high achievers and provide more individual attention, repetition, and review for low achievers. It is supposed to provide a spur to high achievers by making them work harder to succeed, and to place success within the grasp of low achievers, who are protected from having to compete with more able agemates (Atkinson & O'Connor, 1963).

The principal arguments against ability grouping have to do with the fact that this practice must create classes or groups of low achievers. These students are deprived of the example and stimulation provided by high achievers, and the fact of being labeled and assigned to a low group is held to communicate low expectations for students which may be self-fulfilling (see, for example, Good & Marshall, 1984; Persell, 1977).

Further, homogeneously low performing reading groups (Allington, 1980; Barr, 1975) and classes (Oakes, 1985; Evertson, 1982) have been observed to experience a slower pace and lower quality of instruction than do students in higher achieving groups. A lack of appropriately behaving models may lead to "behavioral contagion" among homogeneously grouped low achievers (Felmlee & Eder, 1983), so that these groups may spend less time on task than other groups.

However, perhaps the most compelling argument against ability grouping has little to do with its effects on achievement. This is that ability grouping goes against our democratic ideals by creating academic elites (Persell, 1977; Rosenbaum, 1976; Sorensen, 1970). According to this line of reasoning, all students need opportunities to interact with a wide range of peers. Because ability groupings often parallel social class and ethnic groupings, disproportionately placing low SES, Black, and Hispanic students in low tracks (e.g., Rist, 1970; Haller & Davis, 1980; Heyns, 1974), the use of ability grouping may serve to increase divisions along class, race, and ethnic group lines (see Rosenbaum, 1980).

Comprehensive Ability Grouping in the Elementary School

This review focuses on research on ability grouping at the elementary level. This restriction is made primarily because so many characteristics of elementary schools and the students they serve are unique to this level of schooling. Also, this review focuses on comprehensive ability grouping plans, which involve all students at particular grade levels.

This excludes studies of special classes for the gifted (e.g., Atkinson & O'Connor, 1963) and for low achievers. Gifted and special education programs may be conceived of as one form of ability grouping, but they also involve many other changes in curriculum, class size, resources, and goals that make them fundamentally different from comprehensive ability grouping plans. Further, non-randomized evaluations of gifted and special education/mainstreaming studies suffer from serious problems of selection bias which are less problematic in similar studies of comprehensive ability grouping plans (Slavin, 1984a; Madden & Slavin, 1983). For reviews of research on gifted and accelerated programs see J. Kulik & C. L. Kulik, 1984 or Passow, 1979; for special education/mainstreaming see Leinhardt & Pallay, 1982 or Madden & Slavin, 1983.

One of the most important characteristics of elementary schools for comprehensive ability grouping is that they tend to be small, rarely having more than three classes at each grade level. This means that if ability grouping is done within grade levels, the resulting reduction in heterogeneity may be slight. In fact, several studies (e.g., Clarke, 1958; Balow, 1962; Balow & Curtin, 1966; Goodlad & Anderson, 1963) have demonstrated that grouping students within grades into two or three homogeneous groups brings about a minimal reduction in total heterogeneity, particularly if grouping is done on the basis of IQ or general achievement.

Another important feature of elementary schools is that students are traditionally taught in self-contained classes, remaining with the same teacher all or most of the school day, and correspondingly teachers must attend to only one class. This situation is conducive to the use of <u>within-class</u> ability grouping (reading or math groups). In contrast, a secondary teacher with five or six sections is much less able to use within-class ability grouping, principally because the number of preparations required to teach twelve or more different subgroups would require superhuman effort.

A third characteristic of elementary schools is that while students at the elementary level have widely diverse skills from the first days of school, they are much less heterogeneous than are students at the secondary level (see, for example, Coleman, Campbell, Hobson, McPartland, Mood, Weinfeld, and York, 1966). Perhaps for this reason, between class ability grouping is far less universal in elementary than in secondary schools.

There are differences in the curriculum and goals of elementary and secondary schools which have an important bearing on ability grouping. By far the most important goal of the elementary school is to ensure that all students are able to read and compute.

Reading and mathematics are subjects that, at least in theory, lend themselves especially well to homogeneous groupings, as they are hierarchically organized subjects in which the learning of one skill depends on mastery of earlier skills. In a heterogeneous reading class it is unlikely that a single level of basal reader could be used, as it is probably unrealistic to expect low achievers to read and understand material a grade level or [illegible] above their reading level or to expect high achievers to prof[illegible]om material a grade level or more below their reading leve[illegible]ilarly, it is difficult to give an effective mathematics lesson to a class which includes some students who have not mastered subtraction, multiplication, or simple division, and some who have already mastered division or could do so very rapidly.

Comprehensive ability grouping plans used in elementary schools are adapted to the unique characteristics of that level of schooling, and often have no correlaries at the secondary level. There are three principal factors at issue in elementary ability grouping: 1) whether ability grouping is done within- or between-classes (or both); 2) whether between-class ability grouping is done to assign students to relatively ability-homogeneous self-contained classes, or is done only for selected subjects, with students remaining in heterogeneous classes most of their school day; and 3) whether ability grouping is restricted to one grade level or may combine students of similar performance level regardless of grade level.

Five comprehensive ability grouping plans predominate both in the literature and in practice: ability grouped class assignment, regrouping for reading and/or mathematics, Joplin and nongraded plans, and within-class ability grouping (reading or math groups within the class). The relationship between the three factors listed above and the five principal comprehensive grouping plans is depicted in Figure 1.

Figure 1 Here

It is important to note that other combinations of the same factors can produce additional grouping plans, and such plans have sometimes been studied. For example, ability grouped class assignment is usually done within grades, but Rankin, Anderson, & Bergman (1936) evaluated a "vertical grouping" plan in which upper elementary students were assigned to classes on the basis of IQ without regard to grade level lines. Cross-grade combination classes (e.g., 3-4, 4-5) often resemble vertical grouping, as students assigned to such classes are usually the higher achievers in the lower grade and the lower achievers in the higher grade. However, cross-grade groupings have rarely been studied, as they usually result from administrative needs for equal-sized classrooms rather than from a plan to improve school organization. For example, When a principal has forty-five fifth graders and forty-five fourth graders a combination 4-5 class is a likely result.

Also, it is important to note that many elementary classes use between <u>and</u> within-class grouping. For example, reading groups in the primary grades are virtually universal, regardless of whether or not the classes are grouped by ability. On the other hand, some between-class ability grouping plans (especially the Joplin Plan) are explicitly designed to reduce or eliminate the need for grouping within the classroom.

The following sections discuss the research on comprehensive ability grouping in elementary schools according to the four principal categories discussed above. Each section contains a table summarizing the principal studies on the ability grouping strategy being presented and a discussion of the studies and the methodological and substantive issues they raise. The criteria for inclusion of studies are presented below.

Criteria for Study Inclusion

The studies on which this review is based had to meet a set of *a priori* criteria with respect to germaneness and methodological adequacy. As stated earlier, all studies had to involve comprehensive studies of ability grouping in elementary schools (grades 1-6). European studies of eleven and twelve year olds who were in secondary schools (e.g., Douglas, 1973) are excluded, even though studies of students of the same age in elementary schools were included. Studies of within-class ability grouping were included, but other programs related to within-class grouping were excluded. Examples of such excluded programs are mastery learning, individualized and continuous-progress instruction, cooperative learning, multi-age grouping not done for the purpose of reducing student heterogeneity, open classrooms, and team teaching. No restrictions were placed on year of publication, and every effort was made to locate dissertations and other unpublished documents relating to ability grouping.

Methodological Requirements for Inclusion. One key element of best-evidence synthesis is the *a priori* establishment of inclusion criteria based on substantive and methodological adequacy. In the present case, criteria were established as follows:

1. Ability grouped classes were compared to heterogeneously grouped control classes. This requirement excludes several studies which compared achievement gains in experimental classes to "predicted" gains (e.g., Ramsey, 1962) and studies which correlated "degree of heterogeneity" with achievement gains without identifying classes as ability-grouped or heterogeneous (e.g., Leiter, 1983).

2. Achievement data from standardized achievement tests were presented. This excluded scores of anecdotal accounts and several studies of student or teacher attitudes toward ability grouping.

3. Initial comparability of samples was established by use of random assignment, matching of classes, or matching of students within equivalent classes. In cases of matching of classes or students, evidence had to be presented which established that the classes were in fact initially equivalent in IQ or achievement level (within 20% of a standard deviation). Studies in which experimental and control classes were not initially equivalent but gain scores or analyses of covariance were used to adjust scores for these differences (e.g., Moorhouse, 1964) are listed in tables in a separate category, and results of these studies should be interpreted cautiously.

Several cross-sectional studies that provided little evidence of initial equality were excluded. For example, Powell (1964) compared achievement scores of one school using the Joplin Plan to another using a self-contained model, with no evidence that the two schools were in fact comparable. Some studies (e.g., Hart, 1959) compared achievement under ability grouping to that under heterogeneous grouping in earlier years in the same schools. Such studies were included if there was evidence that the samples in the earlier years were equivalent in ability or achievement. However, comparison with previous classes was limited to two years, on the assumption that too many unrelated changes could take place over longer periods. This excluded one study that made a comparison over a ten-year period (Cushenberry, 1964) and restricted attention to the first two years of an eight-year study by Tobin (1966).

4. Ability grouping was in place for at least a semester. This requirement excluded only one very brief study (Piland & Lemke, 1971).

5. At least three experimental and three control teachers were involved in all included studies. The purpose of this requirement was to minimize the influence of teacher and class effects in small studies (see Slavin, 1984b) on study outcomes. This caused a few very small studies to be excluded (e.g., Johnston, 1973; Putbrese, 1972; Williams, 1966).

Literature Search Procedures

The studies reviewed here were located in an extensive search. Principal sources included the Education Resources Information Center (ERIC), Psychological Abstracts, Education Index, and Dissertation Abstracts. In these sources, the keywords "ability grouping," "classroom organization," "Joplin Plan," "nongraded," and related descriptions produced hundreds of citations. In addition, all citations in other reviews and meta-analyses were located, and citations made in primary sources were followed up. Every attempt was made to obtain a complete set of published and unpublished studies which met the substantive and methodological criteria outlined above. Further, in a few cases where clarifications were needed about important studies, authors were contacted directly for additional information.

Computation of Effect Sizes

Throughout this review, effects of various ability grouping strategies are referred to in terms of effect size. Effect sizes were generally computed as the difference between the experimental and control means divided by the control standard deviation (Glass, McGaw, & Smith, 1981). The control group was always the heterogeneous grouping plan unless otherwise noted, so that a positive effect size implies greater learning in an ability grouped plan and a negative value indicates an advantage for heterogeneous grouping. When means or standard deviations were omitted in studies which met inclusion criteria, effect sizes were estimated when possible from t's, F's, or exact p values (see Glass et al., 1981).

Many of the studies in this review presented data on gain scores without presenting pre- or posttest data. Effect sizes from achievement gain scores are typically inflated, as standard deviations of gain scores are less than those of pre- or posttest scores to the degree that pre-post correlations exceed 0.5. If pre-post correlations are known, effect sizes from gain scores can be transformed to the scale of posttest values using the following multiplier:

$$ES = (ES_{gain})(\sqrt{2(1-r_{pre-post})})$$

Because few studies presenting gain scores also provide pre-post correlations, a pre-post correlation of +0.8 was assumed. This figure is a characteristic correlation between fall and spring scores on alternate forms of the California Achievement Test in the upper elementary grades (CTB/McGraw-Hill, 1979). Substituting 0.8 in the formula, a multiplier of 0.632 is derived, which was used to deflate effect size estimates from gain score data. Because this value is only a rough approximation, effect sizes from gain score data should be interpreted with even more caution than is warranted for effect sizes in general.

In studies in which pretest data were provided, effect sizes were computed as <u>gain</u> scores, divided by the control group's post-test standard deviation. This procedure adjusts effect sizes for any differences in pretest scores. In a few cases, pretest and posttest scores were from different tests. In these studies (e.g., Flair, 1964), experimental-control differences divided by control standard deviations were computed for pre- and posttests, and the difference between these is reported as the study's effect size. Since all studies which met inclusion criteria presented either gain scores or pre- and post-test scores or matched on pretests, all effect sizes were adjusted for initial starting points.

If studies did not present enough data to allow for computation of effect size but otherwise met criteria for inclusion, they were included in tables with an indication of the direction and consistentcy of any achievement differences. In some cases only grade equivalent differences were given, and these are presented in the table. Because the standard deviations of grade equivalents are around 1.0 in upper elementary school, grade equivalent differences may be considered very rough approximations of effect size.

In general, one overall effect size is presented for each study, unless two or more different ability grouping plans were compared to heterogeneous control groups in the same study (e.g., Cartwright and McIntosh, 1972) or two distinct samples were studied (e.g., Borg, 1965). Multiple effects within a study were averaged to obtain the overall effect size estimate (see Bangert-Drowns, in press). If studies presented adequate data, overall effect sizes were also broken down by subject (e.g., reading, mathematics) and by achievement or ability level. Effect sizes by ability level should be interpreted with particular caution, as they are often inflated because standard deviations within subgroup categories are restricted in range.

In this best-evidence synthesis, every effort was made to make each effect size be a meaningful representation of the effect of ability grouping on student posttest achievement, holding the posttest standard deviation as the common metric. In all tables, randomized studies are listed first, followed by matched studies presenting evidence of initial equality between experimental and control groups, and then matched studies lacking evidence of initial equality. Within categories, studies with the largest samples sizes are listed first. These procedures mean that effect sizes from studies listed earlier in each table should generally be given more weight than those listed later.

However, it is important to remember that any effect size is only a rough indicator of the effect of a treatment. Many factors may influence effect size, such as differences in subjects, measures, experimental procedures, and study durations. Often, there are substantial nonsystematic differences in effect sizes for subgroups or for similar measures within the same study. For example, one study by Breidenstine (1936) had a mean effect size of -.08, but effects at particular grade levels ranged from -.89 to +.54. In another study by Slavin and Karweit (1985), effects of within-class ability grouping were +.64 for mathematics computations but .00 for mathematics concepts and applications. Had Breidenstine (1936) studied only one grade level or had Slavin and Karweit (1985) used only one mathematics achievement measure, their results would have appeared quite different. For these reasons, effect size data should always be interpreted cautiously, in light of the quality and consistency of the studies from which they were derived.

Research on Comprehensive Ability Grouping in Elementary Schools

Ability Grouped Class Assignment

A total of fourteen studies of comprehensive ability grouped class assignment plans were located.

The major characteristics and findings of the fourteen studies are summarized in Table 1. The randomized study is listed first, followed by matched studies in descending order of sample size.

Table 1 Here

Inspection of Table 1 clearly indicates that the effects of comprehensive ability grouped class assignment on student achievement are zero. The median total effect size across the seventeen comparisons in fourteen studies is exactly zero, and the effect sizes cluster closely around this value; of thirteen comparisons which yielded effect size data, eight fell in the range -.10 to +.10, and eleven in the range -.15 to +.15. Effect sizes for reading and mathematics did not exhibit any pattern different from that for overall effects. Further, little support appears in Table 1 for the assertion that high achievers benefit from ability grouping while low achievers suffer. Three studies (Borg, 1965; Flair, 1964; Tobin, 1966) found such a pattern, but three others (Bremer, 1958;

Hartill, 1936; Morganstern, 1963) found just the opposite, and Barker-Lunn (1970), Goldberg et al. (1966), Loomer (1962), and Rankin et al. (1936) found no differences according to achievement level.

The only randomized study of ability grouped class assignment is one by Cartwright and MacIntosh (1972), who compared three grouping methods in a school in Honolulu attended by disadvantaged students from a housing project. The students were ethnically diverse, and most came to school speaking Pidgin English and had to learn standard English as a second language. Students in grades 1-2 were randomly assigned to one of three treatments: Self-contained heterogeneous grouping, self-contained ability grouping, and flexible. The ability grouped students were assigned to relatively homogeneous classes according to intellectual ability and reading achievement without regard for grade level, so that the individual classes were somewhat heterogeneous in chronological age. The flexible classes were grouped for various subjects according to their performance level in those subjects, again without regard for grade level, and were frequently regrouped as their progress during the year warranted. All three treatments were begun when students entered the first and second grades and were continued for two years.

The dependent measures were scores on the Metropolitan Achievement Test. As shown in Table 1, the heterogeneous classes had higher scores in reading (ES = -.17) and in mathematics (ES=-.52) than the ability grouped classes. The heterogeneous classes also achieved more in reading than the flexible classes (ES =-.28), but there were no differences in mathematics.

Even though the Cartwright and MacIntosh (1972) study used random assignment to treatments, its results cannot be considered conclusive evidence against the use of ability grouped class assignment. First, there was only one class at each grade level in each treatment, restricting the possibilities for reducing heterogeneity by ability grouping even more than is usually the case in elementary ability grouping studies. Second, the population involved is quite atypical, and generalization to other settings, even other disadvantaged schools, is difficult.

Since there is only one randomized study of ability grouped class assignment and it has some important limitations, we must look at the best of the nonrandomized studies, those which used matching procedures to equate nonrandomly assigned groups and presented data to indicate that the groups were, in fact, initially equivalent in achievement or ability.

Three large, longitudinal studies done in the 1960's stand out in the study of ability grouped class assignment: Barker-Lunn's (1970) study of streaming in English and Welsh junior schools, the Goldberg et al. (1966) study of different grouping patterns in New York City schools, and Borg's (1965) study in two Utah school districts.

Of these three studies, the Goldberg et al. (1966) study is perhaps the most remarkable. This study involved eighty-six grade five classes in forty-five New York City elementary schools. Principals of all New York schools submitted Otis IQ distributions of their fourth grades. Only those schools with at least four students with

IQ's of at least 130 were included in the study sample, which had the effect (according to the authors) of restricting the sample to more middle-class parts of the city. The principals of the selecte schools were asked to assign students to classes for the fifth grade to conform to any of fifteen grouping patterns, ranging frcm extremely narrow to extremely broad. "Extremely narrow" classes included students falling within one IQ decile (e.g., 120-130), or those restricted to IQ 130 and up or 99 and below. "Extremely broad" classes included a full range of students from below 99 to above 130. Between these extremes were various moderately narrow and moderately broad patterns, classes containing students in two to four contiguous IQ deciles.

The principals were asked to keep students in the designated grouping patterns for two years, throughout grades 5 and 6, and only those students who were in the same schools for the entire two-year period (79% of the original sample) were included in the data analyses.

With classes in fifteen grouping patterns, Goldberg et al. were able to simulate many alternative grouping arrangements. For example, they could compare very homogeneous to very heterogeneous grouping plans by comparing the achievement gains of all students in one-decile classes to those in five-decile classes. They could simulate provision of special classes for the gifted by comparing five-decile (heterogeneous) classes to combinations of one-decile (130+) and four decile (less than 99 to 130) classes, and so on.

Unfortunately, Goldberg et al. do not present their actual achievement data, but only describe significant differences and patterns of findings. However, the patterns they present consistently favor broad, heterogeneous grouping plans for all students except for the most gifted (130+) who did equally well in broad- or narrow-range classes. Presence of gifted students was beneficial for the achievement of most students in most subjects, while the presence of low achievers was neither beneficial nor detrimental overall.

The Goldberg et al. (1966) study is arguably the best evidence in existence against the possibility that reductions in IQ heterogeneity can enhance student achievement in the upper elementary grades. The size and rigor of the experiment make it highly unlikely that any non-trivial positive effect of ability grouping could have been missed. While most achievement comparisons in the Goldberg et al. study were non-significant, the patterns of mean differences and of those differences which were statistically significant support heterogeneous rather than ability grouped class assignments.

The Barker-Lunn (1970) study in England and Wales similarly provides little support for ability grouped class assignment. This study compared the achievement gains of students in 36 streamed junior schools (serving students aged 7 to 11) and 36 unstreamed schools, matched on social class. The streamed schools had a slight advantage in achievement after one year, so the initial comparability of the samples was questioned by the author; the year-to-year scores and four-year longitudinal comparisons used first-year scores as covariates to control for this initial difference.

As in the case of the Goldberg et al. (1966) study, Barker-Lunn presents only statistically significant differences and trends. Overall, there were no meaningful trends favoring streamed or non-streamed schools. All but a few comparisons were non-significant, and those which were significant were equally likely to favor streamed or unstreamed schools. Again, if there were any consistent effect of ability grouped class assignment on student achievement, a study the size and quality of Barker-Lunn's would be very likely to find it.

The third large, longitudinal study is one by Borg (1965), who compared achievement gains in two adjacent districts in Utah, one of which used heterogeneous grouping. Two elementary cohorts were studied. One began in the fourth grade and was followed through grade seven. Another began in the sixth grade and was followed through grade nine. The results for the sixth grade sample presented in Table 1 are only for the first year, as these students went on to junior high school beginning in grade seven.

Even though this study used as large a sample as Goldberg et al. (1966) and Barker-Lunn (1970) and was also carefully controlled, the nature of the samples involved make the results of this study less conclusive. First, only two districts were involved, and any differences between the districts other than the use of ability grouping are completely confounded with grouping practice. One district served a small city, while the other served its outlying area, so unmeasured population differences may have been operating. Second, while the two districts' mean pretest scores were equal within each IQ category, the proportion of high-IQ students was higher in the district using heterogeneous grouping, particularly in the sixth grade sample. Third, the districts involved had been using their respective grouping methods for many years, so that the students being studied had been in ability grouped or heterogeneous classes for three years (the grade four sample) or five years (the grade six sample). Any effects of grouping may have already been registered before the study began.

The results of the Borg (1965) study were inconsistent, but in general the longitudinal data following fourth graders indicated that ability grouping was beneficial for the achievement of high-IQ students, detrimental for that of low-IQ students, and neutral for average-IQ students. After one year, high- and average-IQ students scored higher in ability-grouped than in heterogeneous classes, but by the seventh grade this difference had disappeared. High- and average achieving sixth graders gained more in ability grouped than in heterogeneous classes, but these differences also dissipated in junior high school.

One well designed (but fifty year old) study by Hartill (1936) compared ability grouped to heterogeneous class assignment in fifteen New York City schools. Students in grades five and six were assigned to ability grouped or heterogeneous classes for one semester, and were then reassigned to classes according to the opposite grouping pattern for a semester. Not only were the students their own controls (since they experienced both grouping plans) but they were individually matched with one another, so that the group which experienced ability grouping first was identical in IQ to that which experienced heterogeneous grouping first. The results indicated that low-IQ students achieved slightly better in ability grouped classes (ES = +.18), high-IQ students achieved slightly better in heterogeneous classes (ES = -.12), and average-IQ students achieved equally well in the two grouping plans. Overall, achievement gains were identical in the ability grouped and heterogeneous classes.

Another important early study was one by Rankin et al. (1936), who compared students matched on achievement level in three programs. One was traditional ability grouped class assignment done within grade levels, except that in mathematics these classes used a program essentially identical to modern group-paced mastery learning (see Block & Anderson, 1976). Another, called "vertical grouping," assigned students to classes according to their level of achievement without regard for their grade level. This procedure produced such homogeneous classes that reading groups within the classes were considered unnecessary. The third plan involved heterogeneous grouping of classes, with the additional requirement that within-class ability grouping (including use of reading groups) was not allowed. Teacher and administration attitudes toward this heterogeneous plan were quite negative, as the degree of heterogeneity in these classes was great and teachers were unable to use any form of grouping to accommodate student differences. However, achievement differences between the two ability grouped plans and the heterogeneous classes were small for the ability grouped plan (ES = +.05) and for the "vertical" plan (ES = +.07).

A four-year study of streamed and unstreamed junior schools in England by Daniels (1961) compared two pairs of schools. The schools themselves were selected from the same districts and were found to have nearly identical mean IQ's. Students were individually matched on IQ within the schools. After three and one-half years, the students in the unstreamed schools were achieving at a significantly higher level than those in the streamed schools (ES = -.26). However, other multi-year studies (Breidenstine, 1936; Tobin, 1966) found effects near zero, and Morganstern, who followed students from the fourth to the sixth grade, found a small benefit of ability grouping (ES = +.15).

Some authors (e.g., Good and Brophy, 1984) have suggested the use of a modified ability grouping plan in which high and average achievers are mixed and average and low achievers are mixed. However, a study by Loomer (1962) found no achievement benefits of such a plan (ES = -.04).

In addition to the studies listed in Table 1, a few studies have correlated the degree of heterogeneity in classes with student achievement. Justman (1968) found that the reading achievement of third graders increased slightly more in heterogeneous than homogeneous classes, with average and low achieving classes gaining the most from heterogeneity. Leiter (1983) found no correlation between class homogeneity and third grade reading and mathematics achievement, controlling for the previous year's scores, although there was a nonsignificant trend toward higher reading achievement and lower mathematics achievement in more homogeneous classes. Edminston and Benfer (1949) divided sixteen classes of fifth and sixth graders into classes with wide and narrow IQ ranges. Over six months, students in the wide range classes gained significantly more in composite achievement than did students in narrow range classes.

__Summary and Discussion: Ability Grouped Class Assignment__. Given the persistance of the practice over time and the belief teachers typically place in its effectiveness, it is surprising to see how unequivocally the research evidence refutes the assertion that ability grouped class assignment can increase student achievement in elementary schools. There is a considerable quantity of good quality research on this topic, such that any impact of grouping on achievement would surely have been detected.

Several earlier reviews have made the claim that ability grouping is beneficial for high-ability students and detrimental for low-ability students (e.g., Eash, 1961; Esposito, 1973; Begle, 1975). This claim is not clearly supported by the present review. It is possible that a clearer pattern emerges in secondary studies, but it is more likely that confusion arises when studies of special programs for the gifted and for low achievers are included in ability grouping reviews. Studies of special programs for the gifted tend to find achievement benefits for the gifted students (J. Kulkik and C. L. Kulik, 1984; Passow, 1979), while studies of mainstreaming vs.

special education for students with learning problems tend to favor regular class placement (Madden & Slavin, 1983). For this reason, including studies of special programs for the gifted and learning disabled in reviews of ability grouping, as was done by Begle (1975), Borg (1965), Findley & Bryan (1970), and others, would give the impression that ability grouping is beneficial for high achievers and detrimental for low achievers. However, it is likely that characteristics of special accelerated programs for the gifted account for the effects of gifted programs, not the fact of separate grouping per se (see Fox, 1979). Also, problems of selection bias in nonrandomized studies of programs for the gifted and for students with learning problems bias the results of these studies toward the higher placement, spuriously favoring separate programs for the gifted and mainstream placement for low achievers (see Borg, 1965; Slavin, 1984a).

Regrouping for Reading and Mathematics

In many elementary schools, reading and/or mathematics is scheduled at the same time for all students in a particular grade. At that time, students leave their heterogeneous homeroom classes to receive reading or mathematics instruction in a class that is more homogeneous in the skills in question.

Previous reviews and meta-analyses (e.g., Borg, 1965; Findley & Bryan, 1970; C. L. Kulik and J. Kulik, 1984) have not made a clear distinction between regrouping and ability grouped class assignment. Yet there are several important theoretical reasons to do so.

First, regrouping minimizes the presumed negative psychological effects of ability grouping. Students spend most of the day in heterogeneous classes, with which they almost certainly identify. Second, regrouping is always done on the basis of actual performance in reading or mathematics, not on IQ, and is usually flexible, so that any errors in assignment or changes in student level of achievement can be easily accommodated by moving students to different sections. For these reasons, it is likely that regrouping can produce much more homogeneity in the skills being taught than can ability grouped class assignment, which is usually based on IQ or general achievement and is relatively inflexible.

Unfortunately, there is neither the number nor the quality of studies of regrouping to enable definitive conclusions concerning the effectiveness of such plans. Only three studies used matching and presented evidence of initial equality. Four additional studies lacked evidence of initial equality but did adjust posttests for pretests and other variables. Overall, five of the seven studies found that students learned more in regrouped than in heterogeneous classes, while two found the opposite trend. (See Table 2.)

Table 2 Here

Two of the studies investigated the practice of regrouping for reading only. One of these was a large study by Moses (1966) involving 54 classes in rural Louisiana. This carefully controlled study held constant time and instructional materials in matched experimental and control classes. No consistent differences were found in reading achievement. However, a study by Berkun, Swanson, & Sawyer (1966) did find significantly greater gains for regrouped than for self-contained reading classes (ES = +.32). However, this article provides few details of the treatment procedures, and may suffer from pretest differences between the experimental and control groups (all data presented are posttests adjusted for pretests).

A study by Provus (1960) of regrouping for mathematics provides the best evidence in favor of this practice. Experimental students in eleven classes in a suburb of Chicago were regrouped from their heterogeneous homerooms into relatively homogeneous mathematics classes at the same grade level. Achievement gains for students in these classes were compared to those of students matched on IQ who remained in heterogeneous classes all day. One effect of the regrouping was to allow high achievers to be exposed to material far above th .ade level; there were cases of fourth graders finishing the ye working on eighth grade material. Perhaps for this reason, achievement gains for high ability students in the regrouping program were much greater than those of comparable control students (ES = +.79), but the program was less spectacularly beneficial for average ability (ES = +.22) and low ability students (ES = +.15).

In contrast, Davis & Tracy (1963) found that regrouping for mathematics was detrimental to the achievement of students in a rural North Carolina town. However, this study compared only two schools and there were substantial achievement differences at pretest. Also, it is important to note that no attempt was made to provide differentiated materials to the regrouped classes; all classes used grade level-appropriate texts.

Finally, three studies investigated the effects of regrouping in multiple subjects. In a study by Koontz (1961) in Norfolk, Virginia, experimental students were separately grouped according to their achievement in reading, mathematics, and language. At other times students remained in "intact classes," but it is unclear whether these were ability grouped or heterogeneous. This method approaches a departmentalized arrangement, as students changed classes three or four times each day. Its effects on all three subjects involved turned out to be negative, particularly for reading, where the heterogeneous, self-contained classes gained .42 grade equivalents more than the regrouped students. A study by Balow and Rudell (1963) evaluated regrouping for reading and math, and found positive effects in both subjects for average and low achievers. However, pretest differences favoring the experimental (regrouped) classes throw some doubt on these findings.

Finally, Morris (1969) studied a program in which regrouping was done for reading and math. The program was called a "nongraded primary plan" by the author, but since regrouping was done within grade levels, it was categorized as a regrouping program. Overall student achievement at the end of three years was higher in the regrouped classes than in heterogeneous control groups, controlling for IQ (ES = +.43). After two more years during which all students, experimental as well as control, were in a regrouping plan, the former experimental students had greatly increased their advantage over the control group (ES = +1.20).

<u>Summary and Discussion</u>: <u>Regrouping for Reading and Mathematics</u>. Overall, the results of studies of regrouping for reading and mathematics are inconclusive. None of the grouping patterns evaluated were consistently successful, although one study (Provus, 1960) gave strong evidence favoring the use of regrouping in mathematics if students are given materials appropriate to their levels of performance. Another study (Morris, 1969) found strong positive effects of regrouping for reading and mathematics. This study also emphasized adaptation of the level of instruction to accommodate student differences. In contrast to the situation with ability grouped class assignment, where there is adequate high quality evidence to conclude that no important effects of ability grouping exist, it is still quite possible that regrouping for one or two subjects is instructionally effective, and evidence from studies of Joplin and nongraded plans, summarized in the following section, provides some support for this possibility. However, more research is needed to establish the achievement effects of regrouping within grade levels.

Joplin and Nongraded Plans

The Joplin plan (Floyd, 1954) is in its simplest form an extension of regrouping for reading to allow for grouping by reading level across grade level lines. This practice typically creates reading classes in which all students are working at the same or at most two reading levels, so that within-class ability grouping may be reduced or eliminated. The tradeoff between within-class (reading groups) and between-class (Joplin) ability grouping is a pivotal issue in studies of the Joplin plan, which may be conceptualized not as ability grouping versus heterogeneous grouping but as between- versus within-class grouping (see, for example, Newport, 1967). In contrast, studies of regrouping for reading within grades maintain reading groups within the class, although there may be some reduction in the number of reading groups used.

Nongraded plans share with the Joplin plan the idea of grouping students according to performance level in a specific skill, ignoring grade level or age. Some forms of nongraded grouping are very similar to the Joplin plan, except that they are applied in the primary rather than intermediate grades and have been utilized in subjects other than reading. Some nongraded plans incorporated the practice of allowing students to spend two or four years in the primary grades if their progress warranted acceleration or additional time, respectively, but it is unclear how often students actually deviated from the three-year norm. One study (McLoughlin, 1970) found that students in nongraded plans hardly ever completed the primary grades in less than three years and took four years to do so no more often than did students in graded classes.

Table 3 Here

Table 3 summarizes the research on the Joplin Plan and Joplin-like nongraded plans. The Table includes several studies (e.g., Hillson et al., 1964) whose experimental groups were described as nongraded or ungraded but which more closely resembled the Joplin Plan, in that only one subject (usually reading) was involved. The studies listed in Table 4 evaluated more comprehensive nongraded pla involving several subjects and such additional features as team teaching, individualized instruction, and learning centers. However, it should be noted that the division of studies of nongraded plans into Tables 3 and 4 is not exact. Several studies do not adequately describe their nongraded plans, and others vary on a continuum from completely Joplin-like (e.g., Skapski, 1960) to highly complex flexible grouping plans (e.g., Bowman, 1971). As nongraded plans incorporate more of the features proposed by Goodlad and Anderson (1963), they cease to be just ability grouping plans, but come to resemble forms of the open classroom (Giaconia and Hedges, 1982) or of Individually Guided Education (Klausmeier, Rossmiller, & Saily, 1977).

Overall, the evidence in Table 3 strongly supports the use of the Joplin Plan. Joplin classes achieved more than control classes in eleven of fourteen comparisons, with the remaining three studies finding no differences. The median effect size for these studies is about +.45. Further, the quality of these studies is high, with two using random assignment and ten using matching with good evidence of initial equality.

Morgan and Stucker (1960) conducted a randomized study of the Joplin Plan in rural Michigan. Fifth and sixth grade students were matched on reading achievement and then randomly assigned to four Joplin and four control classes. Teachers were also randomly assigned to treatments. Because there were only two Joplin classes at each grade level, the amount of cross-grade grouping that could be done was limited, and control groups were ability grouped (within grade), yet the authors still document a considerable reduction in class heterogeneity as a result of cross-grade assignment.

Results indicated significantly higher achievement in the Joplin Plan for high and low achievers in fifth grades and low achievers in the sixth grades. The authors explain the failure to find experimental-control differences for high achieving sixth grades by noting that because of the small number of classes involved in the study, high achieving sixth graders could not be accelerated as much as would have been possible with larger numbers of classes. Whatever the explanation, larger experimental-control differences for low achievers (ES = +.94) than for high achievers (ES = +.32) are entirely due to the lack of differential gains for high achieving sixth graders.

Hillson, Jones, Moore, and Van Devender (1964) studied a non-graded program similar to the Joplin Plan. They randomly assigned students and teachers to nongraded or traditional classes. Students in the nongraded classes were assigned to heterogeneous classes but regrouped across grade levels for reading. They proceeded through nine reading levels, and were continually regrouped on the basis of their reading performance. Within each reading class teachers had multiple reading groups and used traditional basal readers and instructional methods (J. W. Moore, Personal Communication, January 23, 1986).

The results of this study supported the efficacy of the nongraded program. After three semesters, reading scores for experimental students on three standardized scales were considerably higher than for control students (ES = +.72, or about .41 grade equivalents). After three years in the program, experimental-control differences had diminished, but were still moderately positive (ES = +.33) (Jones, Moore, and Van Devender, 1967).

Ten studies compared Joplin or Joplin-like nongraded classes to matched control classes and presented evidence of initial comparability. The largest of these (Russell, 1946) was done before Floyd (1954) first described the Joplin plan, but evaluated a very similar intervention. Students in grades 4-6 were regrouped for reading without regard to grade level. This created relatively homogeneous groups, but homogeneity was increased still further by the use of reading groups within the classes (usually two) and by reviewing and modifying group assignments four times per year. Students in this plan, called "circling," were matched with students in other schools which did not regroup for reading and followed for two years, from the beginning of grade 4 to the beginning of grade 6. Results indicated no differences between the two types of grouping plans (ES = .00).

It is interesting to note that the only other matched equivalent study to find no advantage for the Joplin plan also used reading groups within the regrouped reading classes. This was a study by Carson and Thompson (1964), in which students in grades 4-6 were regrouped across grade lines for reading but were still assigned to reading groups within their reading classes. These students' gains in reading achievement were compared to those of students assigned by ability (within grade) to self-contained classes.

The eight remaining matched equivalent studies all found positive effects of Joplin or Joplin-like nongraded plans on student achievement. For example, Green and Riley (1963) compared the Joplin Plan to the traditional methods in use in the same schools during the previous year. Students in the Joplin classes gained significantly more in reading achievement than did students in the earlier years (ES = +.36).

In a study by Hart (1959), grade 4-5 students were regrouped into nine reading classes. Seven of these had only one reading level, one had two, and one had four very low achieving groups (but may have used fewer than four reading groups in the classroom). The top class was reading at the seventh grade level, and the bottom class contained students ranging from primer to second grade, second semester. Students' scores were compared to those of students taught by the same teachers the previous year. Gains on the California Achievement Test strongly favored the Joplin approach (ES = +.89). In both the fourth and the fifth grades, Joplin groups gained about a full grade equivalent more than did heterogeneously grouped classes in earlier years.

Rothrock (1961) also compared Joplin Plan classes to heterogeneous classes which used within-class ability grouping, and found significantly positive effects on student reading achievement and work-study skills, averaging .44 grade equivalents more than in heterogeneous classes. An individualized reading program fell between the Joplin and heterogeneous programs in achievement effects. Green and Riley (1963) found consistently greater reading achievement gains in Joplin Plan classes than in matched heterogeneously grouped classes in different schools (ES = +.36). Every experimental class gained significantly more than its corresponding control class, and the average experimental-control difference in grade equivalents was .54. Anastasiow (1968) found no significant differences between a Joplin-type regrouping plan and heterogeneous grouping, but the trends favored the Joplin Plan groups (ES = +.15).

Moorhouse (1964) compared a school using the Joplin Plan to a heterogeneously grouped control school. The students in grades 4-6 were grouped in seven reading classes. Three classes contained one reading level, three contained two, and one (the lowest) contained students from five levels, with a range from first to third grade levels. The top class had students working at the seventh and eighth grade level. The authors note that three quarters of all students in the Joplin classes were working at a level different from the ones usually used at their grade level. Unfortunately, the results of the Moorhouse study are marred by pretest differences favoring the control groups in grades 4 and 6. However, at all three grade levels (including grade five, where there were no pretest differences), Joplin classes gained considerably more in reading achievement than heterogeneous control classes, averaging gains of 1.24 grade equivalents in one semester, more than twice the gains seen in control classes (.61 GE).

Experimental and control classes were followed for a total of five semesters. By the end of sixth grade, fourth graders had gained a total of .50 grade equivalents more than control. Fifth graders had gained about .40 grade equivalents by the end of sixth grade, but lost this advantage by the eighth grade. Sixth graders, who made the greatest gains initially, maintained most of that gain through the eighth grade. Overall, the patterns of results indicate that achievement gains due to the Joplin Plan were primarily seen early in the program implementation and then diminished as students entered the junior high school.

Ingram (1960) evaluated a nongraded program very similar to that studied by Hillson et al. (1964). Students in grades 1-3 were assigned to one of nine reading levels without regard to age. As students moved from grade to grade, they picked up where they had left off in the previous year. Teachers generally had more than one reading group within their reading classes. As in the Hillson et al. (1964) study, the results strongly supported the nongraded approach. By the end of three years, students in the nongraded program were achieving approximately .7 grade equivalents ahead of similar students in earlier years before nongrading (ES = +.55) on standardized reading, spelling, and language tests.

Halliwell (1963) evaluated a nongraded primary program that was virtually identical to the Joplin Plan. Students in grades 1-3 were regrouped for reading only, and remained in heterogeneous classes the rest of the day. Spelling was also included in the regrouped classes for second and third graders. The article is unclear as to whether within-class grouping was used in regrouped reading classes, but there is some indication that reading groups were not used. Results indicated considerably higher reading achievement in nongraded classes than in the same school the year before nongrading was introduced (ES = +.59). Scores were higher for nongraded students at every grade level, but by far the largest differences were for first graders, who exceeded earlier first grade classes by .94 grade equivalents (ES = +1.22).

It is important to note that mathematics achievement, measured at the second and third grade levels, also increased significantly more in the nongraded classes than in previous years (ES = +.51). Since mathematics was not part of the nongraded program, this finding suggests the possibility that factors other than the nongraded program might account for the increases in student achievement. However, the author notes that teachers claimed to have been able to devote more time to mathematics because the nongraded program required less time for reading, spelling, and language instruction than they had spent on these subjects in previous years.

A study by Skapski (1960) also evaluated the use of nongraded organization for reading only. The details of the nongraded program were not clearly described, but it appears that reading groups were not used within regrouped classes and that curricula and teaching methods were traditional. Two comparisons were made. First, the reading scores of students in the nongraded program were compared to the same students' arithmetic scores, on the assumption that since arithmetic was not involved in the nongraded plan any differences would reflect an effect of nongrading. Results of this comparison indicated that second and third grade-aged students achieved an average of 1.1 grade equivalents higher in reading than in arithmetic.

Further, scores of third graders who had spent three years in the nongraded program were compared to those of students in two control schools matched on IQ. Results indicated that the nongraded students achieved at a much higher level in reading than did control students (ES = +.57), but there were no differences in arithmetic. Differences were particularly large for students with IQ's of 125 or higher (ES = +.97), but were still quite substantial for students with IQ's in the range 88-112 (ES = +.52).

Only one study evaluated the use of a nongraded program in mathematics. This study (Hart, 1962) took place in the same school which evaluated the Joplin Plan in reading in its intermediate grades (Hart, 1959). Experimental students were regrouped for arithmetic instruction across grade lines, and were taught as a whole class. Students were frequently assessed on arithmetic skills and reassigned to different classes if their performance indicated that a different level of instruction was needed. Experimental students who had spent three years in the nongraded arithmetic program were matched on IQ, age, and socioeconomic status with students in similar schools using traditional methods. It is not stated whether control classes used within-class ability grouping for arithmetic instruction. Results indicated an advantage of about one-half grade equivalent for the experimental group (ES = +.46).

Summary and Discussion: Joplin Plan. Considered together, the results of the Joplin and Joplin-like nongraded plans are remarkably strong. Both randomized studies found positive effects on student achievement, as did all but two of the ten matched equivalent studies.

Only four studies presented results according to student ability levels. Morgan and Stucker (1960) found stronger positive effects of the Joplin plan for low than for high achievers, while Moorhouse (1964) found the largest gains for high and average achievers, and Kierstead (1963) found no experimental-control differences at any level of ability. Skapski's (1960) results indicated that very able students benefited the most from a Joplin-like nongraded reading program. In no case did one subgroup gain at the expense of another; either all ability levels gained more than their control counterparts or (in the case of the Kierstead study) none did.

Comprehensive Nongraded Plans

Throughout its history, the concept of "nongradedness" has been presented as an ideal to which schools may aspire rather than as a as a specific program which they may implement. Many of the studies of nongraded plans, especially of the Joplin-like variations, apologize for their failure to fully live up to the "nongraded" ideal. As noted earlier, implementations of programs described as nongraded have ranged from simple regrouping plans for reading to very complex interventions. For example, Carbone (1961, p. 88) poses the following "six questions to be considered in discussing the concept of nongrading:

1. Do we have clear statements of our instructional objectives organized in a realistic sequence and covering the entire span of our program? (Objectives)

2. Do we have a sufficient variety of instructional materials on different levels of sophistication so that each teacher can adjust instruction to the range of abilities found in each classroom? (Instructional materials)

3. Are we able to move toward greater individualization of instruction so that pupils can actually progress at individual rates (Individualized instruction)

4. Are we willing to use grouping practices that are flexible enough to allow easy movement from group to group within a class and from class to class within a school? (Grouping practices)

5. Do we have evaluation devices, based on our instructional objectives, that will provide clear evidence of pupil attainments and thus facilitate our decisions on grouping and progress? (Evaluation devices)

6. Are we sufficiently committed to that educational shibboleth -- recognizing individual differences -- to do something about the the differences that we have so long only "recognized"? (Human factors)"

Examples of recommended practices grouped under these six questions include use of self-teaching and self-testing materials, independent study on projects appropriate to students' interest, abilities, and needs, use of independent study or instruction to very small groups (2-6 students) at least two-thirds of the day, and variable amounts of time in which students graduate. In this ideal form, the nongraded elementary school is closer in conception to individualized instruction or to the open school than it is to the Joplin Plan, which does not use individualized instruction and reduces or eliminates within-class ability grouping.

Unfortunately, studies of nongraded programs often do not specify exactly what was implemented. The studies reviewed in this section evaluated nongraded plans which clearly involved several subjects and a comprehensive approach to nongrading as well as a few in which the nongraded plan was only briefly described.

Overall, the studies of comprehensive nongraded plans are less consistent in finding benefits of these programs than are studies of Joplin-like nongraded plans, but the median effect size is still moderately positive (ES = +.29). However, there is a tendency for the higher-quality studies to produce larger effect sizes than the lower-quality ones. For example, a large matched-equivalent study by Hickey (1963) found that students in nongraded primaries in seven Catholic schools learned significantly more after three years than did students in similar graded schools (ES = +.46). Similar results were obtained in matched equivalent studies by Buffie (1962; ES = +.35), Remacle (1971; ES = +.31 grade equivalents), and Machiele (1965; ES = +.50). However, in none of these studies were the nongraded programs clearly described.

Brody (1970) evaluated a nongraded program in which first and second graders had to pass a series of sequential steps in several subjects at 90% mastery, and were placed in groups according to their mastery of specific skills (regardless of grade level). Vertical advancement of students was strongly emphasized. At the time of assessment, first graders had been in this program one year and second graders two years, but both groups of students gained significantly more than did students matched on IQ in graded classes (ES = +.28). Effects were particularly large in mathematics (ES = +.52). This study was somewhat flawed by the fact that before matching, the nongraded students were 5.4 points higher in IQ than their graded counterparts.

The only matched equivalent study to find no differences in achievement between nongraded and graded programs was one by Otto (1969), which took place in a laboratory school at the University of Texas. Unlike most of the studies of comprehensive nongraded plans, the Otto study fully described the nongraded intervention, which was designed to be a full-scale implementation of the Goodlad and Anderson (1963) nongraded plan.

Unfortunately, experimental and control groups did not differ on many elements held to be essential to the nongraded program. Teachers of the nongraded classes did assign students to instructional groups across grade lines and did have students use more individualized materials and provided less whole-class instruction than did teachers in the graded program. However, the nongraded classes did not use more subgroups than graded classes and did not reduce the heterogeneity of subgroups. Because the experiment took place in a laboratory school, it may be that control classes were of high quality and control teachers may have used many aspects of nongrading in their classes. It is interesting to note that another study in a university laboratory school by Ross (1967) also failed to find differences between nongraded and graded classes (ES = +.06 GE).

Perhaps the most comprehensive nongraded program evaluated outside of a university laboratory school was one studied by Bowman (1971), in which individualized instruction, team teaching, flexible grouping, and learning centers were used. This one-year study found strong positive effects on the achievement of interediate students (ES = +.52) but not of primary students (ES = .06). Killough (1972) also found significantly positive effects of a comprehensive nongraded program implemented in an open-space school, although the details of the intervention are not described.

The only study to find higher achievement in graded than ungraded schools was also perhaps the lowest in methodological quality. This is a study by Carbone (1961) which compared the achievement of students in traditional graded schools to those in schools mentioned by Goodlad and Anderson (1959) as nongraded, controlling for IQ scores. The students involved were in grades four, five, and six, which is to say one, two, or three years (respectively) after their experience in the nongraded primary. Further, there were substantial IQ differences between the two sets of students, and teacher questionnaires indicated very few differences between the two sets of teachers in reported classroom practices.

Another study, by Hopkins, Oldridge, and Williamson (1965), found no achievement differences between graded and nongraded classes controlling for IQ. The article does not specify the nongraded program used, but the first author (K. D. Hopkins, Personal Communication, February 12, 1986) indicated that the nongraded teachers were "trying to implement nongrading as suggested by Goodlad and Anderson."

It is interesting to note that, probably because it appeared early in the nongrading movement, the Carbone (1961) study was taken by several revieweres of this literature as a serious counterweight to the positive findings of other studies. For example, this was the only negative evidence cited in a review done by the National Education Association (1967), yet the review concluded that "no conclusive data favoring nongraded organization over the graded or graded over the nongraded can be found in studies made so far, but the preponderance of studies appears to be favorable (page 168).

<u>Summary and Discussion</u>: <u>Comprehensive Nongraded Plans</u>. Overall, the data from studies of comprehensive nongraded plans supports the use of this grouping plan (also see Pavan, 1973). Excluding studies done in university laboratory schools and the seriously flawed Carbone (1961) study, the median effect size rises to about +.33. Two studies (Hickey, 1963; Buffie, 1962) found that the effects of nongraded programs were particularly positive for high achievers, and Bowman (1971) found that older students benefitted more than younger ones. It may be that students need a certain level of maturity or self-organizational skills to benefit from a continuous-progress program which includes a good deal of independent work.

Within-Class Ability Grouping

Research on within-class ability grouping differs from that on between-class ability grouping in many important ways. First, this research is considerably more likely to use random assignment than is research on between-class ability grouping. Five randomized studies met the criteria for inclusion applied in this review (one additional study, by Putbrese (1972), also used random assignment but was omitted because it had only one experimental and one control class). Second, the duration of within-class ability grouping studies is shorter, with most studies lasting about one semester.

Table 5 here

Research on within-class ability grouping is summarized in Table 5. Every study which met the inclusion criteria involved the use of math groups, although Jones (1948) also studied grouping in reading and spelling. The lack of studies of grouping in reading is surprising. It may be that this practice is so widespread that formation of ungrouped control groups is difficult to arrange, even on an experimental basis.

Every study of within-class ability grouping in mathematics favored the practice, though not always significantly. The median effect size for the five randomized studies is +.32; including matched studies makes the median only slightly higher. Five studies broke down achievement effects by student ability levels, but there is no clear pattern of results favoring ability grouping for one or another subgroup. Every subgroup gained more in classes using within-class ability grouping than in control (ungrouped) treatments. However, it is interesting to note that the median effect size for low achievers (ES = +.65) was higher than that for average (ES = +.27) or high achievers (ES = +.41).

Slavin and Karweit (1985) conducted two large randomized studies of within-class ability grouping, one in highly heterogeneous, racially mixed schools in Wilmington, DE (Experiment 1), and one in relatively homogeneous, predominately white schools in and around Hagerstown, MD (Experiment 2).

In Experiment 1, grade 4-6 classes were randomly assigned to one of three treatments. One, an individualized model, is not considered here. A second was a whole-class instructional model called the Missouri Mathematics Program (MMP), which had been found in earlier research (Good & Grouws, 1979) to be more effective than traditional whole-class instruction. The MMP, based on the findings of studies of the practices of outstanding elementary mathematics teachers (e.g., Good & Grouws, 1977), uses a regular sequence of teaching, controlled practice, independent seatwork, and homework, with an emphasis on a high ratio of active teaching to seatwork, teaching mathematics in the context of meaning, and management strategies intended to increase student time on-task (Good, Grouws, &

Ebmeier, 1983). The third treatment was a form of within-class ability grouping based on the MMP, but using two ability groups. Specific management strategies were incorporated in this model to help teachers deal with the problem of having one group working independently while the other was receiving instruction. Details of this treatment are presented in Slavin & Karweit (1983).

Results of the semester-long study indicated significantly higher achievement in ability grouped than in whole-class instruction (ES = +.32). Effects were large for mathematics computations (ES = +.64), but there were no differences in concepts and applications (ES = .00).

Experiment 2 involved the same comparisons, except that an untreated control group was also added. The results very closely parallelled those of Experiment 1; students in the ability grouped model achieved significantly more than those in the two whole-class methods (ES = +.27), with differences larger in mathematics computations (ES = +.37) than concepts and applications (ES = +.17).

Dewar (1964) randomly assigned sixth grade mathematics classes and their teachers in a suburb of Kansas City, KS to use within-class ability grouping or whole-class instruction for a full school year. Three math groups were used in the grouped classes. Results strongly favored the grouped classes for students who had been in the top, middle, and low groups in comparison to their counterparts in the control group (ES = +.55). In a similar study, Smith (1960) randomly assigned grade 2-5 classes to grouped or control conditions for five months. Within each grade level students in experimental and control groups were matched on sex, age, achievement level, and IQ. Results favored within-class ability grouping (ES = +.41), especially for students assigned to the lowest group (ES = +.69).

The smallest positive effect of any study of within-class ability grouping was reported in a study by Wallen & Vowles (1960), who randomly assigned four sixth grade mathematics classes to ability grouped or whole-class treatments for one semester and then had all classes experience the opposite treatment for one semester. After one semester the scores of the ability grouped students were higher than those of the control students (ES = +.30), but second semester results nearly wiped out this difference (ES = +.07). It is important to note that this was the only study to use four math groups rather than two or three.

Three nonrandomized studies generally supported the results of the randomized ones. Spence (1958) compared the achievement gains of students in mathematics classes using within-class ability grouping to that of students in control classes matched on IQ and arithmetic achievement. Results indicated significantly greater gains for the grouped students. Stern (1972) compared the achievement of low achievers in classes using math groups to that of students matched on achievement pretests. Despite matching, pretest differences favored students in the control conditions, but gain scores clearly favored the grouped classes (ES = +.36). Note that these low achievers were not in homogeneously low classes, but were selected from among heterogeneous math classes using or not using within-class ability grouping.

A study by Jones (1948) also compared matched students in different grouping arrangements. The experimental treatment used in this study is not well specified, but did involve within-class grouping of some kind for reading, spelling, and mathematics. Results favored grouped classes for all three subjects and for three levels of ability (overall ES = +.26). Information adequate for computation of effect sizes was only provided for composite achievement.

Summary and Discussion: Within-Class Ability Grouping. Research on the use of math groups consistently supports this practice in the upper elementary grades. Among research on ability grouping in general, this research is of exceptional quality. Five well-controlled studies used random assignment of classes to treatments and sample sizes large enough to minimize the potential impact of teacher effects.

There is no evidence to suggest that achievement gains due to within-class ability grouping in mathematics are achieved at the expense of low achievers; if anything, the evidence indicates the greatest gains for this subgroup. This finding is surprising in light of several studies of ability grouping in reading suggesting that students in low reading groups experience a lower quality of instruction than do those in higher groups (see, for example, Rist, 1970; Allington, 1980; Eder, 1981). Time on task is generally lower in low than in high reading groups (Gambrell, Wilson, & Gantt, 1981;

Good & Beckerman, 1978; Martin & Evertson, 1980), and there is some evidence that low reading groups receive lower-level questions (Seltzer, 1976) and more teacher interruptions (Eder, 1981) than high groups. A few studies (e.g., Weinstein, 1976) find that reading group membership predicts student performance even if student ability is statistically controlled.

Yet comparisons of high and low reading groups are largely comparisons of more and less able or proficient students, not comparisons of different classroom organization methods. It is hardly surprising that high and low achievers differ and that their teachers' behaviors differ accordingly. Comparisons of achievement gains in high and low reading groups are bound to show an advantage of being in the high group because high achievers learn more rapidly than low achievers, and unless measures used to control for initial ability are perfectly reliable and perfectly predictive of later reading achievement, assignment to the high reading group will appear to lead to higher achievement (see Reichardt, 1979).

However, comparison of relative achievement gains or other differences between high and low reading groups only indirectly address the critical question: What are the most effective instructional arrangements for low achievers (as well as average and high achievers)? In elementary mathematics, the evidence presented here supports the use of within-class ability grouping for all students, especially low achievers. It cannot be assumed that results in mathematics can be applied to reading, but it is certainly the case that only experimental comparisons of grouped and ungrouped reading classes of the kind done to study math grouping can determine the achievement effects of within-class ability grouping in reading.

The previous discussion is based on the assumption that any effect of within-class ability grouping is in large part due to the reductions in heterogeneity it brings about. However, there is one study which raises some question about this assumption. This study, a dissertation by Eddleman (1971), compared within-class ability grouping in mathematics to a within-class grouping plan in which students were assigned to three heterogeneous subgroups. There was some differentiation of instructional level for students in the ability grouped classes, but in all other respects the teacher's methods in the two groups were identical, with instruction given to one group at a time while the other two groups worked problems at their desks. Classes were randomly assigned to treatments, with the same teachers teaching classes using homogeneous and heterogeneous subgroups.

Results of the nine-week study slightly favored the heterogeneous grouping plan (ES = -.16). Unfortunately, there was no ungrouped control condition, so it is impossible to determine whether the two forms of subgrouping were equally effective or equally ineffective; the brevity of the study suggests the latter. However, if future research were to establish that within-class ability grouping and within-class heterogeneous grouping were equally effective (and more effective than ungrouped arrangements), we would have to reconceptualize the usual explanations for the effectiveness of within-class ability grouping. For example, it may be that within-class ability grouping increases achievement by reducing the size of instruction groups (say, from thirty to ten) or by structuring the teacher's instructional time more effectively rather than having anything to do with reducing homogeneity (see Slavin and Karweit, 1984). Clearly, research directed at explaining the achievement effects of within-class ability grouping is needed.

Discussion

Many previous reviewers of the ability grouping literature have characterized the evidence as a muddle or a maze (e.g., Borg, 1965; Passow, 1962). However, earlier reviewers have generally combined elementary with secondary research, good quality with hopelessly biased studies, research on comprehensive ability grouping plans with that on special programs for the gifted or learning disabled, and in some cases, research on between-class ability grouping with that on within-class grouping.

When the scope of the review is limited to methodologically adequate studies of comprehensive ability grouping at the elementary level and different types of ability grouping are reviewed separately, the results are surprisingly clear cut for most types of grouping. The best evidence from randomized and matched equivalent studies unequivocally supports the positive achievement effects of the use of within-class ability grouping in mathematics and of Joplin and nongraded plans in reading. In contrast, there is no support for the practice of assigning students to self-contained classes according to general ability or performance level, and there are enough good quality studies of this practice that if there were any effect, it would surely have been detected. Evidence on the effects of regrouping within grade levels for reading and mathematics is unclear, and there is no methodologically adequate evidence concerning the use of reading groups.

The conclusion of the research reviewed here for practice may be quite simple: Use the grouping methods which have been found to be effective (within-class ability grouping in mathematics, Joplin and nongraded plans in reading), and avoid those which have not been found to be effective. In particular, there is good reason to avoid ability grouped class assignment, which seems to have the greatest potential for negative social effects since it entirely separates students into different streams (see Rosenbaum, 1980). However, there is much more we must understand about how various ability grouping plans have their effects. A theory able to encompass the research findings is needed. The remainder of this paper explores the findings and other evidence in an attempt to extract general principles of grouping for instruction in the elementary school.

Accommodating Instruction to Student Differences

The primary reason educators group students according to ability or performance level is to enable teachers to provide instruction closely suited to the readiness and needs of different students. In a highly diverse class, it is argued, one level and pace of instruction is likely to be too easy for some students and/or too difficult for others. Ability grouping is supposed to reduce student heterogeneity so that an appropriate pace and level of instruction is provided for most students.

Having instruction be carefully accommodated to students' level of readiness is probably more important in some subjects than in others. In general, subjects in which skills build upon one another in a hierarchical fashion (e.g., mathematics, reading) should require more accommodation to individual differences in learning rate than subjects in which learning the next skill or concept is less clearly dependent on mastery of earlier material (e.g., social studies, science). The reason for this is that with hierarchically organized subjects, there is a risk that if the teacher proceeds too rapidly, some students will lack the prerequisite skills needed to learn new material, while if the teacher takes the time needed to ensure that all students have prerequisite skills, the more able students will waste a great deal of time.

Ability grouping is one logical way out of the dilemma posed by having to choose one instructional pace for a diverse group in a hierarchical subject. Yet if an ability grouping plan is to have the desired effect, there are at least three criteria it must satisfy:

1. The grouping plan must measurably reduce student heterogeneity in the specific skill being taught;
2. The plan must be flexible enough to allow teachers to respond to missassignments and changes in student performance level after initial placement; and
3. Teachers must actually vary their pace and level of instruction to correspond to students' levels of readiness and learning rates.

As noted earlier, research on the effect of grouping on class heterogeneity has found that in the situation typical of elementary schools where students are divided into two or three "homogeneous" groups, the actual reduction in heterogeneity brought about may be quite minimal. This is particularly true when students are assigned to classes on the basis of IQ or of a general measure of performance, as imperfect correlations between these measures and actual performance in any particular subject leave a great deal of heterogeneity in the supposedly homogeneous classes (Goodlad & Anderson, 1963; Balow, 1962; Clarke, 1958; Balow & Curtin, 1966).

Thus, ability grouped class assignment generally fails to meet the first of the three criteria listed above; a one-time assignment by general ability is unlikely to create enough homo ity on any particular skill to make an instructional differenc

Ability grouped class assignment is also unlikely to fulfill the second criterion, flexibility. Transferring students between self-contained classes is difficult to arrange, so students who are misassigned or whose achievement level markedly changes over time are likely to remain in the self-contained class. In contrast, regrouping and Joplin and nongraded plans group students based on their performance in a *specific* skill, and are inherently more flexible than ability grouped class assignment, as changing students between regrouped classes only involves one subject, not a change in students' main class identification. Similarly, within-class ability grouping is done based on performance in a particular skill, and is the easiest grouping plan to alter based on changes in student performance.

To what extent do teachers adapt their level and pace of instruction to the needs of different ability groups? Research comparing alternative grouping arrangements has not examined this question in any depth, but there are some clues. Studies by Barr and Dreeben (1983) found that teachers do adapt their instructional pace to accommodate the aptitudes of reading groups, but they also found considerable variation from school to school and teacher to teacher in pacing for groups of similar aptitudes.

Some indirect evidence suggests the importance of adapting instruction to student differences. One form of grouping often seen in mathematics instruction involves assigning students to three ability groups within the class. The teacher presents one lesson to the class as a whole, and while students are doing seatwork, visits with the low ability group to provide additional explanation, the high group to provide enrichment, and the middle group to provide some of each. Note that this strategy does not adapt the pace or level of instruction to student needs. Two studies at the secondary level (Bierden, 1969; Mortlock, 1970) and one in a community college (Merritt, 1973) found no significant differences between this type of within-class ability grouping and traditional whole-class instruction, much in contrast to the studies of within-class ability grouping plans in which level and pace of instruction were adapted to student performance levels.

One critical feature of the successful Joplin/nongraded plans is frequent, careful assessment of student performance levels and provision of materials appropriate to these levels regardless of students' grade levels. In these plans, adaptation of instructional pace and level to student needs is as great as it could possibly be short of individualization; in one study of the Joplin Plan (Moorhouse, 1964) it was noted that "three-quarters of the (grade 4-6 experimental) students were reading material either above or below the grade level they would usually be asked to attempt in the graded system" (pp. 281-282).

In contrast, a study of regrouping for reading by Moses (1966) instructed experimental teachers to use only materials appropriate to students' grade levels and to follow the school district's usual course of study. This study found no significant advantages of ability grouping (ES = +.05). Similarly, Davis and Tracy (1963), in a study of regrouping for mathematics, held experimental and control teachers to the same grade-level textbooks. In this study, control students gained more in achievement than did students in the within-grade regrouping plan.

An otherwise similar study of within-grade regrouping for mathematics by Provus (1960) did allow for differentiation of level and pace of instruction, noting that "...it was possible for a fourth grader... to advance to sixth grade or even eighth grade work by the end of the school year." Control (ungrouped) students were also able to go beyond their designated grade level, but presumably did not do so as often as did experimental students. This study found strong positive effects of regrouping on mathematics achievement (ES = +.39).

Of course, Joplin and nongraded plans can be seen as forms of regrouping for reading and/or mathematics which go to great lengths to adapt the level and pace of instruction to that of the regrouped classes. In fact, it could be argued that it is not the cross-grade aspect of Joplin/nongraded plans that accounts for their effects, but rather the fact that students in these plans are carefully assessed and given instruction appropriate to their needs in the regrouped classes.

Taken together, the evidence points to a conclusion that for ability grouping to be effective at the elementary level, it must create true homogeneity on the specific skill being taught and instruction must be closely tailored to students' levels of performance.

Class Identification

Critics of ability grouping (e.g., Oakes, 1985; Schafer & Olexa, 1971; Rosenbaum, 1980) have often noted the detrimental psychological effect of being placed in a low achieving class or track. An interview with a former delinquent about his discovery that he had been assigned to the "basic track" in junior high school illustrates this theme (from Schafer & Olexa, 1971, pp. 62-63):

> I felt good when I was with my (elementary) class, but when they went and separated us -- that changed us. That changed our ideas, our thinking, the way we thought about each other, and turned us to enemies toward each other -- because they said I was dumb and they were smart.
>
> When you first go to junior high school you do feel something inside -- it's like a ego. You have been from elementary to junior high, you feel great inside... you get this shirt that says Brown Junior High.. and you are proud of that shirt. But then you go up there and the teacher says -- "Well, so and so, you're in the basic section, you can't go with the other kids." The devil with the whole thing -- you lose -- something in you -- like it goes out of you.

The anguish expressed by the student who was assigned to the basic track is interesting in light of the high probability that the student had been in low reading groups in elementary school, but he still perceived being "separated" into different classes as a completely new and much more serious affront to his self-esteem. Within-class grouping generally takes place within the context of a more or less heterogeneous class, and a student still identifies with the class as a whole. Ability grouped class assignment, however, creates a situation in which a low achiever's only reference group is other low achievers (see Richer, 1976).

It may be that flexible skill-based ability grouping has a different impact on students' self perceptions and on teachers' expectations than does ability grouped class assignment. Students in low performing classes would seem likely to come to see themselves as a different species of human being, where those assigned to low reading or math groups in heterogeneous classes may see this placement as being done to help them, particularly if assignment to groups is flexible and is clearly focused on achievement in a particular subject.

Teachers' expectations and behaviors may also be different in different types of ability grouping. Not surprisingly, teachers prefer to teach higher-achieving students (NEA, 1968) and have higher expectations for their achievement. These expectations can have an impact on teachers' behaviors and students' achievement (see Good and Brophy, 1984). For example, in a study of Air Force training, Schrank (1969) had students randomly assigned to classes, but told instructors that the classes were grouped by ability. Classes which had been (falsely) identified as high achieving in fact achieved more than did classes identified as low achieving.

The problems of teachers' low expectations for students in low-track classes and their dislike for being assigned to these low achieving classes are largely alleviated in Joplin and nongraded plans, in which reading classes are formed across age lines. This means that a class may contain high-achieving young students and low-achieving older students, so homogeneity for instruction may be achieved without establishing classes that teachers do not want to teach and from whom teachers expect little. Also, as noted by one of the authors of the Hillson et al. (1964) study of a nongraded plan (J.W. Moore personal communication, January 23, 1986), low achieving students in nongraded plans progress from reading level to reading level rather than remaining year after year in the low reading group.

Teachers may have low expectations for students in low reading or math groups, but there is some evidence that they try to bring low groups up to the level of the rest of the class. For example, Rowan & Miracle (1983) found that in between-class ability grouping teachers tended to maintain a slow pace of instruction for low achieving classes, but tended to allocate more time and a more rapid pace of instruction for low reading groups in heterogeneous classes. This and other research (e.g., Alpert, 1974) suggests that in within-class ability grouping, teachers tend to try to equalize the achievement of all students by assigning smaller numbers of students to low groups.

Another issue relating to students' identification with their class is the question of how many times students are regrouped each day. When students are regrouped for reading and/or mathematics, they still typically spend the rest of their school day in heterogeneous homeroom classes, which probably remain as their primary reference group. However, as the number of regroupings increases, the situation comes to resemble departmentalization, where students move from teacher to teacher and have no one group with which to identify. Unfortunately, there is no research on achievement outcomes of departmentalization at the elementary level, although one study of seventh and eighth graders by Spivak (1956) found that students in self-contained classes learned more than matched students in departmentalized settings.

Departmentalization might reduce students' attachment to school by diffusing their attachments to particular teachers. Indirect evidence of this is a finding by Slavin and Karweit (1982) that student truancy in an urban school district rose from about 8% in the fifth and sixth grades to 26% in the seventh grade, the time of first exposure to a departmentalized (and tracked) school in which no one teacher takes responsibility for any one student. It may be that the Koontz (1961) study, in which students were separately regrouped for reading, mathematics, and spelling, deprived students of an opportunity to identify with a single teacher and a heterogeneous class. In this study, heterogeneous control students gained more in achievement than did those who were regrouped, with low achievers suffering most from regrouping.

The evidence on the importance of having students principally identify with a heterogeneous class is more speculative than conclusive, but several indirect indications support the following conclusion: Students should be assigned to heterogeneous classes for as much of the school day as possible, so that it is the heterogeneous homeroom with which students principally identify. Even if for no reason than that separating students by ability goes against the grain of our democratic, egalitarian ideals, any ability grouping plan must have clear educational benefits it it is to be justified. Because no achievement benefits of ability grouped class assignment have been identified, and because more effective grouping methods exist, use of this strategy should be avoided.

Instructional Time

One issue of considerable importance in relation to within-class ability grouping relates to a tradeoff between providing students with instruction appropriate to their needs on one hand, and providing adequate instructional time on the other (see Slavin, 1984a). When a teacher uses a within-class ability grouping plan with three groups, this means that students must spend at least two-thirds of their instructional time working without direct teacher instruction or supervision. Several studies have found that large amounts of unsupervised seatwork are detrimental to student achievement (see Brophy and Good, 1986). Transition times between ability groups further reduce instructional time (Arlin, 1979).

The amount of instructional time lost due to use of within-class ability grouping depends directly on the number of groups in use. Division of students into large numbers of ability groups forces the teacher to spend less time with each group and to assign large amounts of independent seatwork, much of which may be of little value beyond keeping students quiet and busy (see Anderson, Brubaker, Alleman-Brooks, and Duffy, 1985). It may not be a coincidence that Wallen and Vowles (1960), the study with the smallest effect size favoring within-class ability grouping in mathematics (ES = +.07), is also the only one to use four (rather than two or three) ability groups.

The evidence summarized in Table 5 clearly indicates that regardless of any losses in instructional time associated with within-class grouping, this strategy is instructionally effective in elementary mathematics. Mathematics instruction does require a certain amount of time for students to work problems on their own, so followup time (the time during which some students must work by themselves while others are working with the teacher) may be less of a problem in mathematics than in other subjects. However, it still seems apparent that the requirement for large amounts of followup time is a drawback in any within-class grouping arrangement.

The problem of followup time may be important in explaining the effectiveness of Joplin plans for reading. The studies of this program do not typically compare the numbers of ability groups used in experimental and control groups, but it is clear that there are smaller numbers of reading groups in Joplin than in traditional classes, and that in some cases Joplin Plan classes do not use reading groups at all. In fact, some authors (e.g., Newport, 1967) clearly describe studies of the Joplin Plan as comparisons of intervs. intra-class ability grouping, since control groups in Joplin studies always used reading groups. The cross-grade grouping plans create such homogeneous classes that the need for within-class grouping is diminished or eliminated. The time savings of such plans are therefore considerable, and they may produce their effects in part because students in them receive a greater amount of direct instruction from the teacher and supervision during seatwork than do students in control classes using the more typical three or more reading groups.

This line of reasoning may justify use of smaller numbers of ability groups in heterogeneous reading classes. Unfortunately there is no direct experimental evidence on the optimum number of reading or math groups; the number three is treated as though it were handed down from Mount Sinai. In heterogeneous classes it may be that small numbers of ability groups do not provide adequate homogeneity for effective instruction. However, it should be noted that two studies involving only two ability groups in heterogeneous mathematics classes found significant benefits of ability grouping for student achievement (Slavin and Karweit, 1985).

Conclusions

This paper reviewed the best evidence concerning the achievement effects of comprehensive ability grouping plans in elementary schools. Five principal grouping plans were examined: ability grouped class assignment, regrouping for reading and/or mathematics, Joplin and nongraded plans, comprehensive nongraded plans, and within-class ability grouping. The effects of these grouping methods on student achievement from methodologically adequate studies are summarized below:

Ability Grouped Class Assignment. Evidence from fifteen comparisons in twelve matched equivalent and one randomized study clearly indicates that assigning students to self-contained classes according to general achievement or ability does not enhance student achievement in the elementary school (median ES = .00).

Regrouping for Reading and Mathematics. Research is unclear on the achievement outcomes of grouping plans in which students remain in heterogener asses most of the day and are regrouped by ability within grade levels for reading and/or mathematics. There is some evidence that such plans can be instructionally effective if the level and pace of instruction is adapted to the achievement level of the regrouped class and if students are not regrouped for more than one or two different subjects.

Joplin Plan. There is good evidence that regrouping students for reading across grade lines increases reading achievement. The Joplin Plan (Floyd, 1954) and essentially similar forms of nongraded plans (e.g., Hillson et al., 1964) have had relatively consistent positive effects on reading achievement (median ES = +.44), and one study (Hart, 1962) found that a similar program could also be effective in mathematics.

Comprehensive Nongraded Plans. Evidence from studies of nongraded plans closer to those suggested by Goodlad and Anderson (1963) has been less consistent than for Joplin-like nongraded plans, but the preponderance of the evidence is still positive (median ES = +.29). In particular, the best evidence from well-controlled studies in regular schools supports the use of comprehensive nongraded plans.

Within-Class Ability Grouping. Research on within-class ability grouping is unfortunately limited to mathematics in upper elementary school. However, this research clearly supports the use of within-class grouping (approximate median ES = +.34), especially if the number of groups is kept small. Achievement effects of within-class ability grouping are slightly larger for low than for high or average achievers.

In addition to conclusions about the effects of particular grouping strategies, several general principles of ability grouping were proposed on the basis of the experimental evidence. The following are advanced as elements of effective ability grouping plans:

1. Students should remain in heterogeneous classes at most times, and be regrouped by ability only in subjects (e.g., reading, mathematics) in which reducing heterogeneity is particularly important. Students' primary identification should be with a heterogeneous class.

2. Grouping plans must reduce student heterogeneity in the specific skill being taught (e.g., reading, mathematics).

3. Grouping plans must frequently reassess student placements and must be flexible enough to allow for easy reassignments after initial placement.

4. Teachers must actually vary their level and pace of instruction to correspond to students' levels of readiness and learning rates in regrouped classes.

5. In within-class ability grouping, numbers of groups should be kept small to allow for adequate direct instruction from the teacher for each group.

Future Directions

One great danger in reviewing any voluminous literature is that the review will discourage further work in the area, as researchers question the value of one more study. We hope the present review will have the effect of stimulating rather than inhibiting additional research on ability grouping in the elementary grades.

There are many fundamental questions yet to be explored. For example, the research on within-class ability grouping is mostly restricted to mathematics. Experimental studies of the use of reading groups are needed, as are studies of optimum numbers of within-class groups for reading and math.

Many studies are needed to understand why and under what conditions various grouping plans produce achievement effects. Simple-appearing changes in grouping are likely to have complex effects, any of which may contribute to ultimate effects on student achievement. For example, different between-class grouping plans (e.g., ability grouped class assignment, Joplin Plan) are likely to have different effects on within-class grouping.

Studies are needed to understand the effects of various grouping plans on what actually happens in the classroom, for example, how different plans affect the teacher's pace of instruction and use of class time and the success rate of students close to and far away from the class' mean aptitude. Component analyses are needed to explore the critical features of various grouping plans. For example, many fundamentally different practices go under the title "non-grading." Which of these account for the positive effects seen in the studies of this practice? Is it simply use of flexible, homogeneous grouping across grade lines (as in the Joplin Plan), or are other factors involved?

There are two particularly important reasons for further investigation of grouping practices in elementary schools. First, every school district, school administrator, and teacher makes decisions about ability grouping at some time, and these decisions should be made in light of reliable evidence. Ironically, the grouping practice with the least support in the research, ability grouped class assignment, is among the most widely used; schools need effective alternatives to this practice.

Second, if educational researchers can identify grouping practices which can accelerate student achievement, this would provide one kind of school reform that would be low in cost, easy to implement, and easy to maintain over time. In a time of increasing demands on education coupled with dwindling resources, research on easily modified school organizational practices seems particularly likely to bear fruit. We have much yet to learn in this area, but this review illustrates that the potential of effective grouping practices for meaningful improvements in the achievement of elementary students is great, and is certainly worthy of further study.

References

Allington, R. L. (1980). Poor readers don't get to read much in reading groups. Language Arts, 57, 872-876.

Alpert, J. L. (1974) Teacher behavior across ability groups: A consideration of the mediation of pygmalion effects. Journal of Educa ional Psychology, 66, 348-353.

Anastasiouw, N. J. (1968). A comparison of two approaches in upgrading reading instruction. Elementary English, 45, 495-499.

Anderson, L. M., Brubaker, N. L., Alleman-Brooks, J., & Duffy, G. G. (1985). A qualitative study of seatwork in first grade classrooms. Elementary School Journal, 86, 123-140.

Anderson, R. H., & Goodlad, J. I. (1962). Self-appraisal in nongraded schools: A survey of findings and perceptions. Elementary School Journal, 62, 261-269.

Arlin, M. (1979). Teacher transitions can disrupt time flow in classrooms. American Educational Research Journal, 16, 42-56.

Atkinson, J. W., & O'Connor, P. (1963). Effects of ability grouping in schools related to individual differences in achievement-related motivation. Final Report, Cooperative Research Project No. OE-2-10-024. U. S. Department of H. E. W.

Balow, B., & Curtin, J. (1966). Ability grouping of bright pupils. Elementary School Journal, 66, 321-326.

Balow, I. H. (1962). Does homogeneous grouping give homogeneous groups? Elementary School Journal, 63, 28-32.

Balow, I. H., & Rudell, A. K. (1963). The effects of three types of grouping on achievement. California Journal of Educational Research, 14, 108-117.

Bangert-Drowns, R. L. (in press). A review of developments in meta-analytic method. Psychological Bulletin.

Barker-Lunn, J. C. (1970). Streaming in the primary school. London: National Foundation for Education Research in England and Wales.

Barr, R. (1975). How children ... taught to read: Grouping and pacing. School Review, 83, 479-498.

Barr, R., & Dreeben, R. (1983). How schools work. Chicago: University of Chicago Press.

Barthelmess, H. M., & Boyer, P. A. (1932). An evaluation of ability grouping. Journal of Educational Research, 26, 284-294.

Begle, E. G. (1975). Ability grouping for mathematics instruction: A review of the empirical literature. Stanford Mathematics Education Study Group, Stanford University. (ERIC No. ED 116 938).

Berkun, M. M., Swanson, L. W., & Sawyer, D. M. (1966). An experiment on homogeneous grouping for reading in elementary classes. Journal of Educational Research, 59, 413-414.

Bierden, J. E. (1969). Provisions for individual differences in seventh grade mathematics based on grouping and behaviorial objectives: An exploratory study. Dissertation Abstracts International, 30, 196A-197A. (University Microfilms No. 69-12042).

Billett, R. O. (1932). The administration and supervision of homogeneous grouping. Columbus: Ohio State University Press.

Block, J. H., & Anderson, L. W. (1975). Mastery learning in classroom instruction. New York: Macmillan.

Bloom, B. S. (1976). Human characteristics and school learning. New York: McGraw-Hill.

Borg, W. R. (1965). Ability grouping in the public schools: A field study. Journal of Experimental Education, 34, 1-97.

Bowman, B. L. (1971) A comparison of pupil achievement and attitude in a graded school with pupil achievement and attitude in a nongraded school 1968-69, 1969-70 school years. Dissertation Abstracts, 32, 86-A. (University Microfilms No. 71-20, 958).

Breidenstine, A. G. (1936). The educational achievement of pupils in differentiated and undifferentiated groups. Journal of Experimental Education, 5, 91-135.

Bremer, N. (1958). First grade achievement under different plans of grouping. Elementary English, 35, 324-326.

Brody, E. B. (1970). Achievement of first- and second-year pupils in graded and nongraded classrooms. Elementary School Journal, 70, (April, 1970), 391-394.

Brophy, J. E., & Good, T. L. (1986). Teacher behavior and student achievement. In M. C. Wittrock (Ed.), Handbook of Research on Teaching. Third Edition. New York: MacMillan.

Buffie, E. G. W. (1962). A comparison of mental health and academic achievement: The nongraded school vs. the graded school. Dissertation Abstracts, 23, 4255-A. (University Microfilms No. 63-2587).

Campbell, A. L. (1965). A comparison of the effectiveness of two methods of class organization for the teaching of arithmetic in junior high school. Dissertation Abstracts International, 26, 813-814. (University Microfilms No. 65-6726).

Carbone, R. F. (1961). A comparison of graded and non-graded elementary schools. Elementary School Journal, 62, 82-88.

Carson, R. M., & Thompson, J. (1964). The Joplin Plan and traditional reading groups. Elementary School Journal, 65, 387-392.

Cartwright, G. P., & McIntosh, D. K. (1972). Three approaches to grouping procedures for the education of disadvantaged primary school children. Journal of Educational Research, 65, 425-429.

Clarke, S. C. T. (1958). The effect of grouping on variability in achievement at the grade 3 level. Alberta Journal of Educational Research, 4, 162-171.

Coleman, J. S., et al. (1966). Equality of educational opportunity. Washington, DC: U. S. Department of Health, Education, and Welfare.

CTB/McGraw-Hill. (1979). California Achievement Tests, Forms C and D, Levels 10-19: Technical Bulletin 1. Monterey, CA: CTB/McGraw-Hill.

Cushenberry, D. (1964). The intergrade plan of grouping for reading instruction as used in the public schools of Joplin, Missouri. Dissertation Abstracts, (University Microfilms No. 64-9352).

Daniels, J. C. (1961). The effects of streaming in the primary school. Comparison of streamed and unstreamed schools. British Journal of Educational Psychology, 31, 119-126.

Davis, O. L., & Tracy, N. H. (1963). Arithmetic achievement and instructional grouping. Arithmetic Teacher, 10, 12-17.

Dewar, J. (1964). Grouping for arithmetic instruction in sixth grade. Elementary School Journal, 63, 266-269.

Douglas, J. F. (1973). A study of streaming at a grammar school. Educational Research, 15, 140-143.

Eash, M. J. (1961). Grouping: What have we learned? Educational Leadership, 18, 429-434.

Eddleman, V. K. (1971). A comparison of the effectiveness of two methods of class organization for arithmetic instruction in grade five. Dissertation Abstracts International, 32, 1744A (University Microfilms No. 71-25035).

Eder, D. (1981). Ability grouping as a self-fulfilling prophecy: A microanalysis of teacher-student interaction. Sociology of Education, 54, 151-162.

Edminston, R. W., & Benfer, J. G. (1949). The relationship between group achievement and range of abilities within groups. Journal of Educational Research, 42, 547-548.

Esposito, D., (1973). Homogeneous and heterogeneous ability grouping: Principal findings and implications for evaluating and designing more effective educational environments. Review of Educational Research, 43, 163-179.

Evertson, C. (1982). Differences in instructional activities in average- and low-achieving junior high English and math classes. Elementary School Journal, 82, 329-350.

Felmlee, D., & Eder, D. (1983). Contextual effects in the classroom: The impact of ability groups on student attention. Sociology of Education, 56, 77-87.

Findley, W. G., & Bryan, M. (1971). Ability grouping: 1970 Status, impact, and alternatives. Athens, GA: Center for Educational Improvement, University of Georgia. (ERIC Document Reproduction Service No. ED 060-595).

Flair, M. D., (1964). The effect of grouping on achievement and attitudes toward learning of first grade pupils. Dissertation Abstracts, 25, 6430. (University Microfilms No. 65-03,259).

Floyd, C. (1954). Meeting children's reading needs in the middle grades: A preliminary report. Elementary School Journal, 55, 99-103.

Fox, L. H. (1979). Programs for the gifted and talented: An overview. In A. H. Passow, The Gifted and Talented: Their Education and Development. Chicago: University of Chicago Press.

Gambrell, L. B., Wilson, R. M., & Gantt, W. N. (1981). Classroom observations of task-attending behaviors of good and poor readers. Journal of Educational Research, 74, 400-404.

Giaconia, R. H., & Hedges, L. V. (1982). Identifying features of effective open education. Review of Educational Research, 52, 579-602.

Glass, G., McGaw, B., & Smith, M. L. (1981). Meta-analysis in social research. Beverly Hills, CA: Sage.

Goldberg, M. L., Passow, A. H., & Justman, J. (1966). The effects of ability grouping. New York: Teachers College Press.

Good, T. L., & Beckerman, T. M. (1978). Time on task: A naturalistic study in sixth grade classrooms. Elementary School Journal, 78, 192-201.

Good, T. L., & Brophy, J. E. (1984). Looking in classrooms. (Third Edition). New York: Harper & Row.

Good, T., & Grouws, D. (1977). Teaching effects: A process-product study in fourth grade mathematics classes. Journal of Teacher Education, 28, 49-54.

Good, T., & Grouws, D. (1979). The Missouri Mathematics Effectiveness Project: An experimental study in fourth grade classrooms. Journal of Educational Psychology, 71, 355-362.

Good, T., Grouws, D., & Ebmeir, H. (1983). Active mathematics teaching. New York: Longman.

Good, T., & Marshall, S. (1984). Do students learn more in heterogeneous or homogeneous groups? In P. Peterson & L. Cherry Wilkinson, Student Diversity and the Organization, Process, and Use of Instructional Groups in the Classroom. New York: Academic Press.

Goodlad, J. I. (1983). A place called school. New York: McGraw-Hill.

Goodlad, J. I., & Anderson, R. H. (1959). The nongraded elementary school. New York: Harcourt, Brace, & World.

Goodlad, J. I., & Anderson, R. H. (1963). (Rev. Ed.). The nongraded elementary school. New York: Harcourt, Brace, & World.

Green, D. R., & Riley, H. W. (1963). Interclass grouping for reading instruction in the middle grades. Journal of Experimental Education, 31, 273-278.

Haller, E., & Davis, S. (1980). Does socioeconomic status bias the assignment of elementary school students to reading groups? American Educational Research Journr' 17, 409-418.

Hallinan, M., & Sorensen, A. (1983) The formation and stability of instructional groups. American Sociological Review, 48, 839-851.

Halliwell, J. W. (1963) A comparison of pupil achievement in gr'ded and nongraded primary classrooms. Journal of Experimental Education, 32, 59-64.

Hart, R. H. (1959). The effectiveness of an approach to the problem of varying abilities in teaching reading. Journal of Educational Research, 52, 228-231.

Hart, R. H. (1962). The nongraded primary school and arithmetic. The Arithmetic Teacher, 9, 130-133.

Hartill, R. (1936). Homogeneous grouping as a policy in the elementary schools in New York City. Teachers College, Columbia Contributions to Education, 690.

Heathers, G. (1969). Grouping. In R. Ebel (Ed.), Encyclopedia of educational research (Fourth Edition). New York: Macmillan.

Hickey, M. P. (1963). An evaluation of pupil progress of the undgraded primary program in the diocese of Pittsburgh. Dissertation Abstracts, 23, 2817-A (University Microfilms No. 63-208).

Hiebert, E. H. (1983). An examination of ability grouping for reading instruction. Reading Research Quarterly, 18, 231-255.

Hillson, M., Jones, J. C., Moore, J. W., & Van Devender, F. (1964). A controlled experiment evaluating the effects of a nongraded organization on pupil achievement. Journal of Educational Research, 57, 548-550.

Hobson v. Hansen, 269 F. Supp. 401 (1967).

Hopkins, K. D., Oldridge, O. A., & Williamson, M. L. (1965). An empirical comparison of pupil achievement and other variables in graded and nongraded classes. American Educational Research Journal, 2, 207-215.

Hull, J. H. (1958). Multigrade teaching. The Nation's Schools, 62, 33-36.

Ingram, V. (1960). Flint evaluates its primary cycle. Elementary School Journal, 61, 76-80.

Johnston, H. J. (1973). The effect of grouping patterns on first-grade children's academic achievement and personal and social development. Dissertation Abstracts International, (University Microfilms No. 73-25, 893).

Jones, C. J., Moore, J. W., & Van Devender, F. (1967). Comparison of pupil achievement after one and one-half and three years in a nongraded program. Journal of Educational Research, 61, 75-77.

Jones, D. M. (1948). An experiment in adaption to individual differences. Journal of Educational Psychology, 39, 257-273.

Justman, J. (1968). Reading and class homogeneity. Reading Teacher, 21, 314-316.

Kierstead, R. (1963). A comparison and evaluation of two methods of organization for the teaching of reading. Journal of Educational Research, 56, 317-321.

Killough, C. K. (1972). An analysis of the longitudinal effects that a nongraded elementary program, conducted in an open-space school, had on the cognitive achievement of pupils. Dissertation Abstracts, 32, 3614. (University Microfilms No. 72-2265).

Klausmeier, H. J., Rossmiller, R. A., and Saily, M. (1977). Individually guided elementary education: concepts and practices. New York: Academic Press.

Koontz, W. F. (1961). A study of achievement as a function of homogeneous grouping. Journal of Experimental Education, 30, 249-253.

Kulik, C.-L., & Kulik, J. (1982). Effects of ability grouping on secondary school students: A meta-analysis of evaluation findings. American Educational Research Journal, 19, 415-428.

Kulik, C.-L., & Kulik, J. A. (1984, August). Effects of ability grouping on elementary school pupils: A meta-analysis. Paper presented at the Annual Meeting of the American Psychological Association, Toronto.

Kulik, J. A., & Kulik, C.-L. (1984). Effects of accelerated instruction on students. Review of Educational Research, 54, 409-425.

Leinhardt, G., & Pallay, A. (1982). Restrictive educational settings: Exile or haven? Review of Educational Research, 52, 557-578.

Loomer, B. M. (1962). Ability grouping and its effects upon individual achievement. Dissertation Abstracts, 23, 1581 (University Microfilms No. 62-4982).

Leiter, J. (1983). Classroom composition and achievement gains. Sociology of Education, 56, 126-132.

Machiele, R. B. (1965). A preliminary evaluation of the non-graded primary at Leal School, Urbana. Illinois School Research, 1, 20-24.

Madden, N. A., & Slavin, R. E. (1983). Mainstreaming students with mild academic handicaps: Academic and social outcomes. Review of Educational Research, 53, 519-569.

Martin, J., & Evertson, C. M. (1980). Teachers' interactions with reading groups of differing ability levels. (Tech. Rep. No. R-4093). Austin: University of Texas, Research and Development Center for Teacher Education.

Martin, W. H. (1927). The results of homogeneous grouping in the junior high school. Unpublished doctoral dissertation, Yale University.

McLoughlin, W. P. (1970). Continuous pupil progress in the non-graded school: Hope or hoax? Elementary School Journal, 71, 90-96.

McPartland, J. (1968). The segregated student in desegregated schools: Sources of influence on Negro secondary students. Baltimore, MD: Johns Hopkins University.

Merritt, H. P. W. (1973). The effects of variations in instruction and final unit evaluation procedures on community college beginning algebra classes. _Dissertation Abstracts International, 33_, 5978A (University Microfilms No. 73-11207).

Miller, R. I. (1967). _The nongraded school: Analysis and study._ New York: Harper & Row.

Miller, W. S., & Otto, H. J. (1930). Analysis of experimental studies in homogeneous grouping. _Journal of Educational Research, 21_, 95-102.

Moorhouse, W. F. (1964) Interclass grouping for reading instruction. _Elementary School Journal, 64_, 280-286.

Morgan, E. F., & Stucker, G. R. (1960). The Joplin Plan of reading vs. a traditional method. _Journal of Educational Psychology, 51_, 69-73.

Morgenstern, A. (1963). A comparison of the effects of heterogeneous (ability) grouping on the academic achievement and personal-social adjustment of sixth grade children. _Dissertation Abstracts, 23_, 1054. (University Microfilms No. 63-6560).

Morris, V. P. (1969). An evaluation of pupil achievement in a nongraded primary plan after three, and also five years of instruction. _Dissertation Abstracts, 29_, 3809-A. (University Microfilms No. 69-7352).

Mortlock, R. (1970). Provision for individual differences in eleventh grade mathematics using flexible grouping based on achievement of behavioral objectives: An exploratory study. _Dissertation Abstracts International, 30_, 3643A (University Microfilms No. 70-4148).

Moses, P. J. (1966). A study of inter-class ability grouping on achievement in reading. Dissertation Abstracts, 26, 4342.

National Education Association (1967). Nongraded schools: NEA research memo. In R. I. Miller (Ed.), The nongraded school, (pp. 157-191). New York: Harper & Row.

National Education Association (1968). Ability Grouping Research Summary. Washington, DC: National Education Association.

Newport, J. F. (1967). The Joplin Plan: The score. Reading Teacher, 21, 158-162.

Oakes, J. (1985). Keeping track: How schools structure inequality. New Haven, CT: Yale University Press.

Otto, H. J. (1969). Nongradedness: An elementary school evaluation. Austin: University of Texas.

Passow, A. H. (1962). The maze of research on ability grouping. Educational Forum, 26, 281-288.

Passow, A. H. (Ed.) (1979). The gifted and talented: Their education and development. Chicago: University of Chicago Press.

Pavan, B. N. (1973). Good news: Research on the nongraded elementary school. Elementary School Journal, 73, 333-342.

Persell, C. (1977). Education and inequality: The roots and results of stratification in America's schools. New York: Free Press.

Piland, J. C., & Lemke, E. A. (1971). The effect of ability grouping on concept learning. Journal of Education Research, 64, 209-212.

Powell, W. R. (1964). The Joplin Plan: An evaluation. Elementary School Journal, 64, 387-392.

Provus, M. M. (1960). Ability grouping in arithmetic. Elementary School Journal, 60, 391-398.

Putbrese, L. (1972). An investigation into the effects of selected patterns of grouping upon arithmetic achievement. Dissertation Abstracts International, 32, 5113A (University Microfilms No. 72-8388).

Ramsey, W. (1962). An evaluation of a Joplin Plan of grouping for reading instruction. Journal of Educational Research, 55, 567-572.

Rankin, P. T., Anderson, C. T., & Bergman, W. G. (1936). Ability grouping in the Detroit individualization experiment. In G. M. Whipple (Ed.), The grading of pupils (35th yearbook of the National Society for the Study of Education, Part I). Bloomington, IL: Public School Publishing Co.

Reichart, C. S. (1979). The statistical analysis of data from nonequivalent group designs. In T. C. Cook & D. T. Campbell, Quasi-experimentation: Design and analysis issues for field settings. Chicago: Rand McNally.

Remacle, L. F. (1971). A comparative study of the differences in attitudes, self-concept, and achievement of children in graded and nongraded elementary schools. Dissertation Abstracts, 31, 5948-A. (University Microfilms No. 71-12,649).

Richer, S. (1976). Reference-group theory and ability grouping: A convergence of sociological theory and educational research. Sociology of Education, 49, 65-71.

Rist, R. (1970). Student social class and teacher expectations: The self-fulfilling prophecy in ghetto education. Harvard Educational Review, 40, 411-451.

Rosenbaum, J. E. (1976). Making inequality: The hidden curriculum of high school tracking. New York: Wiley.

Rosenbaum, J. E. (1980). Social implications of educational grouping. Review of Research in Education, 8, 361-401.

Ross, G. A. (1967). A comparative study of pupil progress in ungraded and graded primary programs. Dissertation Abstracts, 28, 2146-A. (University Microfilms No. 67-16,428).

Rowan, B., & Miracle, A. (1983). Systems of ability grouping and the stratification of achievement in elementary schools. Sociology of Education, 56, 133-144.

Rothrock, D. G. (1961). Heterogeneous, homogeneous, or individualized approach to reading? Elementary English, 38, 233-235.

Russell, D. H. (1946). Inter-class grouping for reading instruction in the intermediate grades. Journal of Educational Research, 39, 462-470.

Schafer, W. E., & Olexa, C. (1971). Tracking and opportunity. Scranton, PA: Chandler.

Schrank, W. R. (1969). Academic stimulation of mathematics pupils from their classroom association with brighter pupils. Mathematics Teacher, 62, 474.

Seltzer, D. A. (1976). A descriptive study of third grade reading groups. Dissertation Abstracts, 36, 5811 (University Microfilms No. 76-6345).

Skapski, M. K. (1960). Ungraded primary reading program: An objective evaluation. Elementary School Journal, 61, 41-45.

Slavin, R. E. (1984a). Meta-analysis in education: How has it been used? Educational Researcher, 13 (8), 6-15, 24-27.

Slavin, R. E. (1984b). Research methods in education: A practical guide. Englewood Cliffs, NJ: Prentice-Hall.

Slavin, R. E. (1986). Best-evidence synthensis: An alternative to meta-analytic and traditional reviews. Manuscript submitted for publication.

Slavin, R. E., & Karweit, N. L. (1982, August). School organizational vs. developmental effects on attendance among young adolescents. Paper presented at the annual convention of the American Psychological Association, Washington, DC.

Slavin, R. E., & Karweit, N. L. (1983). Ability Grouped Active Teaching (AGAT): Teacher's Manual. Baltimore, MD: Center for Social Organization of Schools, Johns Hopkins University.

Slavin, R. E., & Karweit, N. L. (1984, April). Within-class ability grouping and student achievement: Two field experiments. Paper presented at the annual convention of the American Educational Research Association, New Orleans.

Slavin, R. E., & Karweit, N. L. (1985). Effects of whole class, ability grouped, and individualized instruction on mathematics achievement. American Educational Research Journal, 22, 351-367.

Smith, W. M. (1960). The effect of intra-class ability grouping on arithmetic achievement in grades two through five. Dissertation Abstracts International, 21, 563-564.

Sorensen, A. B. (1970). Organiational differentiation of students and educational opportunity. Sociology of Education, 43, 355-376.

Spence, E. S. (1958). Intra-class grouping of pupils for instruction in arithmentic in the intermediate grades of the elementary school. Dissertation Abstracts International, 19, 1682. (University Microfilms No. 58-5635).

Spivak, M. L. (1956). Effectiveness of departmental and self-contained seventh- and eighth-grade classrooms. School Review, 64, 391-396.

Stern, A. M. (1972). Intraclass grouping of low achievers in mathematics in the third and fourth grades. Dissertation Abstracts International, 32, 5539A. (University Microfilms No. 72-11900).

Tobin, J. F. (1966). An eight year study of classes grouped within grade levels on the basis of reading ability. Unpublished Doctoral Dissertation, Boston University (University Microfilms No. 66-345).

Wallen, N. E., & Vowles, R. O. (1960). The effect of intraclass ability grouping on arithmetic achievement in the sixth grade. Journal of Educational Psychology, 51, 159-163.

Weiner, C. B. (1979). The effects on second and third grade mathematics achievement of a teacher's intra-class grouping, lesson activity and written work assignments. Dissertation Abstracts International, 39, 3853A-3854A. (University Microfilms No. 80-00816).

Weinstein, R. S. (1976). Reading group membership in first grade: Teacher behaviors and pupil experinece over time. Journal of Educational Psychology, 68, 103-116.

Wilson, B., & Schmits, D. (1978). What's new in ability grouping? Phi Delta Kappan, 60, 535-536.

Wilt, H. J. (1971). A comparison of student attitudes toward school, academic achievement, internal structures and procedures: The nongraded school vs. the graded school. Dissertation Abstracts, 31, 5105-A. (University Microfilms No. 71-8406).

Figure 1

Typology of Elementary Ability Grouping

Between Class			Within Class
Class Assignment	Selected Subjects		
	Within Grade	Cross Grade	
Ability Grouped Class Assignment	Regrouping for Reading, Math	Joplin, Non-Graded Plans	Reading, Math Groups

Article	Grades	Location	Sample Size	Duration	Grouping Criteria	Design	EFFECT SIZES by Achievement	EFFECT SIZES by Subject	Total
Randomized Studies									
Cartwright & McIntosh, 1972	1-3	Honolulu, Low Income	262 (9 cl.)	2 yrs.	IQ, Ach.	Random asgmt. to 3 trts: Het, XG, "Flexible." XG classes grouped across grade lines		XG vs Het: Rdg -.17 Math -.52	-.34
								Flex vs Het: Rdg -.28 Math .00	-.14
Matched Studies with Evidence of Initial Equality									
Barker-Lunn, 1970	2-5	England, Wales	5500 (72 sch.)	4 yrs.	IQ, Ach.	Matched schools in social class	Hi 0 Av 0 Lo 0	Rdg/Eng 0 Math 0	0
Goldberg, Passow, & Justman, 1966	5-6	New York City, Middle Class	2219 (86 cl., 45 sch.)	2 yrs.	IQ	Compared classes w/ specified IQ ranges. Students kept in same classes for 2 yrs.	V.Hi 0 Hi (-) Hi Av (-) Av (-) Lo (-)	Rdg (-) Math (-)	(-)
Borg, 1965	4-7, 6	Utah	4-667 (22 cl.) 6-875 (28 cl.)	4 yrs. 1 yr.	Gen. Ach.	Compared 2 districts, Het vs AG	Gr. 4: Hi (+) Av 0 Lo (-)	Rdg 0 Math 0	0
							Gr. 6: Hi + Av (+) Lo 0	Rdg (+) Math (+)	(+)
Hartill, 1936	5-6	New York City	1374 (15 sch.)	1 sem. (see design)	Gen. Ach.	Matched groups - each group Het 1 sem, AG 1 sem. Scores are gains	Hi -.12 Av .00 Lo +.18	Rdg +.05 Math +.01	.00
Barthelmess & Boyer, 1932	4-5	Philadelphia	1130 (10 sch.)	1 yr.	IQ	Schools matched, then students matched. Scores are gains.	Hi +.18 Av +.22 Lo +.15	Composite Ach.	+.21

Table 1, continued

Tobin, 1966	2-6	Marshfield, MA (Rural)	989 (10 cl.)	2 yrs.	Rdg. Ach.	Compared AG to year before AG introduced. Only first 2 yrs data used.	Hi +.13 Av +.03 Lo -.46	Gen ach +.05 Rdg +.13	+.05
Breiden-stine, 1936	2-6	Soudersburg, PA (Rural)	714 (11 sch.)	see design	-	Students in AG or Het classes since 1st grade matched on IQ.		Composite Ach.	-.08
Rankin, Andersen, & Bergman, 1936	3-6	Detroit	~600 (7 sch.)	2 yrs.	IQ	Schools matched on student ach. Scores are gains. Het used no rdg groups. AG used mastery learning in math.	AG vs Het- Hi +.10 Av +.03 Lo +.12 XG vs Het- Hi +.10 Av +.05 Lo +.13	Rdg +.03 Math +.07 Rdg +.11 Math +.03	+.05 +.07
Daniels, 1961	2-5	England	521 (4 sch.)	3.5 yrs.	Gen. Ach.	Schools matched, then students matched on IQ.		Rdg/LA -.25 Math -.27	-.26
Bremer, 1958	1	Amarillo, TX (Anglo)	510	1 yr.	Rdg. Readiness	Compared AG to yr. before AG introduced	Hi -.24 Av .00 Lo -.06	Rdg Ach	-.10
Loomer 1962	4-6	Iowa (Rural)	490 (23 cl.)	1 yr.	IQ, Gen. Ach., Judgment	Schools matched on gen. ach.	Hi -.02 Av +.04 Lo -.06	Composite Ach.	-.04
Flair, 1964	1	Skokie, IL (Suburban)	441 (17 cl.)	1 yr.	Kdg. tchr. prognosis	Schools matched on gen. ach.	V.Hi +.54 Hi -.21 Av -.11	Rdg +.03 Math -.14	-.06
Morgan-stern, 1963	4-6	Uniondale, NY (Suburban)	119 (7 cl.)	3 yrs.	-	Schools matched on IQ	Hi -.22 Av +.15 Lo +.64	Rdg/LA +.17 Math +.06	+.15

KEY: AG = Ability grouped class assignment
Het = Heterogeneous class assignment
XG = Ability grouping across grade lines
\+ = Results clearly favor ability grouped classes
(+) = Results generally favor ability grouped classes
0 = No trend in results
(-) = Results generally favor heterogeneous classes
\- = Results clearly favor heterogeneous classes

Table 2

Ability Grouping for Selected Subjects

Article	Grades	Location	Sample Size	Subjects	Grouping Criteria	Duration	Design	EFFECT SIZES by Achievement	EFFECT SIZES by Subject	Total
Matched Studies with Evidence of Initial Equality										
Moses, 1966	4-6	Calcasieu Parish, LA (Rural)	900 (54 cl.)	Rdg	Rdg Ach	1 sem	Students matched on grade, sex, age, rdg ach.	Hi +.10 Av +.04 Lo -.01	Reading	+.05
Provus, 1960	4-6	Homewood, IL (Suburban, Hi Ach.)	240 (19 cl.)	Math	Math Ach	1 yr	Students matched on IQ. Scores are gains.	Hi +.79 Av +.22 Lo +.15	Math	+.39
Koontz, 1961	4	Norfolk, VA	108 (10 cl.)	Rdg Math Lang	- Rdg Ach - Math Ach - Lang Ach	1 yr	Students matched on gen ach. Scores are grade equivalents.	V.Hi -.12 GE Hi -.25 GE Av -.33 GE Lo -.41 GE	Rdg -.42 GE Math -.29 GE Lang -.12 GE	-.28
Matched Studies Lacking Evidence of Initial Equality										
Berkun, Swanson, & Sawyer, 1966	3-5	Monterrey, CA	1098 (10 sch., 45 cl.)	Rdg	Rdg Ach	1 yr	Compared schools using AG or Het rdg classes. No evidence of initial equality - posttests adj. for pre.	Hi +.44 Lo +.32	Reading	+.32
Davis & Tracy, 1963	4-6	North Carolina	393 (2 sch.)	Math	Math Ach	1 yr	Compared schools using AG or Het math classes. Pretest diffs favored AG. Post adj. for pre, IQ, other vars.		Math	(-)
Balow & Ruddell, 1963	6	Southern Calif. (Suburban)	197 (8 cl.)	Rdg Math	- Rdg Ach - Math Ach	1 yr	Compared AG to het schools. Pretest diffs favored AG. Scores are gains.	Hi 0 Av + Lo +	Rdg (+) Math (+)	(+)

Table 2, continued

Morris, 1969	1-3	Upper Moreland, PA (8 cl.) (Suburban)	117	Rdg - Rdg Ach Math - Math Ach	3 yrs	Compared AG to previous year (het.). IQ diffs favored het. Scores are grade equivalents adjusted for IQ.	Composite Ach.	+.43

KEY: AG = Ability group used for selected subjects
Het = Heterogeneous class assignment
\+ = Results clearly favor ability grouped classes
(+) = Results generally favor ability grouped classes
0 = No trend in results
(-) = Results generally favor heterogeneous classes
\- = Results clearly favor heterogeneous classes

Table 3

Joplin Plan

Article	Grades	Location	Sample Size	Subject	Duration	Design	EFFECT SIZES by Achievement	EFFECT SIZES by Subject	EFFECT SIZES Total
Randomized Studies									
Morgan & Stucker, 1960	5-6	Dundee, MI	360 (8 cl.)	Rdg	1 yr	Students & teachers randomly assigned to Joplin or AG. Authors note limited opportunity for hi ach 6th graders to work above grade level.	Hi +.32 Lo +.94	Rdg	+.30
Hillson, Jones, Moore, & Van Devender, 1964; Jones, Moore, & Van Devender, 1967	1-3	Shamokin, PA	52 (6 cl.)	Rdg	1.5 yrs. 3 yrs. (followup)	Students & teachers randomly assigned to NG/Joplin or het. classes for reading only.		Rdg After 3 yrs: +.33	+.72
Matched Studies with Evidence of Initial Equality									
Russell, 1946	4-5	San Francisco	526 (6 sch.)	Rdg	2 yrs	Matched students in Joplin & het. schools on IQ. Study predated use of term "Joplin Plan." Two reading groups used within regrouped classes.		Rdg	.00
Green & Riley, 1963	4-6	Atlanta (whites)	4 sch.	Rdg	1 yr	Compared Joplin to previous year (het.).		Rdg	+.36
Ingram, 1960	1-3	Flint, MI	68 exp 377 cont	Rdg	3 yrs	Compared NG/Joplin in reading only to previous yr (het.).		Rdg	+.55
Halliwell, 1963	1-3	New Jersey	295	Rdg	1 yr	Compared NG/Joplin in rdg & lang to previous yr (het.). Exp. also sig. higher in math, which was not in program.		Rdg +.53 (gr 1-3) Spel +.58 (gr 2-3) Lang +.26 (gr 3)	+.59
Carson & Thompson, 1964	4-6	Sebastopol, CA	250	Rdg	1 yr	Compared Joplin to AG class assignment matching on IQ. Rdg. grps. were used within Joplin classes.		Rdg	0

Table 4, continued

Matched Studies Lacking Evidence of Initial Equality

Bowman, 1971	1-6	Burlington, NC	457	1 yr.	Compared students in NG, G schools, controlling for IQ. NG started higher in IQ. NG classes used indiv. inst., team teaching, learning centers.	Gr. 1-3 +.06 Gr. 4-6 +.52	Rdg +.29 Math .28	+.29
Hopkins, Oldridge, & Williamson, 1965	3-4	Los Angeles Co., CA (Suburban)	330 (45 cl.)	3 yrs.	Compared students in NG, G schools, controlling for IQ. NG started higher in IQ.		Reading	+.02
Ross, 1967	1-3	Bloomington, IN (Lab school)	314	1 yr.	Compared students in NG, G classes in same university lab school, controlling for pretest and IQ. Pretest and IQ means not given. Scores are grade equivalents adjusted for pretests and IQ.		Rdg +.06GE Math +.06GE	+.06GE
Killough, 1972	1-8	Houston, TX	300 (4 sch.)	3 yrs.	Compared students in 1 NG, 3 G schools. No evidence of initial equality. ANCOVA controlled for IQ.		Rdg + Math +	+
Carbone, 1961	see design	-	244 (6 sch.)	see design	Compared students in NG, G schools controlling for IQ. Students were in grades 4-6, NG had been used in grades 1-3. NG started substantially higher in IQ.		Rdg - LA - Math -	-

KEY: NG = Nongraded Program
G = Graded Program (Control)
\+ = Results clearly favor nongraded classes
(+) = Results generally favor nongraded classes
0 = No trend in results
(-) = Results generally favor graded classes
\- = Results clearly favor graded classes

TABLE 4

Comprehensive Nongraded Plans

Article	Grades	Location	Sample Size	Duration	Design	EFFECT SIZES by Achievement	EFFECT SIZES by Subject	Total
Matched Studies with Evidence of Initial Equality								
Hickey, 1963	1-3	Pittsburgh (Catholic Schools)	1348 (14 sch.)	3 yrs.	Compared NG and G schools matched on SES. IQ's identical. NG program not described.	Hi +.61 Av +.51 Lo +.36	Rdg +.24 Math +.68	+.46
Brody, 1970	1-2	Pennsylvania	268 (3 sch.)	Gr. 1 - 1 yr. Gr. 2 - 2 yrs.	Students in NG and G schools matched on IQ.	Hi + Lo +	Rdg +.14 Spel +.18 Math +.52	+.28
Buffie, 1962	1-3	-	234 (8 sch.)	3 yrs.	Schools matched on SES, then students matched on sex, age, IQ.	Hi +.39 Lo +.19	Rdg +.19 LA +.67 Math +.17	+.35
Otto, 1969	3-5	Austin, TX (Upper middle class, lab school)	15 cl.	2 yrs.	Compared NG and G classes within same university lab school. Students matched on gen. ach.		Rdg 0 Math 0	0
Remacle, 1971	5-6	Brookings, SD	128	Gr. 5 - 2 yrs. Gr. 6 - 1 yr.	Compared NG and G schools. Students matched on IQ. Scores are grade equivalents.		Rdg +.24GE LA +.33GE Math +.37GE	+.31GE
Machiele, 1965	1	Urbana, IL	100	1 yr.	Compared NG to year before NG introduced. Students matched on IQ, age.		Rdg +.64 Math +.36	+.50

Table 4, continued

Matched Studies Lacking Evidence of Initial Equality

Bowman, 1971	1-6	Burlington, NC	457	1 yr.	Compared students in NG, G schools, controlling for IQ. NG started higher in IQ. NG classes used indiv. inst., team teaching, learning centers.	Gr. 1-3 +.06 Gr. 4-6 +.52	Rdg +.29 Math .28	+.29
Hopkins, Oldridge, & Williamson, 1965	3-4	Los Angeles Co., CA (Suburban)	330 (45 cl.)	3 yrs.	Compared students in NG, G schools, controlling for IQ. NG started higher in IQ.		Reading	+.02
Ross, 1967	1-3	Bloomington, IN (Lab school)	314	1 yr.	Compared students in NG, G classes in same university lab school, controlling for pretest and IQ. Pretest and IQ means not given. Scores are grade equivalents adjusted for pretests and IQ.		Rdg +.06GE Math +.06GE	+.06GE
Killough, 1972	1-8	Houston, TX	300 (4 sch.)	3 yrs.	Compared students in 1 NG, 3 G schools. No evidence of initial equality. ANCOVA controlled for IQ.		Rdg + Math +	+
Carbone, 1961	see design	-	244 (6 sch.)	see design	Compared students in NG, G schools controlling for IQ. Students were in grades 4-6, NG had been used in grades 1-3. NG started substantially higher in IQ.		Rdg - LA - Math -	-

KEY: NG = Nongraded Program
G = Graded Program (Control)
\+ = Results clearly favor nongraded classes
(+) = Results generally favor nongraded classes
0 = No trend in results
(-) = Results generally favor graded classes
\- = Results clearly favor graded classes

Table 5

Within-Class Ability Grouping

Article	Grades	Location	Sample Size	No. of Groups	Duration (Months)	Design	EFFECT SIZES by Achievement	by Subject	Total
Randomized Studies									
Slavin & Karweit, 1985 (Experiment 2)	3-5	Hagerstown, MD	366 (17 cl)	2	4	Classes randomly assigned to WCAG, MMP, control	Hi +.41 Av +.21 Lo +.29	Math	+.27
Slavin & Karweit, 1985 (Experiment 1)	4-6	Wilmington, DE	231 (11 cl)	2	5	Classes randomly assigned to WCAG, control	Hi +.13 Av +.30 Lo +.65	Math	+.32
Dewar, 1964	6	Johnson Co., KS (middle class)	199 (8 cl)	3	8	Classes randomly assigned to WCAG, control	Hi +.55 Av +.43 Lo +.67	Math	+.55
Smith, 1960	2-5	Lake Charles, LA	180 (8 cl)	3	5	Classes randomly assigned to WCAG, control, then students matched on sex, age, IQ, math ach.	Hi +.28 Av +.25 Lo +.69	Math	+.41
Wallen & Vowles, 1960	6	Salt Lake City, UT	112 (4 cl)	4	4	Classes randomly assigned to WCAG, control, for 1 sem. Then classes switched treatments for second sem.		Math	+.07

Table 5, continued

Matched Studies with Evidence of Initial Equality									
Spence, 1958	4-6	Bethel, PA	400 (25 cl)	3	8	Students in WCAG, control matched on IQ, math ach. Scores are gains.		Math	+.44
Jones, 1948	4	Richmond, IN	250 (31 cl)	--	8	Students in WCAG, control matched on IQ. Scores are gains.	Hi +.48 GE Av +.27 GE Lo +.37 GE	Rdg +.23 Spel +.43 Math +.16	+.26
Matched Studies Lacking Evidence of Initial Equality									
Stern, 1972	3-4	Covina, CA (Low achievers)	217 exp 91 cont (o7 cl)	--	4	Compared WCAG, control matched on general ach level, low achievers only. Control began higher at pretest.		Math	+.36

KEY: WCAG = Within class ability grouping correlations are known, effect sizes from gain scores can be transformed to the scale of posttest values using the following multiplier:

$$ES = (ES_{gain})(\sqrt{2\,(1-r_{pre\text{-}post})})$$

Because few studies presenting gain scores also provide pre-post correlations, a pre-post correlation of +0.8 was assumed. This figure is a characteristic correlation between fall and spring scores on alternate forms of the California Achievement Test in the upper elementary grades (CTB/McGraw-Hill, 1979). Substituting 0.8 in the formula, a multiplier of 0.632 is derived, which was used to deflate effect size estimates from gain score data. Because this value is only a rough approximation, effect sizes from gain score data should be interpreted with even more caution than is warranted for effect sizes in general.

In studies in which pretest da a were provided, effect sizes were computed as gain scores, divided by the control group's post-test standard deviation. This procedure adjusts effect sizes for any differences in pretest scores. In a few cases, pretest and posttest scores were from different tests. In these studies (e.g., Flair, 1964), experimental-control differences divided by control standard deviations were computed for pre- and posttests, and the difference between these is reported as the study's effect size. Since all studies which met inclusion criteria presented either gain scores or pre- and post-test scores or matched on pretests, all effect sizes were adjusted for initial starting points.

If studies did not present enough data to allow for computation of effect size but otherwise met criteria for inclusion, they were included in tables with an indication of the direction and consistentcy of any achievement differences. In some cases only grade equivalent differences were given, and these are presented in the table. Because the standard deviations of grade equivalents are around 1.0 in upper elementary school, grade equivalent differences may be considered very rough approximations of effect size.

In general, one overall effect size is presented for each study, unless two or more different ability grouping plans were compared to heterogeneous control groups in the same study (e.g., Cartwright and McIntosh, 1972) or two distinct samples were studied (e.g., Borg, 1965). Multiple effects within a study were averaged to obtain the overall effect size estimate (see Bangert-Drowns, in press). If studies presented adequate data, overall effect sizes were also broken down by subject (e.g., reading, mathematics) and by achievement or ability level. Effect sizes by ability level should be interpreted with particular caution, as they are often inflated because standard deviations within subgroup categories are restricted in range.

In this best-evidence synthesis, every effort was made to make each effect size be a meaningful representation of the effect of ability grouping on student posttest achievement, holding the posttest standard deviation as the common metric. In all tables, randomized studies are listed first, followed by matched studies pre-